

# A Graphical Framework for Testing Hierarchically Structured Hypothesis Families

Zhiying Qiu

Biostatistics at Insmed  
Bridgewater, NJ 08807, U.S.A.

Li Yu

Biostatistics at Syndax Pharmaceuticals  
New York, NY 10017, 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](mailto:wenge.guo@njit.edu)

April 12, 2026

## Abstract

In clinical trials, hypotheses are frequently organized into hierarchically ordered families, requiring specialized testing strategies that account for these structured relationships. Existing gatekeeping methods—including serial, parallel, and tree-structured approaches—provide important solutions but are often either too rigid or insufficiently intuitive to accommodate increasingly complex logical dependencies among hypothesis families. To address these limitations, we propose a novel family-based graphical approach that unifies the derivation and visualization of diverse gatekeeping strategies. In this framework, procedures are represented as directed, weighted graphs, where nodes correspond to hypothesis families. Two simple updating rules govern the allocation of significance levels within families and the propagation of significance levels between them. We establish that the proposed method strongly controls the familywise error rate (FWER) at a pre-specified level. Simulation studies under representative configurations indicate that the proposed procedure achieves performance comparable to hypothesis-level graphical approaches and competitive with the superchain procedure, while providing a simpler and more interpretable family-level representation. Case studies and a real clinical trial application further illustrate its flexibility and practical advantages, making it a powerful tool for managing hierarchically structured multiple testing in clinical research.

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

## 1 Introduction

In clinical trial research, complex multiple testing problems often arise from hierarchically ordered objectives (Alosh et al., 2014; Tamhane and Gou, 2018). These typically involve hypotheses grouped into families of endpoints—primary, secondary, and sometimes tertiary—that must be tested in a sequential manner (Dmitrienko et al., 2009). Such hierarchical structures require specialized procedures that ensure strong control of the familywise error rate (FWER) while respecting clinical priorities.

To address these challenges, Maurer (1995) and Bauer et al. (1998) introduced the gatekeeping strategy, in which hypotheses in later families may only be tested if earlier families satisfy predefined gatekeeping conditions. Gatekeeping procedures are broadly classified into two types. In *serial*

*gatekeeping* (Westfall and Krishen, 2001), a family is tested only if all hypotheses in the preceding families are rejected. In *parallel gatekeeping* (Dmitrienko et al., 2003), a subsequent family may be tested if at least one hypothesis in the current family is rejected.

More advanced approaches were developed to address complex logical dependencies among hypothesis families. The tree-structured gatekeeping strategy of Dmitrienko et al. (2007) and its extension—the mixture procedure (Dmitrienko and Tamhane, 2011, 2013)—enable flexible testing of hierarchically ordered families under intricate logical constraints. Tree-structured gatekeeping operates within a fixed hierarchical framework with pre-specified logical relationships, and the testing strategy strictly follows this structure. In contrast, mixture procedures are formulated under the closure principle by constructing combined intersection  $p$ -values through valid mixing functions applied to family-specific intersection  $p$ -values. This formulation permits more general parametric multiple testing procedures and greater flexibility in handling logical constraints. However, because mixture procedures rely on the closure principle, they require evaluation of a potentially large collection of intersection hypotheses. The number of such intersections grows exponentially with the number of hypotheses (or families), which can lead to substantial computational burden in large-scale settings.

Subsequent developments include Xi and Tamhane (2014); Dmitrienko et al. (2016); Kordzakhia et al. (2018); Wang et al. (2022). To improve practical implementation, simpler stepwise gatekeeping procedures were proposed (Dmitrienko et al., 2006; Guilbaud, 2007; Dmitrienko et al., 2008). In particular, the general multistage procedure of Dmitrienko et al. (2008) unified earlier methods into a coherent and implementable framework that is easier to communicate in clinical applications. Nevertheless, stepwise procedures typically offer less flexibility than mixture-based methods when handling highly complex logical structures.

As logical restrictions in trial objectives become increasingly intricate, effective visualization and communication tools are essential. Graphical approaches, such as those proposed by Bretz et al. (2009) and Burman et al. (2009), provide not only intuitive representations of sequential testing strategies but also a formal mechanism for encoding multiplicity control. Within these frameworks, transition coefficients play a pivotal role: they determine how unused portions of the significance level are reallocated across hypotheses, thereby governing the propagation of significance levels and explicitly capturing the logical dependencies among testing objectives.

Building on this idea, the superchain procedure of [Kordzakhia and Dmitrienko \(2013\)](#) represents families as vertices and propagates significance levels through pre-specified transition coefficients. Although elegant and flexible, this approach tests all families simultaneously, which may reduce its suitability for strictly hierarchical trial designs. To better accommodate such settings, [Maurer and Bretz \(2014\)](#) proposed a graphical method tailored to serial gatekeeping, where the graph is updated only after all hypotheses within a family are rejected, thus aligning more closely with hierarchical testing structures.

In this paper, we introduce a new *family-based graphical approach* that combines the transparency of graphical methods with the flexibility of stepwise gatekeeping. The method generalizes the multistage gatekeeping framework, enabling a broader range of logical restrictions to be accommodated. For instance, it naturally addresses scenarios in which equally important families (such as co-primary endpoints) must be tested within the same layer. By operating directly at the family level, the proposed algorithms extend existing graphical approaches and enhance the transparency and interpretability of hierarchical testing strategies.

Family-level graphical representations have appeared in clinical trial protocols primarily for communication purposes; for example, the CANTOS trial protocol ([Ridker et al., 2017](#)) employed diagrams to illustrate the hierarchical testing strategy. However, such diagrams are typically conceptual and do not uniquely determine a valid multiple testing procedure, as they do not encode the within-family testing method or the rule governing the redistribution of unused significance levels. Consequently, the same diagram may correspond to different implementations (e.g., serial versus parallel gatekeeping).

The proposed family-based graphical framework shares this general feature in that the graphical structure alone does not uniquely determine a valid testing procedure. Additional specifications are required, including the choice of within-family testing methods and the associated error-rate function bounds that govern the redistribution of significance levels. Our framework makes these components explicit and integrates them with the graphical representation. As a result, once the graph and the associated local procedures and error-rate functions are specified, they together define a well-defined testing algorithm and its level-updating mechanism. This formulation reduces ambiguity in interpretation and facilitates a rigorous analysis of strong FWER control.

Conceptually, the proposed framework may be viewed as a structured subclass of the more

general superchain procedures of [Kordzakhia and Dmitrienko \(2013\)](#), in that both rely on graph-based propagation of significance levels. However, the two approaches differ substantially in scope and design objectives. Superchain procedures permit iterative re-testing and flexible recycling schemes across objectives, whereas our method is formulated as a sequential multi-layer, family-based procedure with a single forward pass, fixed transition coefficients, and no iterative updating. The primary goal of our framework is to provide a regulator-friendly construction that emphasizes transparency, interpretability, and ease of communication at the family level.

From a practical standpoint, this development is particularly relevant in modern clinical trials, where study designs frequently involve multiple co-primary endpoints, composite outcomes, or key secondary objectives that are critical for regulatory approval. Existing methods are often either too rigid to accommodate such complex logical structures or too computationally intensive for routine implementation. Beyond visualization, our family-based graphical approach formalizes the logical flow of hypothesis testing, explicitly encodes the updating and redistribution of significance levels, and enables procedures tailored to specific study objectives and regulatory constraints. This framework assists statisticians in designing and implementing coherent multiple testing strategies, supports clinicians in interpreting testing hierarchies and decision pathways, and facilitates regulatory reviewers in evaluating the logical structure and error-control properties of confirmatory analyses. By combining theoretical rigor with operational transparency, the proposed approach helps bridge the gap between methodological sophistication and practical clinical application, thereby enhancing the interpretability and acceptance of advanced multiple testing strategies.

The remainder of the paper is organized as follows. Section [1.1](#) motivates the proposed approach through two heuristic examples. Section [1.2](#) introduces the notation and assumptions. Section [2](#) presents the general sequential testing algorithm and establishes FWER control. Section [3](#) illustrates the advantages of the approach through three case studies from [Bretz et al. \(2009\)](#). Section [4](#) provides a real data application. Section [5](#) presents a simulation study evaluating the finite-sample performance of the proposed method. Section [6](#) concludes with a discussion of future research directions. Technical proofs are collected in the Appendix.

## 1.1 Heuristics

Bretz et al. (2009) introduced a general graphical framework as a visualization tool for Bonferroni-based gatekeeping procedures. In particular, they described graphical implementations of truncated Holm-type gatekeeping procedures (see Bretz et al. (2009), Section 3), in which the redistribution of unused significance levels is governed by pre-specified truncation parameters. Such graphical representations provide an intuitive way to visualize how significance levels are allocated and propagated across hypotheses.

For completeness, we briefly recall the truncated Holm procedure (Dmitrienko et al., 2008), which serves as a building block in many graphical gatekeeping strategies.

**Definition 1.1** (Truncated Holm procedure). Let  $H_1, \dots, H_m$  be  $m$  hypotheses with corresponding  $p$ -values  $p_1, \dots, p_m$ , and let  $p_{(1)} \leq \dots \leq p_{(m)}$  denote the ordered  $p$ -values, where  $H_{(i)}$  corresponds to  $p_{(i)}$ . For a truncation fraction  $0 \leq \gamma \leq 1$  and overall significance level  $\alpha$ , the truncated Holm procedure compares  $p_{(i)}$  with

$$w_i \alpha = \left( \frac{\gamma}{m - i + 1} + \frac{1 - \gamma}{m} \right) \alpha, \quad i = 1, \dots, m.$$

Starting from  $i = 1$ , reject  $H_{(i)}$  if  $p_{(i)} \leq w_i \alpha$  and continue sequentially until the first non-rejection occurs. When  $\gamma = 0$ , the procedure reduces to the Bonferroni procedure; when  $\gamma = 1$ , it coincides with the classical Holm step-down procedure.

Figure 1 provides a schematic illustration of a parallel gatekeeping strategy based on a truncated Holm procedure (Dmitrienko et al., 2008). In this example, four hypotheses are grouped into two families, and each hypothesis is represented by a vertex in the graph. Directed edges indicate possible transfers of significance levels following rejections. The figure is intended to illustrate the graphical construction rather than to uniquely represent all possible truncated Holm parameterizations.

Compared with conventional multiple testing procedures applied to a single family of hypotheses, hypothesis-based graphical approaches offer a clear and flexible representation of complex testing strategies. However, in many clinical trials multiple hierarchically ordered families of hypotheses must be tested. As the number of families grows, the corresponding hypothesis-level graphs may become increasingly complex and difficult to interpret.

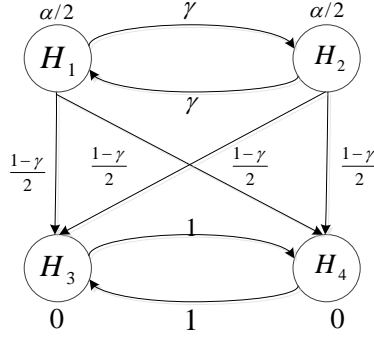


Figure 1: Schematic graphical representation of a gatekeeping procedure based on the truncated Holm method (after [Bretz et al. \(2009\)](#)).

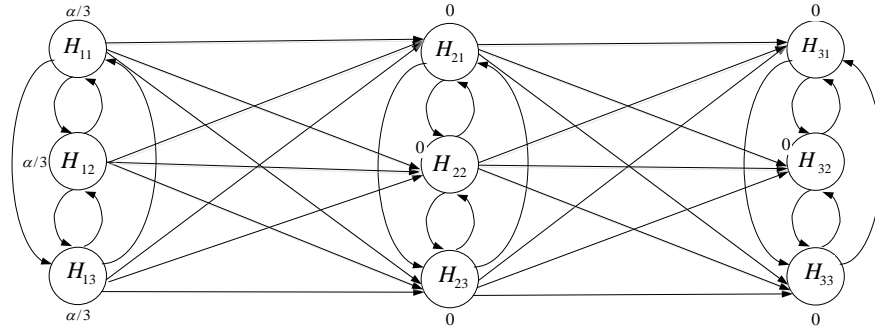


Figure 2: Schematic hypothesis-based graphical representation of a gatekeeping strategy using a truncated Holm procedure with truncation parameter  $\gamma$  (after [Bretz et al. \(2009\)](#)).

To illustrate this issue, consider an example with nine hypotheses grouped into three families, each containing three hypotheses, denoted by  $F_i = \{H_{i1}, H_{i2}, H_{i3}\}$  for  $i = 1, 2, 3$ . Suppose that  $F_1$  and  $F_2$  are tested sequentially using a truncated Holm procedure with truncation parameter  $\gamma$  as described in [Bretz et al. \(2009\)](#), while  $F_3$  is tested using the standard Holm procedure. Furthermore, a subsequent family is tested only if at least one hypothesis in the current family is rejected. Figure 2 shows a schematic hypothesis-based graphical representation of this strategy. To simplify the presentation, the edge weights are omitted. The figure is intended to illustrate the structural complexity of the hypothesis-level representation rather than to correspond to a unique specification of truncation parameters.

In many practical applications, hierarchical logical relationships among families of hypotheses

play a central role in determining the testing strategy. In such settings, it is often more natural and informative to represent the structure at the *family level* rather than at the level of individual hypotheses. Motivated by the superchain framework of [Kordzakhia and Dmitrienko \(2013\)](#), we consider a *family-based graphical representation*, in which each vertex represents a family of hypotheses. Directed edges, together with associated weights, describe how significance levels may be transferred between families following rejections.

For example, the family-based representation corresponding to the strategy in [Figure 2](#) is shown in [Figure 3\(a\)](#), where each family  $F_i$  ( $i = 1, 2, 3$ ) is represented as a vertex. Testing begins with  $F_1$  at significance level  $\alpha$ . The next family,  $F_2$  (or  $F_3$ ), can be tested only if at least one hypothesis in  $F_1$  (or  $F_2$ ) is rejected. Significance levels are transferred between families through transition coefficients assigned to edges: when a rejection occurs within a family, a portion of its significance level may be redistributed to subsequent families according to these coefficients. Detailed updating rules are provided in [Section 3](#).

It is important to note that the graphical structure itself does not uniquely determine the gatekeeping strategy. In particular, the same graph may correspond to different implementations (e.g., serial or parallel gatekeeping) depending on the choice of within-family testing procedures and the associated error-rate-function bounds. Thus, the graph should be interpreted together with these statistical components. The representation in [Figure 3\(a\)](#) is intended to illustrate the hierarchical relationships and a simple forward-propagation mechanism under a specific choice of local procedures.

To further demonstrate the flexibility of this framework, consider a parallel gatekeeping strategy in which the initial significance levels allocated to  $F_1$ ,  $F_2$ , and  $F_3$  are  $4\alpha/5$ ,  $\alpha/10$ , and  $\alpha/10$ , respectively. In addition,  $1/5$  of the significance level from  $F_1$  may be transferred to  $F_3$  if at least one hypothesis in  $F_1$  is rejected. [Figure 3\(b\)](#) shows the corresponding family-based graphical representation. In this case, the parallel gatekeeping structure is induced by the specification of the local testing procedures and error-rate-function bounds, rather than by the graph alone. Even if no rejection occurs in  $F_1$ , the subsequent families  $F_2$  and  $F_3$  may still be tested at their allocated local significance levels.

We note that alternative edge structures (for example, allowing additional propagation between families) can also be specified within the same framework and, together with different choices of

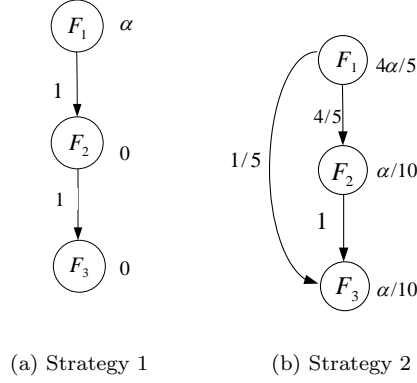


Figure 3: Family-based graphical visualization of gatekeeping strategies: (a) strategy 1 and (b) strategy 2.

local procedures, may lead to procedures with different power characteristics. The present examples are intended primarily to illustrate the structural representation rather than to identify a power-optimal design.

## 1.2 Basic Notation and Framework

We introduce the notation and structural assumptions underlying the proposed procedure. Suppose  $N \geq 2$  null hypotheses are organized into  $m \geq 2$  families, which are further arranged into  $n$  ordered layers.

The  $i$ -th layer is denoted by

$$L_i = \{F_{i1}, \dots, F_{il_i}\}, \quad i = 1, \dots, n,$$

where  $l_i$  is the number of families in layer  $L_i$  and  $\sum_{i=1}^n l_i = m$ . Each family  $F_{ij}$  in layer  $L_i$  contains  $n_{ij} \geq 1$  hypotheses,

$$F_{ij} = \{H_{ij1}, \dots, H_{ijn_{ij}}\}, \quad j = 1, \dots, l_i,$$

so that the total number of hypotheses satisfies

$$\sum_{i=1}^n \sum_{j=1}^{l_i} n_{ij} = N.$$

Each hypothesis  $H_{ijk}$  is associated with a  $p$ -value  $P_{ijk}$ . The overall objective is to control the

familywise error rate (FWER) at a pre-specified level  $\alpha$ .

For any true null hypothesis, the corresponding  $p$ -value is assumed to be stochastically no smaller than  $\text{Uniform}(0, 1)$ . Specifically, letting  $T_{ij}$  denote the set of true nulls in  $F_{ij}$ , we assume that for all  $u \in [0, 1]$ ,

$$\Pr(P_{ijk} \leq u \mid H_{ijk} \in T_{ij}) \leq u, \quad \forall i, j, k. \quad (1)$$

*Remark 1.1* (Definition of families and layers). The partition of hypotheses into families and layers is assumed to be specified prior to testing. In confirmatory clinical trials, this structure is typically determined by the clinical objectives and endpoint hierarchy described in the study protocol or statistical analysis plan (SAP), rather than by statistical convenience alone.

Common grouping principles include: (i) grouping by endpoint type (primary, key secondary, secondary), (ii) grouping dose–placebo comparisons within the same endpoint, (iii) placing co-primary endpoints within the same layer when they share equal priority, and (iv) aligning families with pre-specified gatekeeping logic in the SAP.

More generally, the relationship between hypothesis-based and family-based representations is not one of simple equivalence. A family-based structure can be viewed as a higher-level abstraction of a more detailed hypothesis-based graph. To obtain a family-based representation from a hypothesis-level graph, additional design choices are required, including how hypotheses are grouped into families and how transition weights are aggregated across families. These choices are generally not unique and must be made in a way that preserves the clinical hierarchy and yields a valid testing procedure.

The proposed framework does not prescribe a unique grouping. Instead, it provides a rigorous testing procedure once a clinically meaningful family structure has been specified, together with the associated transition rules, the local testing procedures within each family, and the corresponding error-rate-function bounds.

**Familywise error rate.** The familywise error rate (FWER) is defined as the probability of rejecting at least one true null hypothesis. In the layered setting, the *overall FWER* is the probability of rejecting at least one true null across all families and layers. A procedure is said to *strongly control* the FWER at level  $\alpha$  if this probability does not exceed  $\alpha$  under any configuration of true null hypotheses.

**Significance level allocation.** Let  $\alpha_i$  denote the initial significance level assigned to layer  $L_i$ , and let  $\alpha_{ij}$  denote the portion allocated to family  $F_{ij} \subset L_i$ . These allocations must satisfy:

**Condition 1.1** (significance level allocation).

$$\sum_{i=1}^n \alpha_i \leq \alpha, \quad \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha_i \quad \text{for each } i = 1, \dots, n.$$

Testing proceeds sequentially across layers  $L_1, \dots, L_n$ . Within each layer, families may be tested in arbitrary order using their allocated significance levels. Unused portions of significance levels may be transferred to later families according to transition coefficients defined below.

**Transition coefficients.** Let  $\mathbf{G} = \{g_{ijkl}\}$  denote the transition coefficients, where  $g_{ijkl}$  represents the proportion of the local significance level from family  $F_{ij}$  that may be transferred to family  $F_{kl}$  in a subsequent layer. These coefficients must satisfy:

**Condition 1.2** (Transition coefficients). For all  $i, j$ ,

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

The collections  $\{\alpha_{ij}\}$  and  $\{g_{ijkl}\}$  define a directed acyclic graph (DAG):

- Each vertex corresponds to a family  $F_{ij}$  labeled with its significance level  $\alpha_{ij}$ .
- A directed edge from  $F_{ij}$  to  $F_{kl}$  is present if  $g_{ijkl} > 0$ .

Because vertices represent families rather than individual hypotheses, we refer to this structure as a *family-based graph*.

**Error rate function.** The updating mechanism relies on the error rate function introduced by [Dmitrienko et al. \(2008\)](#).

**Definition 1.2** ([Dmitrienko et al. \(2008\)](#)). For a family  $F = \{H_1, \dots, H_n\}$  with an associated multiple testing procedure, the *error rate function* is defined as

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

for any  $I \subseteq \{1, \dots, n\}$ , where  $H_I = \bigcap_{i \in I} H_i$ .

When  $e(\cdot)$  is not available in closed form, we work with an upper bound  $e^*(\cdot)$ .

**Example 1.1.** For the Bonferroni procedure with local level  $\alpha$ , the error rate function satisfies

$$e(I) \leq \frac{|I|}{n} \alpha, \quad I \subseteq \{1, \dots, n\},$$

since each hypothesis is tested at level  $\alpha/n$ .

For the truncated Holm procedure with truncation fraction  $\gamma$ , an upper bound on the error rate function is given by

$$e^*(I) = \begin{cases} 0, & I = \emptyset, \\ [\gamma + (1 - \gamma)|I|/n]\alpha, & I \subseteq \{1, \dots, n\}, I \neq \emptyset. \end{cases}$$

In particular,  $e^*(I) < \alpha$  for all non-empty  $I$  when  $\gamma \in [0, 1)$ , implying that the procedure is separable. These bounds are used in the level-updating mechanism.

*Remark 1.2.* In [Dmitrienko et al. \(2008\)](#), the error rate function was developed in a parallel gatekeeping setting under a separability condition requiring  $e(I_i) < \alpha_i$  unless all hypotheses in the family are rejected. This condition ensures that the unused portion of the local significance level is well defined for level transfer. In the present layered framework, we do not impose this separability condition as a general requirement. Instead, because significance levels may be partially allocated and redistributed across families, the updating mechanism relies directly on the error-rate-function bound (or its upper bound  $e^*(\cdot)$ ) rather than on separability.

**Local procedures.** Each family  $F_{ij}$  is tested using a local procedure characterized by an error rate function (or bound)  $e_{ij}^*(\cdot)$ . Let  $\alpha_{ij}^*$  denote the effective significance level assigned to  $F_{ij}$  during testing, and let  $A$  denote the set of accepted hypotheses. After testing  $F_{ij}$ , the quantity

$$\alpha_{ij}^* - e_{ij}^*(A)$$

represents the unused portion of the significance level, which is redistributed to subsequent families according to the transition coefficients.

We impose the following standard conditions on local procedures.

**Condition 1.3** (FWER control). For all  $\alpha \in (0, 1)$ ,

$$\text{FWER}(\alpha) \leq \alpha.$$

**Condition 1.4** (Monotonicity). For subsets  $A \subseteq B$ ,

$$e_{ij}^*(A) \leq e_{ij}^*(B).$$

**Condition 1.5** ( $\alpha$ -consistency). For  $\alpha_1 \leq \alpha_2$ ,

$$R(\alpha_1) \subseteq R(\alpha_2),$$

where  $R(\alpha)$  denotes the rejection set at level  $\alpha$ .

## 2 Methodology

In this section, we introduce a novel *family-based graphical approach* and establish its strong control of the overall familywise error rate (FWER). We first describe the two-layer case with four hypothesis families in Subsection 2.1, which serves as a motivating example. The general case with multiple layers and an arbitrary number of families per layer is presented in Subsection 2.2.

The proposed family-based graphical framework relies on the availability of an error-rate function (or a valid upper bound) for the local testing procedure within each family. These local error-rate-function bounds are used to determine how significance levels can be redistributed across families while maintaining global error control.

### 2.1 Two-layer case with four hypothesis families

Consider  $m = 4$  families of hypotheses arranged into two layers,  $L_1$  and  $L_2$ , with two families in each layer.

Using the notation introduced in Section 1.2, the procedure is described formally below.

**Algorithm 3.1** (Two-layer family-based procedure)

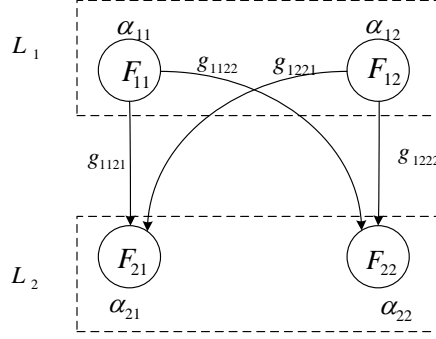


Figure 4: Graphical representation of the two-layer family-based procedure with  $m = 4$  families of hypotheses.

*Inputs.*

- Layers:  $L_1 = \{F_{11}, F_{12}\}$  and  $L_2 = \{F_{21}, F_{22}\}$ .
- Initial significance levels:  $\alpha_{1j}$  for  $F_{1j}$  ( $j = 1, 2$ ) and  $\alpha_{2k}$  for  $F_{2k}$  ( $k = 1, 2$ ).
- Transition coefficients:  $g_{1j,2k} \in [0, 1]$  with  $\sum_{k=1}^2 g_{1j,2k} \leq 1$  for each  $j$ .
- Local FWER-controlling procedures for each family, with error-rate bounds  $e_{1j}^*(\cdot)$ .

*Output.*

- The set of rejected hypotheses across  $L_1$  and  $L_2$ .

*Steps.*

**(S1) First layer.** For each  $j \in \{1, 2\}$  (in any order):

- Test  $F_{1j}$  at level  $\alpha_{1j}$  using an FWER-controlling procedure; let  $A_{1j}$  denote the accepted set.
- Compute the unused portion

$$u_{1j} = \alpha_{1j} - e_{1j}^*(A_{1j}).$$

(c) Update the second-layer levels: for each  $k \in \{1, 2\}$ , set

$$\alpha_{2k} \leftarrow \alpha_{2k} + u_{1j} g_{1j,2k}.$$

(d) Remove outgoing edges from  $F_{1j}$  (no further transfers from  $F_{1j}$ ).

**(S2) Second layer.** For each  $k \in \{1, 2\}$  (in any order):

(a) Test  $F_{2k}$  at level  $\alpha_{2k}$  using an FWER-controlling procedure; record the rejections.

Informally, the procedure proceeds in two stages. The families in layer 1 are tested first at their assigned levels. For each family, the portion of its significance level that is not used is redistributed to the layer 2 families according to the transition coefficients. Once both families in layer 1 have been tested, the families in layer 2 are tested at their updated levels. The overall process is illustrated in Figure 4.

*Remark 2.1.* Since each family is tested only once in the proposed single-pass algorithm, outgoing edges are removed after testing. This may introduce some conservativeness compared with iterative recycling or re-testing schemes.

**Theorem 2.1.** *Suppose Conditions 1.1–1.5 hold. Then the two-layer family-based graphical procedure described in Algorithm 2.1 strongly controls the overall FWER at level  $\alpha$ .*

A proof of Theorem 2.1 is provided in Appendix A.1.

—

## 2.2 General multi-layer family-based graphical approach

The two-layer case illustrates the sequential nature of the family-based graphical procedure. We now generalize this framework to  $n$  layers, each containing an arbitrary number of families.

### Algorithm 2 (Multi-layer family-based procedure)

*Inputs.*

- Layers:  $L_i = \{F_{i1}, \dots, F_{il_i}\}$  for  $i = 1, \dots, n$ , where each layer  $L_i$  contains  $l_i$  families.
- Initial significance levels:  $\alpha_{ij}$  for family  $F_{ij}$ , satisfying  $\sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha_i$  for each  $i$ , and  $\sum_{i=1}^n \alpha_i \leq \alpha$ .
- Transition coefficients:  $g_{ij,kl} \in [0, 1]$  for  $i < k$ , subject to  $\sum_{k=i+1}^n \sum_{l=1}^{l_k} g_{ij,kl} \leq 1$  for each  $F_{ij}$ .
- Local FWER-controlling procedures for each family, with error-rate bounds  $e_{ij}^*(\cdot)$ .

*Output.*

- The set of rejected hypotheses across all layers  $L_1, \dots, L_n$ .

*Steps.*

**(S1) Layer  $i$**  ( $1 \leq i \leq n-1$ ). For each  $j \in \{1, \dots, l_i\}$  (in any order):

- Test  $F_{ij}$  at level  $\alpha_{ij}$  using an FWER-controlling procedure; let  $A_{ij}$  denote the accepted set.
- Compute the unused portion

$$u_{ij} = \alpha_{ij} - e_{ij}^*(A_{ij}).$$

- Update subsequent layers: for each  $k > i$  and each  $l \in \{1, \dots, l_k\}$ , set

$$\alpha_{kl} \leftarrow \alpha_{kl} + u_{ij} g_{ij,kl}.$$

- Remove outgoing edges from  $F_{ij}$  (no further transfers from  $F_{ij}$ ).

**(S2) Final layer  $L_n$** . For each  $j \in \{1, \dots, l_n\}$  (in any order):

- Test  $F_{nj}$  at level  $\alpha_{nj}$  using an FWER-controlling procedure; record the rejections.

Informally, the multi-layer algorithm applies the same logic as the two-layer case but extends it sequentially across  $n$  layers. At each step, families in the current layer are tested at their assigned levels. For each family, the portion of its significance level not used is redistributed to families in

subsequent layers according to the transition coefficients. Once all families in a layer have been tested, the procedure advances to the next layer. This process continues until the final layer is reached, at which point all remaining families are tested at their updated levels.

**Theorem 2.2.** *Suppose Conditions 1.1–1.5 hold. Then the multi-layer family-based graphical procedure described in Algorithm 2.2 strongly controls the overall FWER at level  $\alpha$ .*

A proof of Theorem 2.2 is provided in Appendix A.2.

—

*Remark 2.2.* Consider the special case of  $n$  hierarchically ordered families, with each layer  $L_i$  containing a single family  $F_{i1}$ . The multi-layer family-based graphical approach specializes as follows:

- The initial significance level is  $\alpha_{11} = \alpha$ , and  $\alpha_{i1} = 0$  for  $i > 1$ .
- Transition coefficients are  $g_{i1, (i+1)1} = 1$  for  $i = 1, \dots, n - 1$ , and zero otherwise.

Several implications follow:

1. If each local procedure controls the FWER and satisfies the *separability condition* (i.e., the error rate function is strictly less than  $\alpha$  unless all hypotheses in the family are rejected), then the resulting level-updating mechanism in the multi-layer family-based framework coincides with that of a parallel gatekeeping strategy, and is equivalent to the multistage gatekeeping procedure of Dmitrienko et al. (2008). The separability condition is a property of the local testing procedures rather than of a specific gatekeeping structure. Examples of procedures satisfying this condition include the Bonferroni, truncated Holm, and truncated fallback procedures.
2. If each family uses a local procedure with error rate function upper bound  $e^*(I) = \alpha$  for any non-empty set  $I$ , then the multi-layer approach reduces to a serial gatekeeping strategy. Examples include the Holm and fixed-sequence procedures.
3. If each family contains exactly one hypothesis, the procedure reduces to the conventional fixed-sequence method.

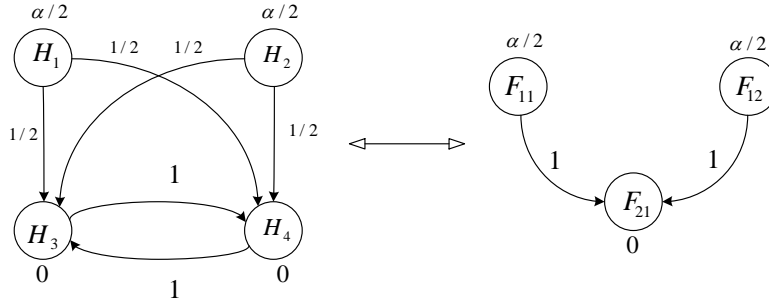


Figure 5: Graphical visualizations of Case 1: Hypothesis-based (left) and family-based (right) approaches.

4. When prior dependence information among  $p$ -values is available, additional local procedures can be used. For instance, if the null  $p$ -values are independent or positively dependent, one may use the Hochberg or truncated Hochberg procedures.

### 3 Case Studies

In this section, we demonstrate the efficiency and simplicity of our proposed family-based graphical approach by comparing it with the conventional hypothesis-based graphical approach in the context of testing multiple families of hypotheses. Three illustrative examples from [Bretz et al. \(2009\)](#) are used for this comparison. Figures 5–7 display the original hypothesis-based graphs (left panels) alongside the corresponding family-based graphs (right panels).

**Case 3.1.** Figure 5 involves four null hypotheses:  $H_1, H_2, H_3, H_4$ . The hypothesis-based procedure is shown on the left, while the equivalent family-based procedure is shown on the right. In the family-based framework, the hypotheses are grouped into  $m = 3$  families:  $F_{11} = \{H_1\}$ ,  $F_{12} = \{H_2\}$ , and  $F_{21} = \{H_3, H_4\}$ , organized across  $n = 2$  layers:  $L_1 = \{F_{11}, F_{12}\}$  and  $L_2 = \{F_{21}\}$ . Initial significance levels are  $\alpha/2$  for  $F_{11}$  and  $F_{12}$ , and 0 for  $F_{21}$ . The transition coefficient set  $\mathbf{G}$  is defined as

$$g_{1121} = g_{1221} = 1, \quad g_{2111} = g_{2112} = g_{1112} = g_{1211} = 0.$$

The procedure begins by testing  $F_{11}$  at level  $\alpha/2$  using the Bonferroni method. If  $H_1$  is rejected, its significance level is transferred to  $F_{21}$ , raising its level from  $\alpha_{21} = 0$  to  $\alpha_{21}^* = \alpha/2$ . Otherwise,

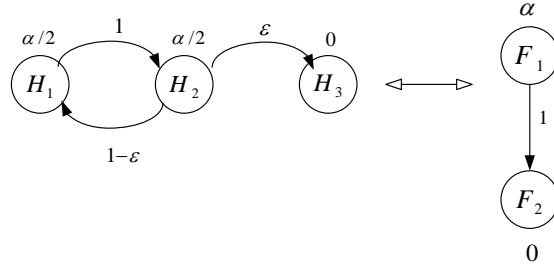


Figure 6: Graphical visualizations of Case 2: Hypothesis-based (left) and family-based (right) approaches.

no transfer occurs. Next,  $F_{12}$  is tested at level  $\alpha/2$ ; if  $H_2$  is rejected,  $\alpha/2$  is added to  $\alpha_{21}^*$ . After testing both families in  $L_1$ ,  $F_{21}$  is tested (if  $\alpha_{21}^* \neq 0$ ) using the Holm procedure at level  $\alpha_{21}^*$ . This procedure is equivalent to the hypothesis-based method, but as shown in the right panel, the family-based representation more clearly conveys the hierarchical structure among families of hypotheses.  $\square$

In some scenarios, hypotheses within one family are testable only if all hypotheses in another family are rejected. Using the hypothesis-based graphical approach to represent such strategies often leads to edges with infinitesimal weights, making the graphs unnecessarily complex and difficult for non-specialists to interpret. The family-based approach avoids this complication, as shown below.

**Case 3.2.** Consider a gatekeeping strategy with three hypotheses:  $H_1$ ,  $H_2$ , and  $H_3$ , where  $H_3$  can only be tested if both  $H_1$  and  $H_2$  are rejected. The hypothesis-based graph (Figure 6, left) uses an edge with an infinitesimally small weight  $\epsilon$ . In contrast, the family-based graph (Figure 6, right) eliminates this edge. The strategy reduces to a two-layer, two-family procedure:  $L_1 = \{F_1\}$  and  $L_2 = \{F_2\}$ , where  $F_1 = \{H_1, H_2\}$  and  $F_2 = \{H_3\}$ . Initial significance levels are  $\alpha$  for  $F_1$  and 0 for  $F_2$ . The gatekeeping procedure is:

1. Test  $F_1$  using the Holm procedure at level  $\alpha$ .
2. If both  $H_1$  and  $H_2$  are rejected, transfer  $\alpha$  to  $F_2$  and test it at level  $\alpha$ .
3. Otherwise, stop the procedure.

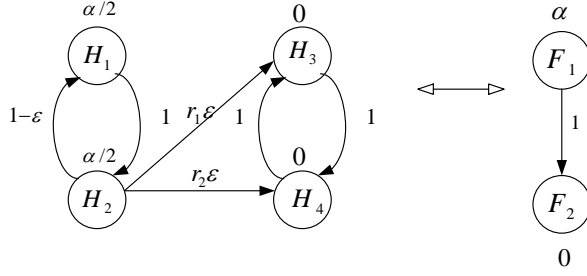


Figure 7: Graphical visualizations of Case 3: Hypothesis-based (left) and family-based (right) approaches.

**Case 3.3.** Now consider a more complex gatekeeping strategy with four hypotheses:  $H_1, H_2, H_3, H_4$ , where  $H_3$  and  $H_4$  can only be tested if both  $H_1$  and  $H_2$  are rejected. The hypothesis-based graph (Figure 7, left) again requires edges with infinitesimally small weights. Specifically, when  $H_1$  and  $H_2$  are both rejected,  $\alpha$  is split between  $H_3$  and  $H_4$  according to weights  $r_1$  and  $r_2$ , so that  $H_3$  receives  $r_1\alpha$  and  $H_4$  receives  $r_2\alpha$ .

In the family-based approach (Figure 7, right), the structure simplifies to two layers with two families:  $L_1 = \{F_1\}$  and  $L_2 = \{F_2\}$ , where  $F_1 = \{H_1, H_2\}$  and  $F_2 = \{H_3, H_4\}$ . Initial significance levels are  $\alpha$  for  $F_1$  and 0 for  $F_2$ . The procedure is:

1. Test  $F_1$  with the Holm procedure at level  $\alpha$ .
2. If both  $H_1$  and  $H_2$  are rejected, transfer  $\alpha$  to  $F_2$  and test it at level  $\alpha$  using a weighted Holm procedure with weights  $r_1$  and  $r_2$ .
3. If either  $H_1$  or  $H_2$  is not rejected, stop the procedure.

*Remark 3.1.* Across the three examples, our family-based graphical approach consistently simplifies the representation and implementation of complex testing strategies. Unlike the hypothesis-based approach—which often requires non-intuitive infinitesimal edge weights ( $\epsilon$ )—the family-based approach offers a clearer and more intuitive framework, facilitating interpretation and communication, especially for non-statisticians.

## 4 A Clinical Trial Example

In this section, we illustrate the application of the proposed family-based graphical approach in a clinical trial setting and compare its performance with the conventional hypothesis-based graphical approach.

We revisit the Type II diabetes clinical trial example from [Dmitrienko et al. \(2007\)](#). The trial evaluates three doses of an experimental drug (Low [L], Medium [M], and High [H]) versus placebo (Plac) across one primary endpoint (P: Hemoglobin A1c) and two secondary endpoints (S1: Fasting serum glucose; S2: HDL cholesterol). Each endpoint is assessed for all three doses, yielding nine null hypotheses, which are grouped into three families:

- $F_1$ : dose–placebo comparisons for the primary endpoint  $P$ , i.e., H vs Plac ( $H_{11}$ ), M vs Plac ( $H_{12}$ ), and L vs Plac ( $H_{13}$ );
- $F_2$ : dose–placebo comparisons for the secondary endpoint  $S1$ , i.e., H vs Plac ( $H_{21}$ ), M vs Plac ( $H_{22}$ ), and L vs Plac ( $H_{23}$ );
- $F_3$ : dose–placebo comparisons for the secondary endpoint  $S2$ , i.e., H vs Plac ( $H_{31}$ ), M vs Plac ( $H_{32}$ ), and L vs Plac ( $H_{33}$ ).

This structure highlights how the family-based graphical approach naturally organizes grouped and hierarchically ordered hypotheses.

The overall Type I error rate is controlled at  $\alpha = 0.05$ , and the raw  $p$ -values for the nine hypotheses are reported in [Table 1](#). Since the primary endpoint ( $P$ ) is considered most critical,  $F_1$  is always tested first, followed by  $F_2$  and  $F_3$ . For the secondary endpoints, we explore two types of hierarchical relationships, leading to two gatekeeping strategies: Procedure 1 and Procedure 2. Both are visualized using the family-based approach, with Procedure 1 also compared to its hypothesis-based counterpart.

**Procedure 1.** Suppose  $S1$  and  $S2$  are equally important. Then  $F_2$  and  $F_3$  are placed in the same layer, with hypotheses within each family tested in a pre-specified order (H vs Plac, M vs Plac, L vs Plac). The fixed-sequence procedure serves as the local test within each family. Initial

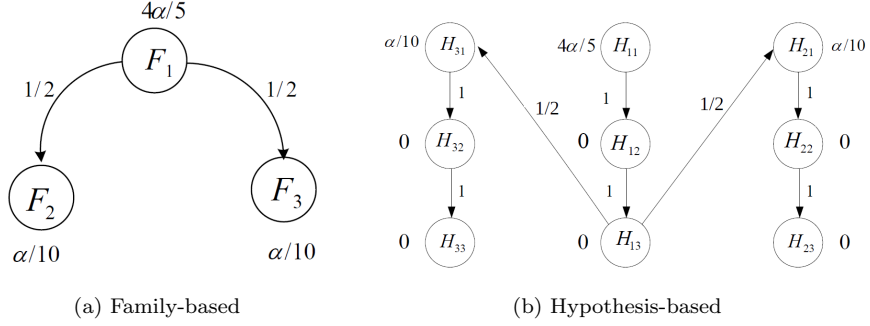


Figure 8: Family-based (a) and hypothesis-based (b) graphical representations of Procedure 1 for the Type II diabetes clinical trial.

significance levels are allocated as 0.04 for  $F_1$ , 0.005 for  $F_2$ , and 0.005 for  $F_3$ . Once  $F_1$  is tested, any significance level it retains is equally split between  $F_2$  and  $F_3$ .

The testing proceeds as follows. At level 0.04, all three hypotheses in  $F_1$  are rejected. Its significance level is then equally divided, updating the levels for  $F_2$  and  $F_3$  to 0.025 each. Testing continues for  $F_2$  and  $F_3$  (order irrelevant), resulting in the rejection of  $H_{21}$ ,  $H_{31}$ , and  $H_{32}$ . Results for Procedure 1 are summarized in Table 1. Figure 8(a) shows the corresponding family-based graph, while Figure 8(b) displays the conventional hypothesis-based graph. The comparison illustrates that the family-based representation more transparently conveys the hierarchical structure.

**Procedure 2.** Suppose  $S_1$  is considered more important than  $S_2$ , yielding a strict hierarchical ordering:  $F_1 \rightarrow F_2 \rightarrow F_3$ . The gatekeeping strategy is shown in Figure 3(b). The truncated Hochberg procedure with truncation parameter  $\gamma = 0.6$  is used for  $F_1$  and  $F_2$ , while the conventional Hochberg procedure is applied to  $F_3$ . Initial significance levels are 0.04 for  $F_1$ , 0.005 for  $F_2$ , and 0.005 for  $F_3$ .

The procedure begins with  $F_1$  at level 0.04, where all hypotheses are rejected. Its significance level is then redistributed:  $0.04 \times 0.8 + 0.005 = 0.037$  to  $F_2$ , and  $0.04 \times 0.2 + 0.005 = 0.013$  to  $F_3$ . Next,  $F_2$  is tested at level 0.037; all hypotheses are rejected, and its entire significance level is transferred to  $F_3$ , increasing its level to 0.05. Finally,  $F_3$  is tested at level 0.05, leading to the rejection of  $H_{31}$  and  $H_{32}$ . Results are reported in Table 1. Notably, this three-layer procedure cannot be conveniently visualized using the conventional hypothesis-based approach, underscoring the added flexibility of the family-based framework.

Table 1: Comparison of results for 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  $\alpha = 0.05$ . 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

*Remark 4.1.* This clinical trial example illustrates that the family-based graphical approach can provide a concise and structured representation of the testing strategy at the family level. Such a representation may be helpful for describing hierarchical relationships among families of hypotheses and for communicating the overall testing strategy.

We note, however, that this does not imply greater power or a broader class of representable procedures than hypothesis-based graphical approaches. In particular, hypothesis-based graphical approaches, possibly combined with local procedures such as Simes or Hochberg (Xi and Bretz, 2019), may yield more powerful tests in some settings. The main advantage of the family-based framework is therefore its interpretability and convenience as a family-level representation.

## 5 Simulation Study

To evaluate the finite-sample performance of the proposed family-based graphical procedure, we conducted a simulation study comparing it with a hypothesis-based graphical procedure and a superchain procedure. The main objective is to assess whether the family-level representation introduces additional conservativeness relative to hypothesis-level procedures while maintaining

strong control of the familywise error rate (FWER).

We considered two layers of hypothesis families,

$$F_1 = \{H_{11}, H_{12}\}, \quad F_2 = \{H_{21}, H_{22}\},$$

with total significance level  $\alpha = 0.05$ . Throughout, the initial family weights were set to

$$w_1 = 1, \quad w_2 = 0,$$

so that the initial local significance levels for the two families were

$$\alpha_{11} = w_1\alpha = \alpha, \quad \alpha_{21} = w_2\alpha = 0.$$

### Compared procedures.

- **Proposed two-layer family-based graphical procedure.** The first family  $F_1$  was tested using the truncated Holm procedure with truncation fraction  $\gamma \in [0, 1]$ , while the second family  $F_2$  was tested using the standard Holm procedure. The procedure operates at the family level. First,  $F_1$  is tested at level  $\alpha$ . Let  $A_1$  denote the set of hypotheses in  $F_1$  that remain accepted after this step. The unused local significance level is then computed through the error-rate bound of the truncated Holm procedure,

$$e_1^*(A_1) = \begin{cases} 0, & |A_1| = 0, \\ \left[ \gamma + (1 - \gamma) \frac{|A_1|}{|F_1|} \right] \alpha, & |A_1| > 0, \end{cases}$$

and the unused amount

$$u_1 = \alpha - e_1^*(A_1)$$

is transferred to  $F_2$ . Since the edge weight from  $F_1$  to  $F_2$  was set to 1 in this study, the updated level for  $F_2$  becomes

$$\alpha_2^{\text{upd}} = u_1.$$

The family  $F_2$  is then tested using the standard Holm procedure at level  $\alpha_2^{\text{upd}}$ .

- **Hypothesis-based graphical procedure.** This procedure is illustrated in Figure 1. Each hypothesis is represented as a node in a graphical multiple testing framework, with initial local significance levels

$$\left( \frac{\alpha}{2}, \frac{\alpha}{2}, 0, 0 \right)$$

assigned to  $(H_{11}, H_{12}, H_{21}, H_{22})$ , respectively. Thus, the hypotheses in  $F_1$  each receive level  $\alpha/2$ , while those in  $F_2$  receive zero initial weight.

The transition weights reflect the family structure: each of  $H_{11}$  and  $H_{12}$  transfers a proportion  $\gamma$  of its local level to the other hypothesis in  $F_1$ , and distributes the remaining proportion  $1 - \gamma$  equally between  $H_{21}$  and  $H_{22}$ . Within  $F_2$ , the two hypotheses are connected by unit transition weights, corresponding to the standard Holm-type updating mechanism.

Testing proceeds sequentially according to the standard graphical multiple testing algorithm, with local significance levels updated after each rejection. In this setting, the initial configuration is equivalent to testing the hypotheses in  $F_1$  using the truncated Holm procedure at level  $\alpha$ , while assigning zero initial significance level to the hypotheses in  $F_2$ . As level propagates to  $F_2$ , testing within  $F_2$  follows the standard Holm procedure, aligning the simulation design with the other procedures.

- **Superchain procedure.** We consider a two-family superchain procedure with  $w_1 = 1$  and  $w_2 = 0$  to ensure comparability with the other procedures. Under this specification, all initial significance level is assigned to  $F_1$ , so that testing starts from the first family, while  $F_2$  receives significance level only through subsequent updating.

In Step 1,  $F_1$  is tested at level  $\alpha$  using the truncated Holm procedure with truncation parameter  $\gamma$ , whereas  $F_2$  is not tested initially. If both families still contain non-rejected hypotheses after Step 1, they are retested in subsequent steps at updated levels  $\alpha_{1k} = \kappa_{1k}\alpha$  and  $\alpha_{2k} = \kappa_{2k}\alpha$ , where

$$\kappa_{1k} = 1, \quad \kappa_{2k} = 1 - f_1(A_{1,k-1} \mid \gamma),$$

and  $A_{1,k-1}$  denotes the set of non-rejected hypotheses in  $F_1$  after Step  $k - 1$ . Thus, partial rejection in  $F_1$  leads to partial transfer of significance level to  $F_2$ , whereas complete rejection of  $F_1$  allows  $F_2$  to be tested at the full level  $\alpha$ . Throughout,  $F_1$  is tested using the truncated Holm procedure, while  $F_2$  is tested using the standard Holm procedure.

If only one family still contains non-rejected hypotheses, that family is retested at level  $\alpha$ , and the algorithm terminates. This specification retains the superchain updating structure while aligning the directional flow of significance level with that of the proposed family-based and hypothesis-based graphical procedures.

**Data-generating mechanism.** For simplicity, we generated independent one-sided  $z$ -tests across

hypotheses. Specifically, for each hypothesis we generated

$$Z_{ij} \sim N(\mu_{ij}, 1),$$

and computed the corresponding one-sided  $p$ -value as

$$p_{ij} = 1 - \Phi(Z_{ij}).$$

A true null hypothesis corresponds to  $\mu_{ij} = 0$ , while a false null corresponds to  $\mu_{ij} = \Delta$ . In the simulations reported below, we set  $\Delta = 2.2$ ,  $\gamma = 0.5$ , and  $\alpha = 0.05$ .

**Simulation scenarios.** We considered the following four configurations:

- **Scenario 1:** all four hypotheses are true;
- **Scenario 2:** only  $H_{11}$  is false;
- **Scenario 3:**  $H_{11}$  and  $H_{21}$  are false;
- **Scenario 4:**  $H_{11}$ ,  $H_{12}$ , and  $H_{21}$  are false, while  $H_{22}$  is true.

This setup allows us to examine performance under varying signal sparsity and across-family signal configurations.

**Performance measures.** For each scenario, we generated 10,000 Monte Carlo replicates. For each procedure, we recorded:

- the familywise error rate (FWER), defined as the probability of rejecting at least one true null hypothesis;
- the average power, defined as the expected proportion of false null hypotheses that are correctly rejected, that is,

$$\text{Average Power} = \mathbb{E} \left( \frac{\#\{\text{correctly rejected false nulls}\}}{\#\{\text{false nulls}\}} \right).$$

When all hypotheses are true, power is not applicable and is therefore omitted.

Table 2: Simulation results under different scenarios.

Scenario 1		
Procedure	FWER	Power
Proposed family-based graph	0.04997	—
Hypothesis-based graph	0.04997	—
Superchain	0.04997	—
Scenario 2		
Procedure	FWER	Power
Proposed family-based graph	0.03951	0.5949
Hypothesis-based graph	0.03951	0.5949
Superchain	0.04704	0.5960
Scenario 3		
Procedure	FWER	Power
Proposed family-based graph	0.03678	0.416860
Hypothesis-based graph	0.03678	0.416860
Superchain	0.04386	0.418355
Scenario 4		
Procedure	FWER	Power
Proposed family-based graph	0.0208	0.56336
Hypothesis-based graph	0.0208	0.56336
Superchain	0.0226	0.58667

Table 2 reports the results. All three procedures approximately control the FWER at the nominal level across all scenarios.

In Scenarios 2–4, the proposed family-based graphical procedure and the hypothesis-based graphical procedure yield identical performance in terms of both FWER and power, indicating that, in this setting, the family-level representation does not introduce additional conservativeness relative to the hypothesis-based graphical approach.

The superchain procedure exhibits slightly higher FWER and consistently achieves marginally higher power, particularly in Scenario 4 where most hypotheses are false. This gain in power reflects the effect of iterative updating and retesting.

The observed pattern of average power reflects a tradeoff between multiplicity and signal strength: as the number of false hypotheses increases, power may initially decrease due to the increased multiplicity burden, and then increase as the overall signal becomes stronger.

Overall, the proposed family-based graphical procedure achieves performance comparable to the hypothesis-based graphical approach, while offering a simpler and more interpretable family-level representation, with only minor differences relative to the superchain procedure.

We also considered the improved parallel gatekeeping procedure presented in Figure 12 of [Bretz et al. \(2009\)](#). Additional simulations (not reported) indicate that its performance is essentially indistinguishable from that of the hypothesis-based graphical approach included here, with negligible differences in both familywise error rate control and power. Accordingly, it is omitted from the simulation tables.

The code used to generate all simulation results will be made publicly available upon publication of the manuscript.

## 6 Conclusions

We introduce a family-based graphical framework for testing hierarchically ordered families of hypotheses. The proposed approach provides strong control of the familywise error rate (FWER)

at a pre-specified level, while offering a transparent and systematic way to construct and visualize a broad class of gatekeeping strategies. As a special case, when each layer contains a single family, the framework recovers the general multistage gatekeeping procedures of [Dmitrienko et al. \(2008\)](#).

The primary contribution of this work lies in providing a unifying family-based graphical framework that enables transparent construction and interpretation of hierarchical testing strategies, rather than introducing a fundamentally new testing algorithm.

Through illustrative examples, simulation studies, and a real clinical trial application, we demonstrate that the proposed framework is often simpler and more interpretable than hypothesis-level graphical approaches (e.g., [Bretz et al. \(2009\)](#)) in settings involving multiple hierarchically ordered families. The simulation results indicate that the proposed procedure achieves performance comparable to hypothesis-level graphical procedures, while remaining competitive with the super-chain procedure. These findings suggest that any potential conservativeness due to family-level aggregation is scenario-dependent and remains limited in practice.

By operating at the family level, the framework aligns naturally with clinical objective structures (such as primary and key secondary endpoints), leading to clearer visualization and improved communication of complex testing strategies, particularly for non-statistical stakeholders. Overall, the proposed framework provides a practical balance between statistical rigor and interpretability, making it well suited for complex hierarchical testing problems in modern clinical trials.

Conceptually, the proposed framework is related to the superchain procedures of [Kordzakhia and Dmitrienko \(2013\)](#), which allow iterative re-testing and flexible propagation of significance levels. In contrast, our approach adopts a sequential multi-layer design with fixed transition coefficients and no iterative updating. This restriction yields a more transparent and regulator-friendly procedure, while maintaining rigorous error control and avoiding the additional complexity of more general recycling schemes.

Like any methodological framework, the proposed approach has certain limitations. In studies with many layers or a large number of families, the construction and specification of family-level graphs may become cumbersome. In addition, practical implementation in large-scale trials may require careful coordination between design-level specification and software implementation.

Another limitation is that the proposed framework relies on the availability of an error-rate

function, or a valid upper bound, for the local multiple testing procedure within each family. These local error-rate-function bounds are essential for determining how significance levels can be redistributed across families while maintaining global error control. Consequently, the approach is most naturally applicable when the within-family procedures admit explicit or easily computable bounds, such as Bonferroni or truncated Holm procedures. When such bounds are not available, implementation may be more challenging, and hypothesis-based graphical approaches may offer greater flexibility in such settings.

**Relation to hypothesis-level graphical tools.** Hypothesis-based graphical procedures and associated software tools (e.g., gMCP and related implementations) provide flexible and widely used solutions for multiple testing in clinical trials. The proposed framework is complementary rather than competitive with these approaches. While hypothesis-level graphs offer fine-grained control, they can become difficult to construct and interpret when numerous families, layers, and within-family procedures are involved.

The family-based graphical framework operates at a higher structural level, encoding family-wise logical relationships directly and integrating explicit updating rules through error-rate-function bounds. As a result, once the family structure, transition rules, and local procedures are specified, the framework defines a well-posed and FWER-valid testing procedure, rather than serving as a purely conceptual diagram. In practice, the family-based design can be implemented using standard within-family procedures and, if desired, translated into a hypothesis-level graph for software execution. Thus, the framework provides a scalable and transparent design structure while remaining compatible with existing implementation tools.

Future work may explore extensions to adaptive designs, more complex dependence structures, and broader regulatory applications, further enhancing the applicability of the family-based graphical approach in modern confirmatory trials.

## Acknowledgments

The authors thank the Editor, the Associate Editor, and the reviewers for their careful evaluation of the manuscript and for their insightful and constructive comments, which have significantly

improved the quality and clarity of the paper.

## Conflict of Interest

The authors declare that there are no competing interests.

## Appendix

### A.1 Proof of Theorem 2.1

We establish strong control of the FWER by partitioning the overall error event into two components: (i) the contribution from the first-layer families, and (ii) the contribution from the second-layer families, conditional on no error occurring in the first layer.

Suppose that family  $F_{ij}$  is tested at level  $\alpha_{ij}^*$ . Then

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

for  $i, j = 1, 2$ , where  $e_{1j}^*$  denotes the upper bound of the error rate function for  $F_{1j}$  and  $A_{1j}$  is the set of accepted hypotheses in  $F_{1j}$ .

Let  $E_{ij}(x)$  denote the event that at least one true null in  $F_{ij}$  is rejected at level  $x$ , and let  $\bar{E}_{ij}(x)$  denote its complement. Then the overall FWER can be decomposed as

$$\begin{aligned}\text{FWER} &= \Pr \left\{ \bigcup_{i=1}^2 \bigcup_{j=1}^2 E_{ij}(\alpha_{ij}^*) \right\} \\ &= \Pr \left\{ \bigcup_{j=1}^2 E_{1j}(\alpha_{1j}^*) \right\} + \Pr \left\{ \left( \bigcap_{j=1}^2 \bar{E}_{1j}(\alpha_{1j}^*) \right) \cap \left( \bigcup_{j=1}^2 E_{2j}(\alpha_{2j}^*) \right) \right\}.\end{aligned}\tag{3}$$

—

**Step 1: Bound for the first term.** For the first term in (3), the Bonferroni inequality yields

$$\begin{aligned} \Pr\left\{\bigcup_{j=1}^2 E_{1j}(\alpha_{1j}^*)\right\} &\leq \sum_{j=1}^2 \Pr\{E_{1j}(\alpha_{1j}^*)\} \\ &\leq \sum_{j=1}^2 e_{1j}^*(T_{1j}), \end{aligned} \quad (4)$$

where the second inequality follows from the definition of the error rate function. Note that  $e_{1j}^*(T_{1j})$  depends on  $\alpha_{1j}^*$ , since it is the error rate function of the specified multiple testing procedure at level  $\alpha_{1j}^*$ .

—

**Step 2: Bound for the second term.** Consider the second term of (3). If  $\bigcap_{j=1}^2 \bar{E}_{1j}(\alpha_{1j}^*)$  holds, then  $T_{1j} \subseteq A_{1j}$ , implying by Condition 1.4 that

$$e_{1j}^*(T_{1j}) \leq e_{1j}^*(A_{1j}), \quad j = 1, 2.$$

From (2), it follows that

$$\begin{aligned} \alpha_{2i}^* &= \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j}^* - e_{1j}^*(A_{1j})) g_{1j2i} \\ &\leq \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i}. \end{aligned} \quad (5)$$

Hence, by Condition 1.5,

$$\left(\bigcap_{j=1}^2 \bar{E}_{1j}(\alpha_{1j}^*)\right) \cap \left(\bigcup_{j=1}^2 E_{2j}(\alpha_{2j}^*)\right) \subseteq \bigcup_{i=1}^2 E_{2i} \left( \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i} \right).$$

Applying the Bonferroni inequality again gives

$$\begin{aligned} &\Pr\left\{\left(\bigcap_{j=1}^2 \bar{E}_{1j}(\alpha_{1j}^*)\right) \cap \left(\bigcup_{j=1}^2 E_{2j}(\alpha_{2j}^*)\right)\right\} \\ &\leq \sum_{i=1}^2 \Pr\left\{E_{2i} \left( \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i} \right)\right\}. \end{aligned} \quad (6)$$

By Condition 1.3, each  $F_{2i}$  is tested using an FWER-controlling local procedure. Thus, the probability inside the sum is bounded by its level, so the right-hand side of (6) is at most

$$\sum_{i=1}^2 \left( \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j}^* - e_{1j}^*(T_{1j})) g_{1j2i} \right). \quad (7)$$

Rearranging gives

$$\begin{aligned}
&= \sum_{i=1}^2 \alpha_{2i} + \sum_{j=1}^2 (\alpha_{1j} - e_{1j}^*(T_{1j})) \sum_{i=1}^2 g_{1j2i} \\
&\leq \sum_{i=1}^2 \alpha_{2i} + \sum_{j=1}^2 \alpha_{1j} - \sum_{j=1}^2 e_{1j}^*(T_{1j}) \\
&\leq \alpha - \sum_{j=1}^2 e_{1j}^*(T_{1j}), \tag{8}
\end{aligned}$$

where the first inequality follows from  $\sum_{i=1}^2 g_{1j2i} \leq 1$ .

—

**Step 3: Combine bounds.** Combining (4) and (8) with (3), we obtain

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

Thus, the two-layer family-based graphical procedure achieves strong control of the FWER at level  $\alpha$ .  $\square$

## A.2 Proof of Theorem 2.2

We prove the result by induction on the number of layers.

Let  $\text{FWER}_n(\alpha_1, \dots, \alpha_n)$  denote the overall familywise error rate (FWER) of the multi-layer family-based procedure when the initial significance levels assigned to the layers  $L_i$  are  $\alpha_i$  for  $i = 1, \dots, n$ . Within each layer  $L_i$ , the families  $F_{ij}$  receive initial significance levels  $\alpha_{ij}$  for  $j = 1, \dots, l_i$ , subject to Condition 1.1. We aim to show that

$$\text{FWER}_n(\alpha_1, \dots, \alpha_n) \leq \sum_{i=1}^n \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha. \tag{9}$$

**Base case ( $n = 2$ ).** By Theorem 2.1 (proved under Conditions 1.1–1.5),

$$\text{FWER}_2(\alpha_1, \alpha_2) \leq \sum_{i=1}^2 \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha.$$

**Induction step.** Assume (9) holds for  $n = k$  ( $k \geq 2$ ), i.e.,

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

We show it holds for  $n = k + 1$ :

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

**Step 1: Decomposition.** Define

$$B_1 = \{\text{at least one true null is rejected in } L_1\}, \quad B_2 = \{\text{at least one true null is rejected in } L_2, \dots, L_{k+1}\}.$$

Then

$$\text{FWER}_{k+1}(\alpha_1, \dots, \alpha_{k+1}) = \Pr\{B_1\} + \Pr\{\bar{B}_1 \cap B_2\}. \quad (10)$$

By the definition of the error-rate function and Bonferroni,

$$\Pr\{B_1\} \leq \sum_{j=1}^{l_1} e_{1j}^*(T_{1j}), \quad (11)$$

where  $T_{1j}$  is the set of true nulls in  $F_{1j}$ .

**Step 2: Bounding  $\Pr\{\bar{B}_1 \cap B_2\}$ .** After testing all families in  $L_1$ , the unused portions  $\sum_{j=1}^{l_1} (\alpha_{1j} - e_{1j}^*(A_{1j}))$  are redistributed forward according to later layers. For any  $F_{ij}$  in  $L_i$  ( $i \geq 2$ ),

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

where  $A_{1l}$  is the accepted set in  $F_{1l}$ . Let  $\alpha_i^* = \sum_{j=1}^{l_i} \alpha_{ij}^*$ .

On  $\bar{B}_1$  no true null is rejected in  $L_1$ , so errors can only occur in  $L_2, \dots, L_{k+1}$ :

$$\Pr\{\bar{B}_1 \cap B_2\} = \text{FWER}_k(\alpha_2^*, \dots, \alpha_{k+1}^*). \quad (12)$$

Moreover,  $\bar{B}_1$  implies  $T_{1j} \subseteq A_{1j}$ ; by Condition 1.4,

$$e_{1j}^*(T_{1j}) \leq e_{1j}^*(A_{1j}), \quad j = 1, \dots, l_1.$$

Applying the induction hypothesis to (12) yields

$$\begin{aligned}
\text{FWER}_k(\alpha_2^*, \dots, \alpha_{k+1}^*) &\leq \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij}^* \\
&= \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} + \sum_{l=1}^{l_1} \alpha_{1l} \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} g_{1lij} - \sum_{l=1}^{l_1} e_{1l}^*(A_{1l}) \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} g_{1lij} \\
&\leq \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} + \sum_{l=1}^{l_1} \alpha_{1l} - \sum_{l=1}^{l_1} e_{1l}^*(A_{1l}) \quad (\text{by Condition 1.2}) \\
&\leq \sum_{i=1}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} - \sum_{j=1}^{l_1} e_{1j}^*(T_{1j}) \quad (\text{by Condition 1.4}).
\end{aligned}$$

Here, Condition 1.2 ensures  $\sum_{i=2}^{k+1} \sum_{j=1}^{l_i} g_{1lij} \leq 1$  for each  $l$ . (Condition 1.5 guarantees that increases in significance levels cannot decrease rejection sets and underpins the forward redistribution argument.)

**Step 3: Conclusion.** Combining (10), (11), and the bound above gives

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

where the final inequality uses Condition 1.1. This completes the induction and establishes (9) for all  $n \in \mathbb{N}$ .  $\square$

## References

- Mohamed Alosch, Frank Bretz, and Mohammad Huque. Advanced multiplicity adjustment methods in clinical trials. *Statistics in medicine*, 33(4):693–713, 2014.
- P Bauer, J Röhmel, W Maurer, and L Hothorn. Testing strategies in multi-dose experiments including active control. *Statistics in medicine*, 17(18):2133–2146, 1998.
- Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine*, 28(4):586–604, 2009.
- C-F Burman, C Sonesson, and O Guilbaud. A recycling framework for the construction of bonferroni-based multiple tests. *Statistics in medicine*, 28(5):739–761, 2009.

- Alex Dmitrienko and Ajit C Tamhane. Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in medicine*, 30(13):1473–1488, 2011.
- Alex Dmitrienko and Ajit C Tamhane. General theory of mixture procedures for gatekeeping. *Biometrical Journal*, 55(3):402–419, 2013.
- Alex Dmitrienko, Ajit C Tamhane, Xin Wang, and Xun Chen. Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(6):984–991, 2006.
- Alex Dmitrienko, Brian L Wiens, Ajit C Tamhane, and Xin Wang. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in medicine*, 26(12):2465–2478, 2007.
- Alex Dmitrienko, Ajit C Tamhane, and Brian L Wiens. General multistage gatekeeping procedures. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(5):667–677, 2008.
- Alex Dmitrienko, Ajit C Tamhane, and Frank Bretz. *Multiple testing problems in pharmaceutical statistics*. CRC press, 2009.
- Alex Dmitrienko, George Kordzakhia, and Thomas Brechenmacher. Mixture-based gatekeeping procedures for multiplicity problems with multiple sequences of hypotheses. *Journal of Biopharmaceutical Statistics*, 26(4):758–780, 2016.
- Alexei 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(15):2387–2400, 2003.
- Olivier Guilbaud. Bonferroni parallel gatekeeping–transparent generalizations, adjusted p-values, and short direct proofs. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 49(6):917–927, 2007.
- George Kordzakhia and Alex Dmitrienko. Superchain procedures in clinical trials with multiple objectives. *Statistics in Medicine*, 32(3):486–508, 2013.
- George Kordzakhia, Thomas Brechenmacher, Eiji Ishida, Alex Dmitrienko, Winston Wenxiang Zheng, and David Fuyuan Li. An enhanced mixture method for constructing gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics*, 28(1):113–128, 2018.
- Willi Maurer. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypothesis. *Biomed. Chem.-Pharm. Ind.*, 6:3–18, 1995.

- Willi Maurer and Frank Bretz. A note on testing families of hypotheses using graphical procedures. *Statistics in Medicine*, 33(30):5340–5346, 2014.
- Paul M Ridker, Brendan M Everett, Tom Thuren, Jean G MacFadyen, William H Chang, Christie Ballantyne, Francisco Fonseca, Jose Nicolau, Wolfgang Koenig, Stefan D Anker, et al. Anti-inflammatory therapy with canakinumab for atherosclerotic disease. *New England journal of medicine*, 377(12):1119–1131, 2017.
- Ajit C Tamhane and Jiangtao Gou. Advances in p-value based multiple test procedures. *Journal of biopharmaceutical statistics*, 28(1):10–27, 2018.
- Huajun Wang, Ellen Ypma, and Uwe Nicolay. Managing multiplicity in clinical vaccine studies—a case study using a gatekeeping testing strategy. *Vaccine*, 40(16):2462–2469, 2022.
- Peter H Westfall and Alok Krishen. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1):25–40, 2001.
- Dong Xi and Frank Bretz. Symmetric graphs for equally weighted tests, with application to the hochberg procedure. *Statistics in Medicine*, 38(27):5268–5282, 2019.
- Dong Xi and Ajit C Tamhane. A general multistage procedure for k-out-of-n gatekeeping. *Statistics in Medicine*, 33(8):1321–1335, 2014.