A Family-based Graphical Approach for Testing Hierarchically Ordered Families of Hypotheses

Zhiying Qiu
Biostatistics and Programming, Sanofi
Bridgewater, NJ 08807, U.S.A.

Li Yu
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, U.S.A.

Wenge Guo*
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, U.S.A.
Email: wenge.guo@njit.edu

December 1, 2018

*The research of Wenge Guo was supported in part by NSF Grant DMS-1309162.

Abstract

In applications of clinical trials, tested hypotheses are often grouped as multiple hierarchically ordered families. To test such structured hypotheses, various gatekeeping strategies have been developed in the literature, such as series gatekeeping, parallel gatekeeping, tree-structured gatekeeping strategies, etc. However, these gatekeeping strategies are often either non-intuitive or less flexible when addressing increasingly complex logical relationships among families of hypotheses. In order to overcome the issue, in this paper, we develop a new family-based graphical approach, which can easily derive and visualize different gatekeeping strategies. In the proposed approach, a directed and weighted graph is used to represent the generated gatekeeping strategy where each node corresponds to a family of hypotheses and two simple updating rules are used for updating the critical value of each family and the transition coefficient between any two families. Theoretically, we show that the proposed graphical approach strongly controls the overall familywise error rate at a pre-specified level. Through some case studies and a real clinical example, we demonstrate simplicity and flexibility of the proposed approach.

KEY WORDS: Graphical approach, gatekeeping strategy, familywise error rate, multiple testing, error rate function.

1 Introduction

In clinical trial research, it is becoming increasingly common to consider the problems of complex multiple testing due to hierarchically ordered multiple objectives. In these problems, the hypotheses to be tested are usually grouped into multiple families, and these families are tested in a sequential manner. For example, there are usually multiple endpoints of interest in clinical trials and these endpoints are generally classified as primary, secondary and sometimes tertiary endpoints which form a natural hierarchical structure. To deal with such structured multiple testing problems, Maurer, Hothorn and Lehmacher (1995) and Bauer et al. (1998) introduced a convenient and efficient way called gatekeeping strategy based on which

hypotheses in one family cannot be tested if the testing results of the previous families do not meet some pre-specified gatekeeping conditions. Basically, there are two types of gatekeeping strategies. One is serial gatekeeping (Westfall and Krishen, 2001) in which each family can be tested using any FWER controlling procedure if and only if all hypotheses in the previous families are rejected. The other is parallel gatekeeping (Dmitrienko, Offen and Westfall, 2003) in which the subsequent family can be tested if and only if at least one hypothesis in current family is rejected.

Tree-structured gatekeeping strategy introduced by Dmitrienko, Wiens and Tamhane (2007) and its extension, mixture procedure, introduced by Dmitrienko and Tamhane (2011, 2013) were also developed for testing hierarchically ordered families of hypotheses with complex logical relationships. However, both the tree-structured gatekeeping strategy and mixture procedure were derived based on the closure principle of Marcus et al. (1976). Thus, to implement these procedures, intensive computation is unavoidable. To avoid such complex computational issue caused by the closure principle, Dmitrienko, Tamhane, Wang and Chen (2006), Guibaud (2007) and Dmitrienko, Tamhane and Wiens (2008) developed a simple stepwise approach for implementing gatekeeping strategies. Dmitrienko, Tamhane and Wiens (2008) introduced a general multistage gatekeeping procedure, which unified the above works. Due to the stepwise shortcut, the multistage gatekeeping procedure is apparently more straightforward and easier to explain to the clinicians in practice. However, to deal with complex logical restrictions, multistage gatekeeping procedure is less flexible compared with the mixture procedure, although the latter is computationally intensive.

With increasing complexity of hierarchically logical restrictions of gatekeeping strategies, the proper visualization and presentation of such strategies will be very helpful for users. To develop such visualization tool, one solution is to employ the idea of graphical approaches proposed by Bretz et al. (2009) and Burman et al. (2009). The graphical approaches have been used to for sequentially testing hierarchically structured hypotheses, such as superchain procedure proposed by Kordzakhia and Dmitrienko (2013), where each family is presented as a vertex and the local significance levels are propagated via transition coefficients between families instead of hypotheses. However, this approach tests all families of hypotheses si-

multaneously at each step which is not suitable in most clinical trial settings, such as families of hypotheses having hierarchical structure. Maurer and Bretz (2014) developed a graphical approach for testing families of hypotheses which is able to visualize the serial gatekeeping procedure in the sense that only if all hypotheses in a single family are rejected, the graph can be updated.

In this paper, we are motivated to propose a new family-based graphical approach which can be more flexible to visualize the hierarchically logical restrictions of the usual gatekeeping procedures than the existing graphical approaches. This approach can serve as an extension of multistage gatekeeping procedure in the sense that it not only takes advantage of the stepwise algorithm but also deals with more general logical restrictions than the multistage gatekeeping procedure. For example, the proposed graphical approach can also be applied to some complex multiple testing problems where equally important families of hypotheses are grouped in the same layer, e.g, primary endpoints and co-primary endpoints.

The rest of the paper is organized as follows. We discuss our research motivation through an example and briefly introduce the idea of our family-based graphical approach in Section 2.1. We then present some basic notations and assumptions in Section 2.2. In Section 3, we introduce the general algorithm for sequentially testing families of hypotheses and show its overall FWER control. In Section 4, we show the advantages of our approach through three case studies in Bretz et al. (2009). A real data analysis is performed in Section 5. Some concluding remarks are made in Section 6 and all proofs are deferred to Appendix.

2 Preliminary

In this section, we will discuss our research motivation though a heuristic example and introduce some basic notations and assumptions.

2.1 Heuristics

Bretz et al. (2009) introduce a general graphical approach which provides a graphical tool to visualize Bonferroni-adjusted gatekeeping procedures. As an example, Figure 1 shows such graphical visualization of the parallel gatekeeping strategy based on

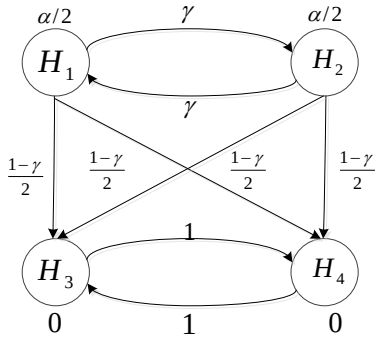


Figure 1: Graphical visualization of the gatekeeping procedure with truncated Holm procedure as gatekeeper.

a truncated Holm procedure that is used for testing four hypotheses grouped as two families, where each hypothesis is represented by a vertex. Compared with the conventional multiple testing procedures for testing a single family of hypotheses, the hypothesis-based graphical approach is indeed explicit and efficient. However, in practice, increasingly complex clinical trials problems often involve testing multiple ordered families of hypotheses, which makes such hypothesis-based graphs more complicated, even not applicable in some settings of a large number of families.

Consider an example that 9 hypotheses are grouped into 3 families where each family consists of three hypotheses, denoted as $F_i = \{H_{i1}, H_{i2}, H_{i3}\}$, for $i = 1, 2, 3$. Suppose that F_1 and F_2 are sequentially tested by a truncated Holm procedure and F_3 is tested by the conventional Holm procedure. The subsequent family of hypotheses can be tested if and only if at least one hypothesis in the current family is rejected. Figure 2 illustrates the hypothesis-based graphical visualization of the parallel gatekeeping strategy. Due to its complexity, the weights on the edges are omitted in this graph. As seen from Figure 2, the hypothesis-based graph is relatively unclear and complicated, although it only involves testing 3 families of hypotheses.

While testing multiple families of hypotheses, hierarchically logical restrictions among the families are often one important aspect. Thus, it is natural for us to focus more on the logical relationships at family level rather than at hypothesis level, to develop a graphical approach for visualizing conventional gatekeeping strategies for testing multiple ordered families of hypotheses. By using the similar idea as in Ko-

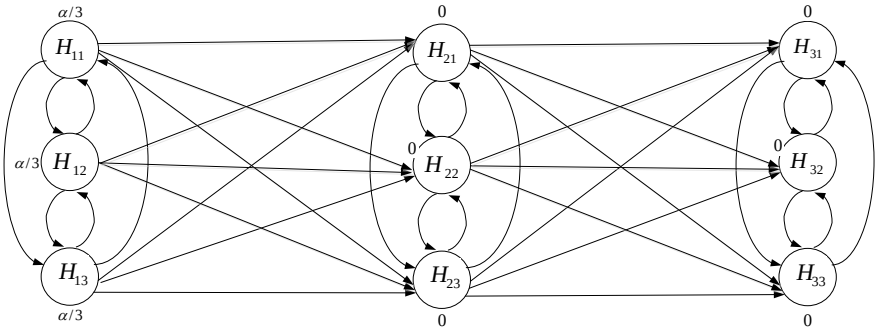
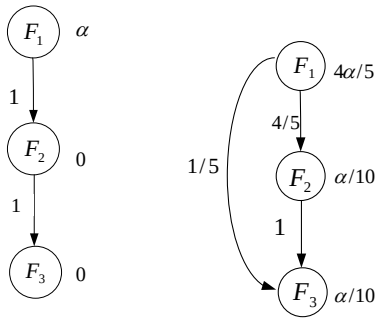


Figure 2: Hypothesis-based graphical visualization of gatekeeping procedure with truncated Holm procedure with truncation parameter γ as gatekeeper.

rdzakhia and Dmitrienko (2013), we use a vertex to represent a family of hypotheses instead of an individual hypothesis and a directed edge with a pre-specified weight associated with it to represent the transition relationship between two families. We term this approach as *family-based graphical approach*. For the example illustrated in Figure 2, an equivalent family-based graph is shown in Figure 3 (a), where the families $F_i, i = 1, 2, 3$ are represented by vertices. As seen from Figure 3 (a), we start testing F_1 at level α; the subsequent family $F_2(F_3)$ can be tested if and only if at least one rejection is made while testing the current family $F_1(F_2)$. The allocation of the critical values among families is via transition coefficients on the edges, that is, after rejections are made in one family, the critical value of this family is proportionally transferred to the subsequent families based on the transition coefficients on the edges from the family to the subsequent families. For more details of the updating rule, see Section 3.

To make the example in Figure 3 (a) more interesting, consider a specific parallel gatekeeping strategy for which the initial critical values of F_1, F_2 and F_3 are respectively $4\alpha/5, \alpha/10$ and $\alpha/10$ and except for transferring to F_2, 1/5 of the critical value of F_1 can be passed down to F_3 if at least one hypothesis is rejected in F_1. Figure 3 (b) illustrates the family-based graph of this parallel gatekeep strategy. As seen from Figure 3 (b), even when there is no rejections in F_1, the subsequent F_2 and F_3 can still be tested at their local critical values.



(a) strategy 1 (b) strategy 2

Figure 3: Family-based graphical visualization of parallel gatekeeping strategies 1 (a) and 2 (b).

2.2 Basic notations

In this subsection, we present some basic notations and definitions. Suppose there are $N \geq 2$ hypotheses divided into $m \geq 2$ families, which are further grouped into n layers, with $L_i = \{F_{i1}, \ldots, F_{il_i}\}$ being the ith ordered layer consisting of l_i families of hypotheses, $i = 1, \ldots, n, \sum_{i=1}^{n} l_i = m$. Each family F_{ij} within layer L_i has $n_{ij} \geq 1$ null hypotheses, denoted as $F_{ij} = \{H_{ij1}, \cdots, H_{ijn_{ij}}\}$, for $j = 1, \cdots, l_i$ such that $\sum_{i=1}^{n} \sum_{j=1}^{l_i} n_{ij} = N$. These families F_{ij} of hypotheses are to be tested based on their respective p-value $P_{ijk}, k = 1, \cdots, n_{ij}$, subject to controlling an overall measure of type I error at a pre-specified level α. Each of the true null p-value is assumed to be stochastically greater than or equal to the uniform distribution on $[0, 1]$; that is, if T_{ij} is the set of true null hypotheses in F_{ij}, then for any fixed $u \in [0, 1]$,

$$\Pr\{P_{ijk} \leq u | H_{ijk} \in T_{ij}\} \leq u, \tag{1}$$

for any $i = 1, \cdots, n$, $j = 1, \cdots, l_i$, and $k = 1, \cdots, n_{ij}$.

The familywise error rate (FWER), which is the probability of incorrectly rejecting at least one true null hypothesis, is a commonly used notion of an overall measure of type I error when testing a single family of hypotheses. Since we have multiple layers with any number of families within each layer, we consider this measure not locally for each family but globally. In other words, we define the overall FWER as the probability of incorrectly rejecting at least one true null hypothesis

across all families of hypotheses for all layers. If it is bounded above by α regardless of which and how many null hypotheses within each family are true for any layer, then this overall FWER is said to be strongly controlled at α.

In this paper, we propose a general procedure, called family-based graphical approach, strongly controlling the overall FWER at α. Given the pre-specified critical value α, let α_i denote the initial critical values assigned to layer L_i with $\sum_{i=1}^{n} \alpha_i \leq \alpha$. Moreover, let α_{ij} denote the initial critical values assigned to families F_{ij} within layer L_i with $\sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha_i$. The procedure starts with testing L_1 to L_n sequentially and within each layer L_i, families F_{ij} are tested in any order using any local procedures based on their own (local) critical values. The critical values used to locally test each family within the current layer is updated from its initially assigned value to one which incorporates certain portions of the critical values used in testing the families within the previous layers. This procedure stops testing when all families of the last layer L_n are tested. The specific updating rule for local critical values is described in Section 3. The distribution of the amount of critical values transferred among families can be pre-fixed by a transition coefficient set $\mathbf{G}$ which is defined as follows.

Let $\mathbf{G} = \{g_{ijkl}\}$ denote a set of all transition coefficients g_{ijkl} which satisfies the following conditions for any $i = 1, \ldots, n$ and $j = 1, \ldots, l_i$:

$$\sum_{k=i+1}^{n} \sum_{l=1}^{l_k} g_{ijkl} \leq 1; \quad 0 \leq g_{ijkl} \leq 1; \quad g_{ijkl} = 0 \text{ if } i \geq k.$$

Note that g_{ijkl} is defined as the proportion of the local critical value that can be transferred from family F_{ij} within layer L_i to family F_{kl} within layer L_k. Figure 4 shows the graphical representation of the general family-based approach.

Based on the initial critical values α_{ij} and the transition coefficients g_{ijkl}, we can construct a *directed acyclic graph* for the aforementioned family-based approach. In this graph, each family F_{ij} is represented by a vertex associated with its initial critical value α_{ij}; for any two vertices corresponding to two respective families F_{ij} and F_{kl}, if the transition coefficient g_{ijkl} from F_{ij} to F_{kl} is positive, then a directed edge between these two vertices is displayed, where F_{ij} and F_{kl} are head and tail vertices, respectively. Since each vertex is associated with a family instead of a hypothesis, we term the graph as a *family-based graph*, which is illustrated in Figure

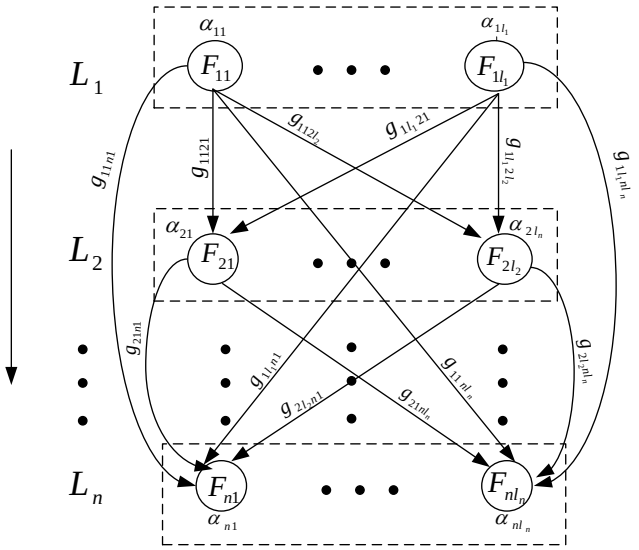


Figure 4: Graphical representation of general family-based graphical approach.

4.

Our specific updating rule for local critical values, which is described in Section 3, is defined based on error rate function introduced in Dmitrienko et al. (2008). The error rate function is defined as follows.

Definition 1 (Dmitrienko et al., 2008) *Consider a single family of hypotheses,* $F = \{H_1, \cdots, H_n\}$ *and a multiple testing procedure for testing the family* F*. The error rate function of this procedure is defined as*

$$e(I) = \sup_{H_I} \Pr\left\{ \bigcup_{i \in I} \{reject\, H_i\} \,\middle|\, H_I \right\}$$

for any $I \subseteq \{1, 2, \cdots, n\}$*, where* $H_I = \bigcap_{i \in I} H_i$ *is the intersection of hypotheses* H_i *with* $i \in I$*.*

Note that in applications, if the error rate function $e(\cdot)$ cannot be calculated easily, we often use one of its upper bounds $e^*(\cdot)$ to replace it.

In the family-based approach, each family is tested by its own local procedure, thus it is associated with a particular error rate function. Let α^*_{ij} denote the local

critical value for testing family F_{ij} and A_{ij} denote the set of accepted hypotheses in F_{ij}. Based on A_{ij}, we can calculate $e^*(A_{ij})$ after testing F_{ij} at level α^*_{ij} and then transfer the remaining amount of its local critical value $\alpha^*_{ij} - e^*(A_{ij})$ to the respective families in the subsequent layers according to the corresponding transition coefficients.

Remark 1 The error rate function introduced in Dmitrienko et al. (2008) was used to develop a simple stepwise approach for parallel gatekeeping strategies. In their discussion, the error rate function is required to be strictly less than α unless all of the hypotheses in one family are rejected, which is termed as separability condition. In this paper, the definition of the error rate function we used is a a little bit more general. For this function, the separability condition is not required when choosing local procedures for our suggested family-based graphical approach.

3 Methodology

In this section, we introduce a new family-based graphical approach and show its overall FWER control. We begin in Subsection 3.1 with a simple case of two layers with two families of hypotheses within each layer. The general case of multiple layers with arbitrary number of families within each layer is discussed in Subsection 3.2.

3.1 Two-layer family-based graphical approach with four families

Consider $m = 4$ families of hypotheses being divided into two layers L_1, L_2 based on their hierarchal relationships, with two families of hypotheses within each layer. By using the notations introduced in Section 2.2, we define a two-layer family-based graphical approach through the following algorithm:

Algorithm 1

Step 1. *Set* $L_1 = \{F_{11}, F_{12}\}, L_2 = \{F_{21}, F_{22}\}$. *Test family* $F_{1j}, j = 1, 2$, *using any FWER controlling procedure at critical value* α_{1j}, *and calculate* $e^*(A_{1j})$.

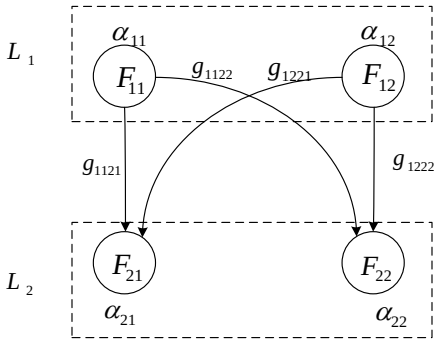


Figure 5: Graph for two layer family-based procedure with m=4.

Update the graph:

$$L_1 \to L_1 \backslash \{F_{1j}\};\ \textit{for } k = 1, 2,\ \textit{let}$$
$$\alpha_{2k} \to \alpha_{2k} + (\alpha_{1j} - e^*_{1j}(A_{1j}))g_{1j2k};$$
$$g_{1l2k} \to \begin{cases} g_{1l2k},\ l \neq j. \\ 0,\ \textit{otherwise}. \end{cases}$$

If $L_1 \neq \emptyset$, go back to step 1; otherwise, go to next step.

Step 2. *Test F_{2k}, $k = 1, 2$, using any FWER controlling procedure at level α_{2k} and update the graph:*

$$L_2 \to L_2 \backslash \{F_{2k}\}.$$

If $L_2 \neq \emptyset$, go back to step 2; otherwise stop.

Algorithm 1 starts the test from the families $F_{1j}, j = 1, 2$, in L_1. Once F_{1j} is tested, the critical value of F_{2k} is updated based on the error rate function $e^*_{1j}(A_{1j})$ and the transition coefficient set $\mathbf{G}$; moreover, $\mathbf{G}$ itself is updated by deleting all the elements associated with F_{1j}. This procedure can be fully described by a graph displayed in Figure 5. For Algorithm 1, we have the following theorem.

Theorem 1 *Under the conditions of the corresponding local procedures controlling the FWER within each family of hypotheses, the two-layer multiple testing procedure described in Algorithm 1 strongly controls the overall FWER at level α.*

For the proof of Theorem 1, see Appendix A.1.

3.2 General multi-layer family-based graphical approach

The aforementioned two-layer four-family case demonstrates the inherent nature of sequential testing of the family-based graphical approach. Now we generalize the graphical approach from two layers with two families of hypotheses in each layer to any n layers with arbitrary number of families of hypotheses within each layer. The general multi-layer family-based graphical approach is defined through the following algorithm:

Algorithm 2

Step i $(\mathbf{1 \leq i \leq n-1})$. *Test family* $F_{ij}, j = 1, \cdots, l_i$ *using any FWER controlling procedure at level* α_{ij}*, and calculate* $e^*_{ij}(A_{ij})$.
Update the graph:

$$
\begin{aligned}
&L_i \to L_i \backslash \{F_{ij}\};\ \textit{for } k = i+1, \cdots, n,\ l = 1, \cdots, l_k,\ \textit{let} \\
&\alpha_{kl} \to \alpha_{kl} + (\alpha_{ij} - e^*_{ij}(A_{ij}))g_{ijkl}; \\
&g_{iskl} \to \begin{cases} g_{iskl},\ s \neq j. \\ 0,\ otherwise. \end{cases}
\end{aligned}
$$

If $L_i \neq \emptyset$*, go back to step i; otherwise, go to next step.*
Step n. *Test* $L_n = \{F_{n1}, \cdots, F_{nl_n}\}$*. Use any FWER controlling procedure at level* α_{nj} *to test* F_{nj} *and update* $L_n \to L_n \backslash \{F_{nj}\}$*. If* $L_n \neq \emptyset$*, go back to step n; otherwise stop.*

For this general multi-layer family-based graphical approach, we have the following theorem.

Theorem 2 *Under the conditions of the corresponding local procedures controlling the FWER within each family of hypotheses, the general multi-layer family - based graphical approach strongly controls the overall FWER at level* α*.*

For the proof of Theorem 2, see Appendix A.2.

Remark 2 Consider a specific problem of testing hierarchically ordered families of hypotheses, where there are n layers, $L_1, \ldots, L_n$ and for each layer L_i, there is

only one family F_{i1}. To deal with this multiple testing problem, consider a multi-layer family-based graphical approach, whose initial critical value for F_{i1} is α if $i = 1$ and 0 otherwise; whose transition coefficients are given by $g_{i1k1} = 1$, if $1 \leq i \leq n-1, k = i+1$ and 0 otherwise. Regarding this graphical approach, we have the following several remarks.

1. If each family is tested using a local procedure controlling the FWER and satisfying separability condition, i.e., the error rate function of the local procedure is strictly smaller than α when at least one hypothesis is not rejected within the family, then the multi-layer family-based graphical approach reduces to a specific parallel gatekeeping strategy, which is in turn equivalent to a general multistage gatekeeping procedure introduced by Dmitrienko et al. (2008). The examples of such local procedures include the conventional Bonferroni procedure, truncated Holm procedure, truncated fallback procedure, etc, see Dmitrienko et al. (2008).

2. If each family is tested using a FWER controlling local procedure for which the upper bound of its error rate function is given by $e^*(I) = \alpha$ for any $I \neq \emptyset$, then the corresponding multi-layer graphical approach is equivalent to a specific serial gatekeeping strategy. The examples of such local procedures including the conventional Holm procedure and fixed sequence procedure, etc.

3. If each family has only one null hypothesis, then the multi-layer graphical approach reduces to the conventional fixed sequence procedure.

4. If some correlation information regarding the null p-values within one family is known in advance, then there are more options for local procedures. For example, if the null p-values in a family are known to be positive dependent or independent, then we can use the conventional or truncated Hochberg procedure as its local procedure.

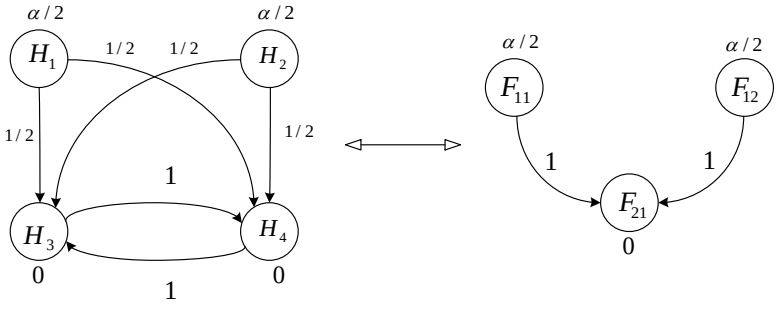


Figure 6: Hypothesis-based (left) and family-based (right) graphical visualization of Case 1.

4 Discussions

In this section, we use three cases shown in Bretz et al. (2009) to illustrate the efficiency and simplicity of our proposed family-based graphical approach as compared to the conventional hypothesis-based graphical approach in dealing with the problem of testing multiple families of hypotheses. These cases are respectively visualized in Figures 6-8, in which the original hypothesis-based graphs in Bretz et al. (2009) are displayed in the left side, and their corresponding family-based graphs are displayed in the right side.

Case 1 Consider a case in Figure 6 with four null hypotheses H_1, H_2, H_3 and H_4. The left side of Figure 6 displays the hypothesis-based graphical procedure and its right side displays an equivalent family-based graphical procedure, where these four null hypotheses are grouped as $m = 3$ families, $F_{11} = \{H_1\}, F_{12} = \{H_2\}$ and $F_{21} = \{H_3, H_4\}$, and $n = 2$ layers, $L_1 = \{F_{11}, F_{12}\}$ and $L_2 = \{F_{21}\}$. The initial critical values allocated to the three families are respectively $\alpha/2, \alpha/2$ and 0, and the transition coefficient set $\mathbf{G}$ is given by

$$g_{1121} = g_{1221} = 1;$$
$$g_{2111} = g_{2112} = g_{1112} = g_{1211} = 0.$$

The family-based procedure starts with testing F_{11} (or F_{12}) using the Bonferroni method at level $\alpha_{11} = \alpha/2$. If H_1 is rejected, the critical value $\alpha/2$ of F_{11} is transferred to F_{21} as indicated by the transition coefficient 1 on the directed edge

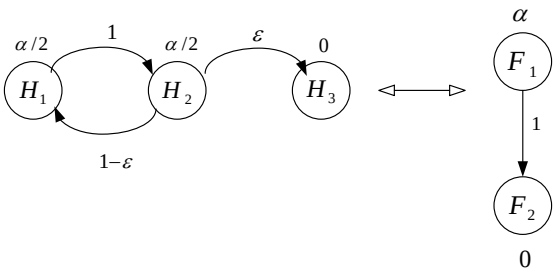


Figure 7: Hypothesis-based (left) and family-based (right) graphical visualization of Case 2.

from F_{11} to F_{21}, such that the critical value $\alpha_{21} = 0$ of F_{21} is updated to $\alpha_{21}^* = \alpha/2$. If H_1 is not rejected, no critical value is transferred to F_{21}. Then, the procedure continues testing F_{12} using the Bonferroni method at level $\alpha_{12} = \alpha/2$. Once H_2 is rejected, its critical value $\alpha/2$ will be added to α_{21}^*. Otherwise, no critical value is transferred to F_{21}. After testing both F_{11} and F_{12} in L_1, if $\alpha_{21}^* \neq 0$, we continue testing F_{21} in L_2 using the Holm procedure at level α_{21}^*. Through the whole testing process, we can see that our family-based graphical procedure is equivalent to the hypothesis-based graphical procedure displayed in Figure 6 (left). It is easy to observe from Figure 6 (right) that family-based graphical visualization describes the hierarchical relationship among the families of hypotheses more simply and clearly, as compared to hypothesis-based graphical visualization. □

There are often some situations where the hypotheses in one family can be tested only if all the hypotheses in another family are rejected. If one uses the original hypothesis-based graphical approach to deal with such multiple testing problems, the generated graphs often include the edges with infinitesimally small weights, which are complex and difficult to communicate to non-statisticians. However, it is shown in the following that the infinitesimally small weights can be removed in the graphs by using our suggested family-based graphical approach.

Case 2 Consider a case of gatekeeping strategy involving testing three hypotheses H_1, H_2 and H_3. Suppose only if both H_1 and H_2 are rejected, H_3 has the chance to be tested. The hypothesis-based graph of this gatekeeping strategy is shown in Figure 7 (left) with an edge associated with an infinitesimally small weight ϵ.

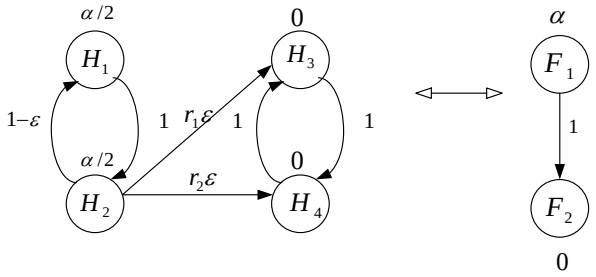


Figure 8: Hypothesis-based (left) and family-based (right) graphical visualization of Case 3.

When using the family-based graphical approach, the generated family-based graph is shown in Figure 7 (right), where the edge with the infinitesimally small weight ϵ is removed. As seen from Figure 7 (right), this method turns out to be a simple two layers, two families procedure with $L_1 = \{F_1\}$ and $L_2 = \{F_2\}$, where $F_1 = \{H_1, H_2\}$ and $F_2 = \{H_3\}$; the initial critical values for F_1 and F_2 are α and 0, respectively. Thus, the specific gatekeeping strategy can be described as follows: start testing F_1 using the conventional Holm procedure at level α. If both hypotheses in F_1 are rejected, then its critical value α are passed on to F_2 such that F_2 is tested at level α. Otherwise, the test stops. □

Case 3 Consider a more complicated case of gatekeeping strategy involving testing four hypotheses H_1, H_2, H_3 and H_4. Suppose that H_3 and H_4 are of interest only if both H_1 and H_2 are rejected. The hypothesis-based graph of this gatekeeping strategy is shown Figure 8 (left) with the edges associated with infinitesimally small weights. As seen from Figure 8 (left), if both hypotheses H_1 and H_2 are rejected, the critical value α is proportionally assigned to H_3 and H_4 according to the weights r_1 and r_2 such that H_3 receives $r_1\alpha$ and H_4 receives $r_2\alpha$. When using the family-based graphical approach, the generated family-based graph is shown in Figure 8 (right). As seen from Figure 8 (right), this method turns out to be a simple two layers, two families procedure with $L_1 = \{F_1\}$ and $L_2 = \{F_2\}$ where $F_1 = \{H_1, H_2\}$ and $F_2 = \{H_3, H_4\}$. The initial critical values for F_1 and F_2 are α and 0, respectively. Thus, the specific procedure can be described as follows: perform the conventional Holm procedure for testing F_1 at level α. If both H_1 and H_2 are rejected, its critical

value α is passed on to F_2 and unlike Case 2, we then perform a weighted Holm procedure with weights r_1 and r_2 for testing F_2 at α. Otherwise, the test stops. $\square$

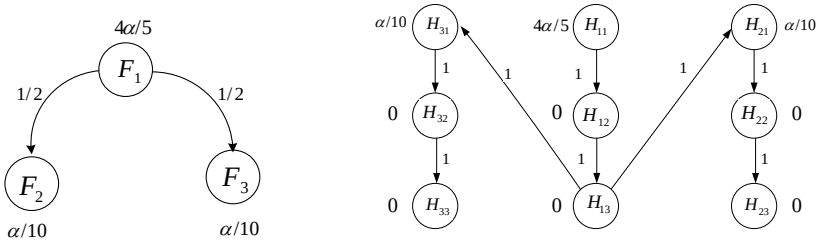
Remark 3 Through discussions of the above three cases, it is easy to see that when dealing with complex problems of testing multiple families of hypotheses, our proposed family-based graphical approach usually makes the whole testing process more clearly and easier to communicate to non-statisticians as compared to the conventional hypothesis-based graphical approach, which often involves with non-intuitive infinitesimally small weights ϵ.

5 A Clinical Trial Example

In this section, we consider a clinical trial example to illustrate the application of our proposed family-based graphical approach and compare its performance with that of the conventional hypothesis-based graphical approach.

We revisit the Type II diabetes clinical trial example in Dmitrienko et al. (2007). The trial compares three doses of an experimental drug (Doses L, M and H) versus placebo (Plac) with respect to one primary endpoint (P: Haemoglobin A1c), and two secondary endpoints (S1: Fasting serum glucose; S2: HDL cholesterol). The three endpoints will be examined at each of the three doses, so a total of nine null hypotheses will be formulated and grouped into three families, F_1, F_2 and F_3. Family F_1 consists of three dose-placebo comparisons corresponding to the primary endpoint (P): H vs Plac (H_{11}), M vs Plac (H_{12}) and L vs Plac (H_{13}). Similarly, family F_2 consists of three dose-placebo comparisons corresponding to the secondary endpoint S1: H vs Plac (H_{21}), M vs Plac (H_{22}) and L vs Plac (H_{23}) and family F_3 consists of three dose-placebo comparisons corresponding to the secondary endpoint S2: H vs Plac (H_{31}), M vs Plac (H_{32}) and L vs Plac (H_{33}).

The overall Type I error rate is pre-specified at $\alpha = 0.05$ and the raw p-values for the nine null hypotheses are given in Table 1. In this example, we assume that the primary endpoint P is more important than the secondary endpoints $S1$ and $S2$, thus F_1 is always tested before testing F_2 and F_3. For F_2 and F_3, we consider two types of hierarchical relationships below and thus discuss two different gatekeeping strategies, Procedure 1 and 2. We visualize these two procedures by using the



(a) Famly-based (b) Hypothesis-based

Figure 9: Family-based graph (a) and hypothesis-based graph (b) for Procedure 1 in the Type II diabetes clinical trial.

family-based and hypothesis-based graphical approaches, respectively.

Procedure 1. Suppose that the secondary endpoints $S1$ and $S2$ are equally important, thus F_2 and F_3 are grouped into the same layer; the dose-placebo comparisons within each family are ordered a priori (H vs. Plac through L vs. Plac). We choose the conventional fixed sequence procedure as local procedure for each family and the initial allocation of critical values for F_1, F_2 and F_3 are $0.04, 0.005$, and 0.005, respectively. Once F_1 is tested and all of its hypotheses are rejected, its critical value is equally allocated to F_2 and F_3. Figure 9 (a) visualizes this gatekeeping strategy. We start testing F_1 at level 0.04; all of three hypotheses in F_1 are rejected using the conventional fixed sequence procedure. Then, all of its local critical value 0.04 is equally assigned to F_2 and F_3 and the updated critical values for F_2 and F_3 become $0.005 + 0.02 = 0.025$. We continue to test F_2 and F_3 at level 0.025 in any order using the conventional fixed sequence procedure; the resulting rejected hypotheses are H_{21}, H_{31} and H_{32}. Finally, the testing results of Procedure 1 are summarized in Table 1. In addition, Figure 9 (b) provides a graphical visualization for Procedure 1 by using the hypothesis-based graphical approach. As seen from Figure 9, compared to the hypothesis-based graph, the family-based graph provides more clear and intuitive illustrations of the hierarchical relationships among the families of hypotheses.

Procedure 2. Suppose that the secondary endpoint $S1$ is more important than $S2$, thus F_1, F_2 and F_3 are tested in a pre-defined order. Consider the gatekeeping

Table 1: Comparison of results of two-layer (Procedure 1) and three-layer (Procedure 2) family-based procedures in the Type II diabetes clinical trial. The overall Type I error rate is $\alpha = 0.05$. Note: S=significant; NS=not significant.

Null hypothesis	**Raw *p*-value**	**Procedure 1**	**Procedure 2**
H_{11}	0.005	S	S
H_{12}	0.011	S	S
H_{13}	0.018	S	S
H_{21}	0.009	S	S
H_{22}	0.026	NS	S
H_{23}	0.013	NS	S
H_{31}	0.010	S	S
H_{32}	0.006	S	S
H_{33}	0.051	NS	NS

strategy visualized in Figure 3 (b) for which the truncated Hochberg procedure with truncation parameter $\gamma = 0.5$ is used as local procedure for testing F_1 and F_2; the conventional Hochberg procedure is used for testing F_3. The initial allocation of critical values for F_1, F_2 and F_3 are $0.04, 0.005$, and 0.005, respectively. We start testing F_1 at level 0.04; all of three hypotheses in F_1 are rejected using the truncated Hochberg procedure; the updated critical values for F_2 and F_3 are $0.04*0.8+0.005 = 0.037$ and $0.04 * 0.2 + 0.005 = 0.013$, respectively. We then test F_2 at level 0.037 using the same truncated Hochberg procedure; all of the three hypotheses in F_2 are rejected as well and its local critical value is transferred to F_3; the updated critical value of F_3 is $0.013 + 0.037 = 0.05$. Finally, we test F_3 at level 0.05; thus H_{31} and H_{32} are rejected. The testing results of Procedure 2 are also summarized in Table 1. We need to note that the conventional hypothesis-based graphical approach is not applicable to visualize Procedure 2.

6 Conclusions

In this paper, we developed a new family-based graphical approach for testing hierarchically ordered families of hypotheses. Theoretically we proved that the proposed graphical approach strongly controls the FWER at a pre-specified level. By using the proposed approach, we can easily develop and visualize various gatekeeping strategies. Specifically, when each layer has only one family, the proposed approach reduces to Dmitrienko et al. (2008)'s general multistage gatekeeping strategies.

Though case studies and a real clinical trial example, we showed that the proposed approach is simpler and more efficient as compared to Bretz et al. (2009)'s hypothesis-based graphical approach when dealing with the problem of testing multiple hierarchically ordered families. In addition, due to its family-based graphical visualization, our proposed approach will be easier to communicate to the non-statisticians than the original hypothesis-based graphical approach when dealing with increasingly complex hierarchical relationships among families of hypotheses.

Appendix

A.1 Proof of Theorem 1

Suppose that the family F_{ij} is tested at level α^*_{ij}, then we know that

$$\begin{aligned} \alpha^*_{1j} &= \alpha_{1j}, \\ \alpha^*_{2i} &= \alpha_{2i} + \sum_{j=1}^{2} (\alpha^*_{1j} - e^*_{1j}(A_{1j}))g_{1j2i}. \end{aligned} \tag{2}$$

For $i, j = 1, 2$, define the event $E_{ij}(x) = \{$at least one true null hypothesis being rejected in F_{ij} at significant level $x\}$. Let $\overline{E}_{ij}(x)$ denote the complement of $E_{ij}(x)$. Thus,

$$\begin{aligned} \text{FWER} &= \Pr\left\{ \bigcup_{i=1}^{2} \bigcup_{j=1}^{2} E_{ij}(\alpha^*_{ij}) \right\} \\ &= \Pr\left\{ \bigcup_{j=1}^{2} E_{1j}(\alpha^*_{1j}) \right\} + \Pr\left\{ (\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha^*_{1j})) \bigcap (\bigcup_{j=1}^{2} E_{2j}(\alpha^*_{2j})) \right\}, \end{aligned} \tag{3}$$

where $\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)$ is the complement set of $\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)$.

Let T_{ij} denote the set of true null hypotheses in F_{ij}, and R_{ij} and A_{ij} denote the sets of rejections and acceptances, respectively.

First of all, let us consider the first term of the right side of (3). Note that

$$\Pr\left\{\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)\right\} \leq \sum_{j=1}^{2} \Pr\left\{E_{1j}(\alpha_{1j}^*)\right\} \leq \sum_{j=1}^{2} e_{1j}^*(T_{1j}). \tag{4}$$

Here, the first inequality follows from the Bonferroni inequality and the second follows from the definition of the error rate function.

Next, we consider the second term of the right side of (3). If $\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)$ is true, i.e., all of the rejected hypotheses in F_{11} and F_{12} are false, then $T_{11} \subseteq A_{11}$ and $T_{12} \subseteq A_{12}$, which implies $e_{11}^*(T_{11}) \leq e_{11}^*(A_{11})$ and $e_{12}^*(T_{12}) \leq e_{12}^*(A_{12})$, respectively. Then, by (2), we have

$$\begin{aligned} \alpha_{2i}^* &= \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e_{1j}^*(A_{1j})) g_{1j2i} \\ &\leq \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i}. \end{aligned}$$

Thus,

$$\begin{aligned} &\left(\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)\right) \bigcap \left(\bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*)\right) \\ \subseteq &\bigcup_{j=1}^{2} E_{2j}\left(\alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i}\right) \end{aligned}$$

and then by the above result and the Bonferroni inequality,

$$\begin{aligned} &\Pr\left\{\left(\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)\right) \bigcap \left(\bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*)\right)\right\} \\ \leq &\Pr\left\{\bigcup_{i=1}^{2} E_{2j}\left(\alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i}\right)\right\} \\ \leq &\sum_{i=1}^{2} \Pr\left\{E_{2j}\left(\alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i}\right)\right\}. \end{aligned} \tag{5}$$

Note that the fact that families F_{2j} are tested by FWER controlling local procedures and the probability inside the sum in the second inequality of (5) is exactly the FWER of the local procedures at level $\alpha_{2i}+\sum_{j=1}^{2}(\alpha_{1j}^{*}-e_{1j}^{*}(T_{1j}))g_{1j2i}$, thus the right side of (5) is bounded above by

$$\begin{aligned}
& \sum_{i=1}^{2}\left(\alpha_{2i}+\sum_{j=1}^{2}(\alpha_{1j}^{*}-e_{1j}^{*}(T_{1j}))g_{1j2i}\right) \\
= & \sum_{i=1}^{2}\alpha_{2i}+\sum_{j=1}^{2}\left(\alpha_{1j}-e_{1j}^{*}(T_{1j})\right)\sum_{i=1}^{2}g_{1j2i} \\
\leq & \sum_{i=1}^{2}\alpha_{2i}+\sum_{j=1}^{2}(\alpha_{1j}-e_{1j}^{*}(T_{1j})) \\
= & \sum_{i=1}^{2}\alpha_{2i}+\sum_{j=1}^{2}\alpha_{1j}-\sum_{j=1}^{2}e_{1j}^{*}(T_{1j}) \\
\leq & \alpha-\sum_{j=1}^{2}e_{1j}^{*}(T_{1j}). \qquad (6)
\end{aligned}$$

The first inequality of (6) follows from the fact that $\sum_{i=1}^{2}g_{1j2i}\leq 1$ for any $j=1,2$.

Therefore, using (4)-(6) in (3), we have

$$\text{FWER} \leq \sum_{j=1}^{2}e_{1j}^{*}(T_{1j})+\alpha-\sum_{j=1}^{2}e_{1j}^{*}(T_{1j})=\alpha.$$

Thus, the desire result is proved. □

A.2 Proof of Theorem 2

Let $\text{FWER}_n(\alpha_1,\cdots,\alpha_n)$ denote the overall FWER of the multi-layer family-based procedure for which the initial critical values assigned to layers L_i are $\alpha_i, i=1,\ldots,n$. Within each layer L_i, suppose that the initial critical values assigned to families F_{ij} are $\alpha_{ij}, j=1,\ldots,l_i$ with $\sum_{j=1}^{l_i}\alpha_{ij}\leq\alpha_i$. We show the following inequality by using induction,

$$\text{FWER}_n(\alpha_1,\cdots,\alpha_n)\leq\sum_{i=1}^{n}\sum_{j=1}^{l_i}\alpha_{ij}\leq\alpha. \qquad (7)$$

If $n = 2$, through the proof of Theorem 1, we can get that $\text{FWER}_2(\alpha_1, \alpha_2) \leq \sum_{i=1}^{2} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha$.
Assume that (7) holds when $n = k, k \geq 2$, which is

$$\text{FWER}_k(\alpha_1, \cdots, \alpha_k) \leq \sum_{i=1}^{k} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha.$$

In the following, we show that (7) also holds for $n = k + 1$, i.e.,

$$\text{FWER}_{k+1}(\alpha_1, \cdots, \alpha_{k+1}) \leq \sum_{i=1}^{k+1} \alpha_i \leq \alpha.$$

Define the events $B_1 = \{$at least one true null being rejected among all the families in layer 1$\}$ and $B_2 = \{$at least one true null being rejected among the families in all the layers except layer 1$\}$. Then we have

$$\text{FWER}_{k+1}(\alpha_1, \cdots, \alpha_{k+1}) = \Pr\{B_1\} + \Pr\left\{\overline{B}_1 \bigcap B_2\right\}. \tag{8}$$

Note that

$$\Pr\{B_1\} \leq \sum_{j=1}^{l_1} e_{1j}^*(T_{1j}), \tag{9}$$

which follows from the definition of error rate function and the Bonferroni inequality. Let us consider the probability of the event $\overline{B}_1 \cap B_2$ below.

After testing all families in L_1, the total significant level $\sum_{j=1}^{l_1}(\alpha_{1j} - e_{1j}^*(A_{1j}))$ of layer L_1 will be transferred to the respective families from L_2 to L_n. Specifically, for family F_{ij} with layer L_i, its updated significant level is

$$\alpha_{ij}^* = \alpha_{ij} + \sum_{l=1}^{l_1}(\alpha_{1l} - e_{1l}^*(A_{1l}))g_{1lij}.$$

Let $\alpha_i^* = \sum_{j=1}^{l_i} \alpha_{ij}^*$ denote the updated critical value for layer L_i.

If $\overline{B}_1$ is true, which means that no true null hypotheses are rejected in any families within L_1, then it implies that type I error can only occur in the families of layers L_2 to L_{k+1}. Thus,

$$\Pr\left\{\overline{B}_1 \bigcap B_2\right\} = \text{FWER}_k(\alpha_2^*, \cdots, \alpha_{k+1}^*). \tag{10}$$

Note $\overline{B}_1$ being true also implies that for any $F_{1j}, j = 1, \cdots, l_1, T_{1j} \subseteq A_{1j}$, which in turn implies $e^*_{1j}(T_{1j}) \leq e^*_{1j}(A_{1j})$ due to the monotonicity condition of error rate function. Thus, by the induction assumption,

$$\begin{aligned}
&\mathrm{FWER}_k(\alpha^*_2, \cdots, \alpha^*_{k+1}) \leq \sum_{i=2}^{k+1}\sum_{j=1}^{l_i} \alpha^*_{ij} \\
= &\sum_{i=2}^{k+1}\sum_{j=1}^{l_i}\left[\alpha_{ij} + \sum_{l=1}^{l_1}(\alpha_{1l} - e^*_{1l}(A_{1l}))g_{1lij}\right] \\
= &\sum_{i=2}^{k+1}\sum_{j=1}^{l_i}\alpha_{ij} + \sum_{l=1}^{l_1}\alpha_{1l}\sum_{i=2}^{k+1}\sum_{j=1}^{l_i} g_{1lij} - \sum_{l=1}^{l_1} e^*_{1l}(A_{1l})\sum_{i=2}^{k+1}\sum_{j=1}^{l_i} g_{1lij} \\
\leq &\sum_{i=2}^{k+1}\sum_{j=1}^{l_i}\alpha_{ij} + \sum_{l=1}^{l_1}\alpha_{1l} - \sum_{l=1}^{l_1} e^*_{1l}(A_{1l}) \\
\leq &\sum_{i=1}^{k+1}\sum_{j=1}^{l_i}\alpha_{ij} - \sum_{j=1}^{l_1} e^*_{1j}(T_{1j}). \qquad (11)
\end{aligned}$$

The second inequality of (11) holds due to the condition of transition matrix that for any fixed $k = 1, \cdots, l_1, \sum_{i=2}^{k+1}\sum_{j=1}^{l_i} g_{1lij} \leq 1$. Therefore, by combining (8)-(11), we have

$$\mathrm{FWER}_{k+1}(\alpha_1, \cdots, \alpha_{k+1}) \leq \sum_{i=1}^{k+1}\sum_{j=1}^{l_i}\alpha_{ij} \leq \alpha.$$

This completes the induction, and show that (7) holds for any positive n. □

References

[1] Bauer P., Rohmel J., Maurer W. and Hothorn L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

[2] Bretz F., Maurer W., Brannath W. and Posch M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **28**, 586–604.

[3] Burman C. F., Sonesson C. and Guilbaud O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* **28**, 739–761.

[4] Dmitrienko A., Offen W. and 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.

[5] Dmitrienko A. and Tamhane A. C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine* **30**, 1473–1488.

[6] Dmitrienko A. and Tamhane A. C. (2013). General theory of mixture procedures for gatekeeping. *Biometrical Journal* **5**, 311–320.

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

[8] Dmitrienko A., Tamhane A. C., Wang X. and Chen X. (2006). Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* **48**, 984–991.

[9] Dmitrienko A., Tamhane A. C. and Wiens B. L. (2008). General multistage gatekeeping procedures. *Biometrical Journal* **50**, 667–677.

[10] Dmitrienko A., Wiens B. L. and Tamhane A. C. (2007). Tree–structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* **26**, 2465–2478.

[11] Guilbaud O. (2007). Bonferroni parallel gatekeeping - transparent generalizations, adjusted p-values, and short direct proofs. *Biometrical Journal* **49**, 917–927.

[12] Kordzakhia G. and Dmitrienko A. (2013). Superchain procedures in clinical trials with multiple objectives. *Statistics in Medicine* **32**, 486–508.

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

[14] Maurer W. and Bretz F. (2014). A note on testing families of hypotheses using graphical procedures. *Statistics in Medicine* **30**, 5340–5346.

[15] Maurer W., Hothorn L. and Lehmacher W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der Chemisch-pharmazeutischen Industrie,* Vollmar J(ed.). Fischer Verlag: Stuttgart, **6**, 3–18.

[16] Westfall P. H. and Krishen A. (2001). Optimally weighted, fixed-sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.