Biostatistics

# MULTIPLICITY

IN CLINICAL TRIALS

Yupeng Li

12-15-2021

## TABLE OF FIGURES

# 1 Introduction

Confirmatory controlled clinical trials, also known as Phase III clinical trials, when successful, are significant achievements in medical research as they provide evidence that new treatments (e.g., test drugs or other types of interventions) studied in these trials are clinically effective in treating targeted diseases.

Unfortunately, many such trials fail and are unable to show that new treatments studied in these trials are better than placebo.

There can be several reasons for such failures:

- Certain weaknesses in the primary endpoints of a trial can jeopardize the success of a trial
  - ✓ e.g., if these endpoints are not objective, or are not validated, or are not in line with the mechanisms of actions of the treatment
- Poor planning or disregarding multiplicity issues with respect to multiple endpoints and multiple comparisons

Clinical trials generally pose multiple questions in the form of *hypotheses* whose evaluations involve

- multiple comparisons
- tests for multiple endpoints

## 1.1 Testing A Single Hypothesis

A statistical test in the absence of a treatment effect can lead to a positive conclusion in favor the treatment effect just by chance. Such an error in the testing of hypotheses terminology is known as a **False positive error** or a **Type I error** (will be defined in Section 1.3).

## 1.2 Testing Multiple Hypotheses

When multiple hypotheses are tested without an appropriate adjustment, this error can become excessive. In other words, the **Familywise error rate (FWER)** defined later can become inflated. There is a simple example with two hypotheses demonstrating Type I error inflation.

Assume that we test a single null hypothesis at significance level $\alpha$ = 0.05. If we have two null hypotheses and do two independent tests, each at level $\alpha$ = 0.05:

The probability of rejecting at least one true null hypothesis is

$$\mathrm{P}(\text{reject at least one true null hypothesis}) = 1 - \mathrm{P}(\text{reject neither true null hypothesis})$$

Which is $1 - 0.95^2 = 0.0975\ (> 0.05)$. The type I error rate is almost doubled.

This situation can then lead to an easy approval of an ineffective treatment. In fact, For large $m$ (the number of null hypotheses) we almost surely reject incorrectly at least one null hypothesis.

Therefore, it is important that trials control this error probability at a prespecified level through appropriate design and analyses strategies that are prospectively planned. Multiple test problems are very common in clinical trials. Example applications include the comparison of a new treatment with

- Several other treatments
- A control for more than one endpoint
- A control for more than one population
- A control repeatedly in time
- … (or any combination thereof)

Multiple test problems in clinical trials are very diverse and many different methods are available. **It should be noted that Regulatory guidance listed below requires a description of the multiplicity adjustment in Phase III study protocols** (NOT FOUND IN 国家药监局药审中心发布的《药物临床试验多重性问题指导原则（试行）》):

- ICH E9 (1998) on "Statistical principles for clinical trials"
- FDA draft guidance for industry on "Multiple endpoint analyses" expected for 2014

**Commented [YL1]:** 摘自诺华 bretz ppt，待确认

## 1.3 Familywise error rate (FWER)

**Type I error rate**

It is defined as the probability of rejecting the null hypothesis when it is true.

Accordingly, the overall Type I error rate is defined as the probability of rejecting at least one true hypothesis. The probability can be computed under the assumption that all hypotheses are simultaneously true. This is known as the weak control of the **familywise error rate (FWER)**.

In the context of clinical trials with multiple endpoints, the weak FWER control can be interpreted as the probability of concluding an effect on at least one endpoint when there is no effect on any endpoint, i.e., the probability of concluding an ineffective treatment has an effect.

> *回顾一下，弱控制就是说，原假设都是真的，控制 I 类错误率就是弱控制，但是这个假设基本在现实中很难成立，所以我们会引出强控制的概念。对于强控制，在原假设有真有假的情况下，控制 I 类错误率不超过给定的 alpha 水平（比如双侧 0.05），就是强控制*

Using mathematical terminology, it can be reformulated as the control of the probability of incorrectly rejecting any true hypothesis regardless of which and how many other hypotheses are true. In other words, if $T$ is the index set of true null hypotheses, we require that

$$\text{supFWER} = \max_{T} \sup_{\{\mu_i(T)\}} \Pr(\text{reject at least one } H_i,\ i \in T) \leq \alpha \ ,\ i = 1,2,\dots,m$$

As an example, consider a **dose-finding study with $m$ doses tested versus placebo**.

- Let $\mu_0$ be the mean improvement in the placebo arm, $\mu_i$ be the mean improvement in the i$^{th}$ dose group.
- $\delta$ is a non-negative constant defining the clinically important difference.

The supremum is taken over all $\mu_i$ satisfying null hypothesis $H_i : \mu_i - \mu_0 \leq \delta$ for $i \in T$ and $H_i : \mu_i - \mu_0 > \delta$ for $i \notin T$, and the maximum is taken over all index sets $T$.

This approach to protecting the overall error rate is known as strong control of the familywise error rate. ***Strong control of the FWER for the primary objectives is mandated by regulators in all confirmatory clinical trials.***

> *In general, a strong control of the FWER means that we need to allocate $\alpha$ in advance. The choice of $\alpha$ allocation depends on specific problem and it may lead to different results. We need to determine the methods of adjustment during trial design.*

## 1.4 False discovery rate (FDR) and false discovery proportion (FDP)

If the number of rejected hypotheses is positive, then the FDP is defined as

$$\text{FDP}=\left(\frac{\#\text{ reject true null hypotheses}}{\#\text{ reject null hypotheses}}\right)$$

The FDR is said to be controlled at the $\gamma$ level if

$$\text{FDR}=\mathbb{E}(\text{FDP}) \leq \gamma$$

To ensure the control of the FDR at $\gamma$ level, one can choose an acceptable probability of exceedance $\alpha$ and require that

$$P(\text{FDP} > \beta) \leq \alpha$$

The interpretation is that of those hypotheses that are rejected, the proportion of false discoveries may exceed a specified fraction $\beta$ with probability no larger than $\alpha$.

Note that control of the FWER is equivalent to control of the FDP with $\beta$=0. Control of the FDR at the $\alpha$ level does not imply control of the FWER at the $\alpha$ level, nor does any ($\beta$>0) control of the FDP at the $\alpha$ level imply control of the FWER at the $\alpha$ level.

Control of the FDP makes sense in many nonconfirmatory settings like genetic or pre-clinical studies, where a certain proportion of errors is considered acceptable. FDR or FDP controlling procedures are not suitable for confirmatory clinical (Finner and Roter, 2001) trials.

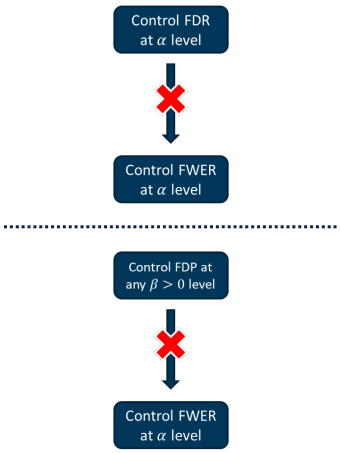



Figure 1 - The relationship among FDR, FDP and alpha

## 1.5 The reasons to consider multiple testing adjustments

When a set of hypotheses are tested simultaneously within the same study, the overall type I error rate (i.e. the probability of rejecting at least one null hypothesis given that all nulls are in fact true) is increased, potentially resulting in an increased risk of a false-positive finding.

If adequate adjustments are not made in multiple testing, findings may be misleading. Besides the increased risk of spurious statistical significance, multiplicity also has important implications for sample size determination and interpretation of study results.

Therefore we need to consider multiplicity adjustments in designing, analyzing and interpreting trials.

### 1.5.1 FDA presenters' view: example of substantial underreporting of true error rate

The details of this section are based on the ***Alpha-Recycling For The Analyses Of Primary And Secondary Endpoints Of Clinical Trials*** proposed by Mohammad Huque, Ph. D and Sirisha Mushti, Ph. D from FDA/CDER/OTS/ Office of Biostatistics.

In the presentation, they also introduced two key statistical approaches for the analyses in clinical trials which will be introduced in Section 3.3.1 and 3.3.2 separately.

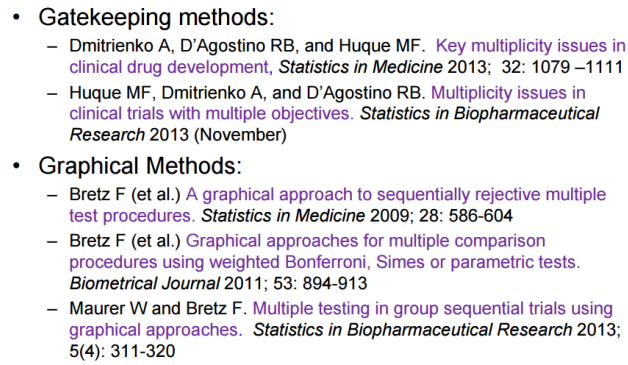



Figure 2 - Two key statistical approaches for the analyses of the Primary Endpoint and Secondary Endpoint hypotheses of clinical trials.

### 1.5.1.1 Example 1

Consider treatment-to-control comparisons in a trial on 4 endpoints (Dmitrienko, D'Agostino, and Huque; 2013):

- A is primary
- B, C and D are secondary

**Test strategy**

- Test for A at level 0.05
- If the test for A is significant, then test for B, C, and D each at level 0.05

Suppose that the global null hypothesis is true, i.e., there is no treatment effects for any endpoint. Then the probability of falsely concluding treatment effect in any endpoint = 0.05. That is FWER = 0.05. Because, tests for endpoints B, C, and D occur only after the test for endpoint A is significant at level 0.05. This renders the size of error rate for secondary endpoints not to exceed 0.05. This testing strategy seems to work well.

Suppose that the null hypothesis for A is false but those for B, C, and D are true.
Then the error rate for the test strategy can be as high as $1-(1-0.05)^3 = 0.142 > 0.05$ (on assuming tests are independent)!!!! **This testing strategy can lead to a substantial underreporting of true error rate!!!**

**Issues: Should the secondary endpoint family be always analyzed at the full alpha level (e.g., at 0.05) after the trial is successful on one or more specified primary endpoints?**

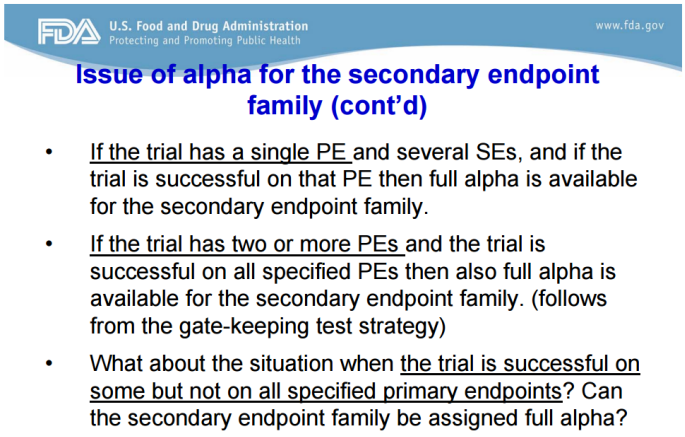



Figure 3 - PE is primary endpoint; SE is secondary endpoint.

### 1.5.1.2 Example 2

Consider a 2-arm trial designed to compare a treatment to control on two PEs (A and B) and on single secondary endpoint C.

Suppose that

- the Bonferroni method is applied for testing for A and B with each test at level 0.025, on splitting the trial alpha of 0.05.
- at the conclusion of the trial the observed treatment effect p-values are: $p_A < 0.001$ and $p_B = 0.20$.

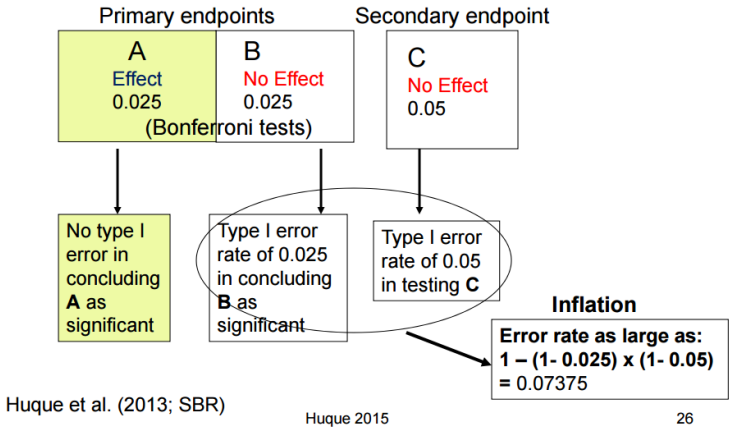



Figure 4 - Test PEs A and B, each at level 0.025, if win in one of them, then tests the secondary endpoint C at level 0.05.

# 2 Multiple testing principles

Two general principles in multiple testing, known as

- Union-intersection testing
- Intersection-union testing

There are also two methods for constructing multiple tests

- Closure principle （闭合原理）
- Partitioning principle （分割原理）

The mathematical definitions are:

- $H_1, \ldots, H_m$ denote the hypotheses corresponding to the multiple objectives and they are tested against the alternative hypotheses $K_1, \ldots, K_m$
- $H_I$ denotes the global hypothesis

## 2.1 Union-intersection testing

The global hypothesis $H_I$, defined as the intersection of the hypotheses, is tested v.s. the union of the alternative hypotheses ($K_U$):

$$H_I: \bigcap_{i=1}^{m} H_i \text{ vs. } K_U: \bigcup_{i=1}^{m} K_i$$

$$Testing\ the\ objective\ O_i\ consists\ in\ testing: \quad H_0^i \ vs. H_1^i \quad \forall\, i = 1, \ldots, m$$
$$and\ the\ FWER\ for\ the\ set\ of\ objectives\ \{O_1, \ldots, O_m\}\ is:$$
$$\alpha_{FWER} = P(\{reject\ H_0^1\} \cup \ldots \cup \{reject\ H_0^m\} \,|\, \{H_0^1\ true\} \cap \ldots \cap \{H_0^m\ true\})$$

Figure 5 - The FWER is the probability of rejecting at least one hypothesis when all hypotheses are true.

In the context of union-intersection testing, carrying out the individual tests at an unadjusted α level leads to an inflated probability of rejecting $H_I$ and can compromise the validity of statistical inferences.

To address this problem, a multiplicity adjustment method needs to be utilized to control the appropriately defined probability of a Type I error.

$$\begin{aligned}
\alpha_{FWER} &= P(\{reject\ H_0^1\} \cup \ldots \cup \{reject\ H_0^m\} \,|\, \{H_0^1\ true\} \cap \ldots \cap \{H_0^m\ true\}) \\
&= P(\{reject\ H_0^1 \,|\, H_0^1\ true\} \cup \ldots \cup \{reject\ H_0^m \,|\, H_0^m\ true\}) \\
\Leftrightarrow\ 1 - \alpha_{FWER} &= \mathrm{P}(\{accept\ H_0^1 | H_0^1\ true\} \cap \ldots \cap \{accept\ H_0^m | H_0^m\ true\}) \\
&= (1 - \alpha_1) \ldots (1 - \alpha_m) \\
\Leftrightarrow \alpha_{FWER} &= 1 - (1 - \alpha_1) \ldots (1 - \alpha_m)
\end{aligned}$$

Figure 6 - In this case, we are allowed to fail rejecting one or more hypotheses and still have conclusions for the others which are rejected. This kind of scenario is less restrictive and so, it is more used in clinical

trials where several objectives are tested.

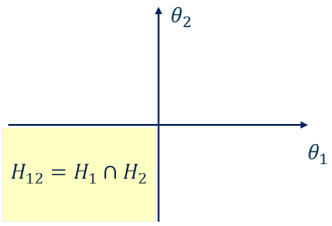
### 2.1.1 When will Type I error occur

Suppose a significant outcome about 2 objectives is required to declare study successful with Union-intersection testing method. We will reject $H_I$ iff $t_i \geq c$ which is the rejection region. $t_i$ is the statistics for hypothesis $i$ and $c$ is the $(1-\alpha)$ quantile from the marginal distribution of $t$; $i = 1,2$

In this example, we will test

$$H_1 : \theta_1 \leq 0$$
$$H_2 : \theta_2 \leq 0$$

We suppose $H_{12} = H_1 \cap H_2$ is true so that Type I error will occur when $H_1$ or $H_2$ is rejected.




Reject $H_1$ : evidence to show $\theta_1 > 0$
Reject $H_2$ : evidence to show $\theta_2 > 0$
$H_{12}$ deems to be rejected! Type I error occurs!

Figure 7 - Graphical presentation

## 2.2 Intersection-union testing

Intersection-union testing arises naturally in studies when a significant outcome with respect to two or more objectives is required in order to declare the study successful. For example, new therapies for the treatment of Alzheimer's disease are required to demonstrate their effects on both cognition and global clinical scores.

The global hypothesis $H_U$, defined as the union of the hypotheses, is tested v.s. the union of the alternative hypotheses ($K_I$):

$$H_U : \bigcup_{i=1}^{m} H_i \text{ vs. } K_I : \bigcap_{i=1}^{m} K_i$$

When the global hypothesis $H_U$ is rejected, one concludes that all $K_I$s are true, i.e., there is evidence of a positive effect with respect to all of the *m* objectives.

An interesting feature of intersection-union tests is that no multiplicity adjustment is necessary to control the size of a test but the individual hypotheses cannot be tested at levels higher than the nominal significance level either.

## 2.3 Closure principle

The closure principle proposed by Marcus, Peritz and Gabriel (1976) plays a key role in the theory

of multiple testing and provides a foundation for virtually all multiple testing methods arising in pharmaceutical applications. Marcus et al. (1976) showed that this closed testing procedure for the hypotheses controls the FWER in the strong sense at the α level.

This principle has been used to construct a variety of **stepwise** testing procedures for Union-intersection testing problems. In the general case of testing m hypotheses, the process of constructing a closed testing procedure goes through the following steps:

- Define the closed family of hypotheses. For each non-empty index set $I \subseteq \{1, \dots, m\}$, consider an intersection hypothesis defined as
$$H_I = \bigcap_{i \in I} H_i$$
- Establish implication relationships. An intersection hypothesis that contains another intersection hypothesis is said to imply it, i.e., $H_I$ implies $H_J$ if $J \subset I$
- Define local $\alpha$ level tests for individual intersection hypotheses. Let $p_I$ denotes the p-value produced by the associated local test and reject $H_I$ iff (if and only if) $p_J \leq \alpha$ for all $J \subseteq I$.

In particular, reject $H_i$ iff all intersection hypotheses containing $H_i$ are rejected by their local tests.

$\Leftrightarrow$ reject $H_i$ iff $p_I \leq \alpha$ for all index sets $I$ that include $i$

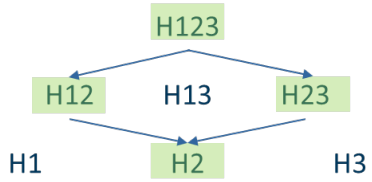
## 2.3.1 The mechanism of controlling $\alpha$ level

Suppose three null hypotheses to be tested using closure principle

$$H_1, H_2, H_3$$

Here $H_{ij} = H_i \cap H_j$ where $i \neq j$; And a fully closure set is $\mathcal{H} = \{H_1, H_2, H_3, H_{12}, H_{13}, H_{23}, H_{123}\}$. Obviously any $H_{ij} \subsetneq \mathcal{H}$.

Based on the closure principle, the highlighted items need to be significant if we want to conclude that $H_2$ is statistically significant as illustrated below:




**Mathematically, with Union-intersection testing:**

{reject $H_2$ using closure} = {reject $H_{123}$ using closure} ∩ {reject $H_{12}$ using closure} ∩ {reject $H_{23}$ using closure} ∩ {reject $H_2$ using closure} ⊆ {reject $H_{123}$ using closure} ;

Similarly, we have

{reject $H_1$ using closure} ⊆ {reject $H_{123}$ using closure};

{reject $H_3$ using closure} ⊆ {reject $H_{123}$ using closure}.

Therefore,
{reject $H_1$ using closure} ∪ {reject $H_2$ using closure} ∪ {reject $H_3$ using closure} ⊆ {reject $H_{123}$ using closure}.
That is,
$P$({reject $H_1$ using closure} ∪ {reject $H_2$ using closure} ∪ {reject $H_3$ using closure}) ≤ $P$({reject $H_{123}$ using closure}) = $\alpha$

## 2.3.2 Bonferroni-based closed testing procedures

Commented [YL2]: 可以移到 bonferroni 下面

Commented [YL3]: 阿斯利康 ppt 用的 recycling framework 的基础?

Understanding the closure principle (Section 2.3) enables one to take full advantage of its flexibility and to tailor the multiple testing procedure to the study objectives.

In the following we will:

- describe the class of Bonferroni-based closed testing procedures;
- give a sufficient characterization to derive sequentially rejective multiple testing procedures and demonstrate that many common procedures are in fact special cases thereof;
- provide graphical tools that facilitate the derivation and communication of Bonferroni-based closed testing procedures based on sequentially rejective rules that are tailored to study objectives.

### 2.3.2.1 Class of Bonferroni-based closed testing procedures

**Problem**

- Testing $m$ hypotheses $H_1, \dots, H_m$ and let $I = \{1, \dots, m\}$.
- Applying the closure principle leads to consideration of the intersection hypotheses $H_J = \bigcap_{j \in J} H_j$, where $J \subseteq I$.
- For each intersection hypothesis $H_J$ we assume a collection of non-negative weights $w_j(J)$ (These weights quantify the relative importance of the hypotheses $H_j$ included in the intersection $H_J$), where $0 \leq w_j(J) \leq 1$ and $\sum_{j \in J} w_j(J) = 1$.

In this section we assume that each intersection hypothesis is tested with a weighted Bonferroni test[A].

Consequently, we obtain the multiplicity adjusted p-values

$$p_J = \min\{q_j(J) : j \in J\}$$

For the weighted Bonferroni test for $H_J$, where

$$q_j(J) = \begin{cases} \dfrac{p_j}{w_j(J)} & \text{if } w_j(J) > 0 \\ 1 & \text{if } w_j(J) = 0 \end{cases}$$

[A] With the weighted Bonferroni test, we reject $H_J$ if $p_j \leq \alpha_j(J) = w_j(J)\alpha$ for at least one $j \in J$, where $p_j$ denotes the unadjusted p-value for $H_j$.

This defines Class $\mathcal{B}$ of all closed testing procedures that use weighted Bonferroni tests for each intersection hypothesis.

Any collection of weights subject to the constraints given above can be used and thus one can choose the weights and tailor the closed testing procedure to the given study objectives.

The **Shaffer procedure**, **fixed-sequence procedure**, **fallback procedure**, and all **Bonferroni-based gatekeeping procedures** are examples of multiple testing procedures from Class $\mathcal{B}$.

**Example**

Consider the simple two-hypothesis problem where the intersection hypothesis is $H_J$ with $J = \{1,2\}$ and $w_1(J) = w_2(J) = \frac{1}{2}$.

This results in the regular Bonferroni test and the adjusted p-value

$$p_J = \min\{q_1(J), q_2(J)\} = 2 \times \min\{p_1(J), p_2(J)\}.$$

If $H_J = H_{\{1,2\}} = H_1 \cap H_2$ is rejected, so is either $H_1$ or $H_2$, since they are tested subsequently at level $\alpha$. In other words, if $H_{\{1,2\}}$ is rejected (the smaller of the two p-values is less than $\alpha/2$), the remaining elementary hypothesis is tested at level $\alpha$.

Suppose the significance level is $\alpha$.
If $H_J = H_{\{1,2\}} = H_1 \cap H_2$ is rejected when

$$p_J = \min\{q_1(J), q_2(J)\} = 2\min\{p_1(J), p_2(J)\} < \alpha,$$

It means that

$$\min\{p_1(J), p_2(J)\} < \frac{\alpha}{2}.$$

Based on this derivation, we can conclude that $p_1(J) < \frac{\alpha}{2}$ or $p_2(J) < \frac{\alpha}{2}$. Therefore we can say that If $H_J = H_{\{1,2\}} = H_1 \cap H_2$ is rejected at level $\alpha$, so is either $H_1$ or $H_2$.

- Closed testing considers {$H_1 \cap H_2$, $H_1$ and $H_2$}
- Suppose we use Bonferroni test for $H_1 \cap H_2$. That is, reject $H_1 \cap H_2$ if unadjusted $p_j < \alpha/2$ for some $j \in \{1, 2\}$.
- Suppose that $H_1 \cap H_2$ is rejected for $j = 2$. Then by the consonance property of the test for $H_1 \cap H_2$, the hypothesis $H_2$ is rejected
- Consequently, by the CTP, the test for $H_1$ is at level $\alpha$ and not at level $\alpha/2$.

Figure 8 - closure principle with 2 hypotheses and it connection to $\alpha$-recycling and the graphical method.

### 2.3.2.2 Sequentially rejective Bonferroni-based closed testing procedures

It can further be shown that under a mild monotonicity condition on the weights $w_j(J)$ the closure principle leads to powerful **consonant** [B] multiple testing procedures.

Commented [YL4]: 补充到 closure principle 那里

Short-cut versions can thus be derived, which substantially simplify the implementation and interpretation of the related procedures.

Hommel, Bretz and Maurer showed that all the procedures mentioned previously (with the notable exception for the Shaffer procedure) belong to a subclass $\mathcal{S} \subset \mathcal{B}$ of shortcut procedures characterized by the property

$$w_j(J') \geq w_j(J) \quad \text{for all } J' \subseteq J \subseteq I \text{ and } j \in J'.$$

This condition ensures that if an intersection hypothesis $H_J$ is rejected, there is an index $j \in J$ such that $\frac{p_j}{w_j(J)} \leq \alpha$ ($p_j \leq w_j(J)\alpha \leq \alpha$) and the corresponding elementary $H_j$ is deemed to be rejected at level $\alpha$ by the closed testing procedure.

Therefore, short-cut procedures of order $m$ can be constructed:
**Instead of testing $2^m - 1$ hypotheses (as usually required by the closure principle), it is sufficient to test the elementary hypotheses $H_1, \ldots, H_m$ in $m$ steps.**

**Algorithm**
Therefore, shortcut procedures from $\mathcal{S}$ can be carried out with the following $m$-step procedure:
**Start:** testing the global intersection $H_I$, $I = \{1, \ldots, m\}$;
⇓
**Subsequent Step:** If it is rejected, there is an index $i \in I$ such that $H_i$ is rejected by the closed testing procedure;
⇓
**Subsequent Step:** continues testing the global intersection $H_{I\setminus i}$ until the first non-rejection.

Commented [YL5]: 每次排斥一个 elementary hypothesis，所以只需要做 m 步

Sequentially rejective (SR) graphical procedures are (implicitly) related to closed testing procedures that satisfy this property.

The paper by Bretz et al. (Bretz, Frank, Maurer, Willi, Brannath, Werner, & Posch, Martin, 2009) graphical approach (will be introduced in Section 3.3.2) satisfies this condition for all intersection hypotheses $H_J$.

[B] A closed testing procedure is termed **consonant** if the rejection of an intersection hypothesis $H_I$ with $I \subseteq \{1, \ldots, m\}$ and $|I| > 1$ always leads to the rejection of at least one $H_J$ implied by $H_I$, i.e., $H_J$ with $J \subset I$.

An example (Mohammad Huque, Ph.D & Sirisha Mushti, Ph.D, 2015) is provided in Figure 9, Figure 10 and Figure 11. Note that we can use graphical tools to visualize the specific rules in Figure 10.

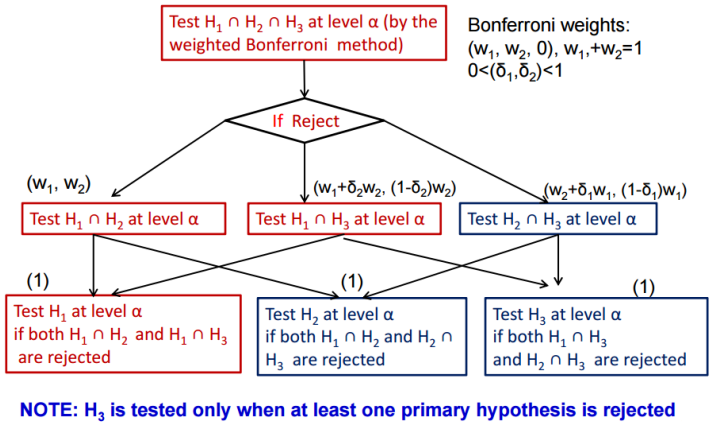
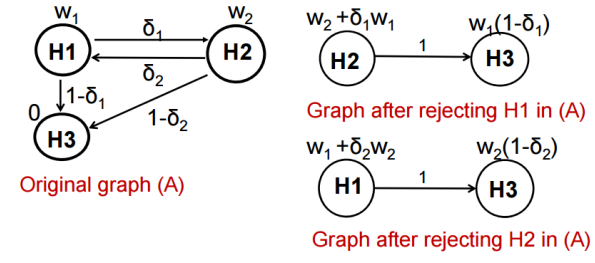



Figure 9 - $H_1$ and $H_2$ are primary hypotheses and $H_3$ is the secondary hypothesis. Note that $H_3$ is tested only when at least one primary hypothesis is rejected.

| Hypotheses | H1 | H2 | H3 |
|---|---|---|---|
| **H123** | **$w_1$** | **$w_2$** | 0 |
| H12 | $w_1$ | $w_2$ | 0 |
| **H13** | **$w_1+\delta_2 w_2$** | - | **$(1-\delta_2)w_2$** |
| H1 | 1 | - | - |
| **H23** | - | **$w_2+\delta_1 w_1$** | **$(1-\delta_1)w_1$** |
| H2 | - | 1 | - |
| H3 | - | - | 1 |

Figure 10 - Closure principle tables for Figure 9 with Bonferroni weights satisfying consonance.




Transition matrix of g-values in (A)

$g_{12} = \delta_1$, $g_{13} = 1-\delta_1$;

$g_{21} = \delta_2$, $g_{23} = 1-\delta_2$; $g_{31} = g_{32} = 0$

Figure 11 - Graphical representation of the Figure 9.

Details for Figure 9, Figure 10 and Figure 11 are derived in Appendix (Section 5.4).

#### 2.3.2.3 Graphical visualization

More details of this section will be provided in Section 3.3.2.

It was shown above that Class $\mathcal{S}$ includes a variety of Bonferroni-based testing procedures, such as

- fixed-sequence;
- fallback;
- gatekeeping procedures.

Using procedures in this class $\mathcal{S}$, one can map the difference in importance as well as the relationship between various study objectives onto a suitable multiple test procedure.

However, since the procedures are based on the closure principle, one needs to specify the weights $w_j(J)$ for each of the $2^m - 1$ intersection hypotheses $H_J,\ J \subseteq I$.

Unless these weights follow some simple and well-known specification rules (e.g., in Holm procedure), the underlying test strategy may be difficult to communicate to clinical trial teams.

Graphical tools have been proposed instead, which help visualizing different sequentially rejective test strategies and thus to best tailor a multiple testing procedure to given study objectives.

### 2.3.3 Properties of closed testing procedures

This section briefly describes important properties of closed testing procedures.

#### 2.3.3.1 Monotone procedures

A monotone procedure rejects a hypothesis whenever it rejects another hypothesis with a larger p-value. For example, if $p_i < p_j$ then the rejection of $H_j$ automatically implies the rejection of $H_i$.

Monotonicity helps to avoid logical inconsistencies; as such it is an essential requirement for multiple testing procedures. When a procedure does not have this property, monotonicity needs to be enforced by updating adjusted p-values. The Shaffer procedure (will be introduced later) serves as an example of a procedure that requires monotonicity to be enforced.

**Commented [YL6]:** 待添加

### 2.3.3.2 Consonant procedures

A closed testing procedure is termed **consonant** if the rejection of an intersection hypothesis $H_I$ with $I \subseteq \{1,\ldots,m\}$ and $|I| > 1$ always leads to the rejection of at least one $H_J$ implied by $H_I$, i.e., $H_J$ with $J \subset I$.

While consonance is generally desirable, non-consonant procedures can be of practical importance. The Hommel procedure defined later is an example of a non-consonant closed testing procedure. It is possible for this procedure to reject the global null hypothesis $H_I$, $I = \{1,\ldots,m\}$, without rejecting any other intersection hypotheses.

### 2.3.3.3 $\alpha$-exhaustive procedures

An $\alpha$-exhaustive procedure is a closed testing procedure based on intersection hypothesis tests the size of which is exactly $\alpha$ (Eugene Grechanovsky & Yosef Hochberg, 1999). In other words, $P(Reject\ H_I) = \alpha$ for any intersection hypothesis $H_I = \bigcap_{i \in I} H_i, I \subseteq \{1,\ldots,m\}$.

If a procedure is not $\alpha$-exhaustive, one can construct a uniformly more powerful procedure by setting the size of all intersection hypothesis tests at $\alpha$.

It is worth noting that some popular multiple testing procedures, for example, the fallback and Hochberg procedures described later, respectively, are not $\alpha$-exhaustive. These procedures are used in pharmaceutical applications due to other desirable properties such as computational simplicity.

Commented [YL7]: 使用 2.3 节标记

**Commented [YL8]:** 意思是，如果一个 proc 不是 alpha-exhaustive 的，那么可以构建一个更 powerful 的方法，其所有 intersection 假设都能在 alpha 水平上检验。据此说法，fallback proc 不是 alpha-exhaustive 的 procedure.

## 2.4 Partitioning principle

The partitioning principle was introduced by Stefansson, Kim and Hsu (1988) and Finner and Strassburger (2002). The advantage of using this principle is two-fold:

- It can be used to construct procedures that are **more powerful** than procedures derived using the closed testing principle.
- Partitioning procedures are easy to invert in order to **set up simultaneous confidence sets** for parameters of interest.

To illustrate the process of carrying out partitioning tests, consider the clinical trial example with two doses and a placebo. The first step involves partitioning the union of the hypotheses

$$H_1 : \mu_1 \leq \mu_0 \,,\ H_2 : \mu_2 \leq \mu_0$$

Into 3 mutually exclusive hypotheses:

$$H_1^* : \mu_1 \leq \mu_0 \ and\ \mu_2 \leq \mu_0 \quad \text{(I)}$$
$$H_2^* : \mu_1 \leq \mu_0 \ and\ \mu_2 > \mu_0 \quad \text{(II)}$$
$$H_3^* : \mu_1 > \mu_0 \ and\ \mu_2 \leq \mu_0 \quad \text{(III)}$$

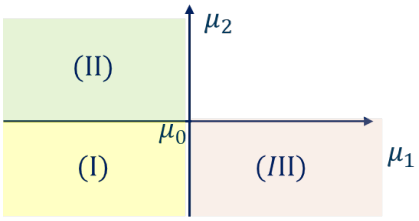


Figure 12 - Graphical presentation

Since the three hypotheses are disjoint, each one of them can be tested at level α without compromising the FWER control. The final decision rule is constructed by considering all possible outcomes for the three mutually exclusive hypotheses. For example,

A. If $H_1^*$ is rejected, we conclude that $\mu_1 > \mu_0$ or $\mu_2 > \mu_0$;
B. If $H_1^*$ and $H_2^*$ are rejected, we conclude that $\mu_1 > \mu_0$ (similarly, rejecting $H_1^*$ and $H_3^*$ implies that $\mu_2 > \mu_0$);
C. If $H_1^*$, $H_2^*$ and $H_3^*$ are all rejected, we conclude that $\mu_1 > \mu_0$ and $\mu_2 > \mu_0$.

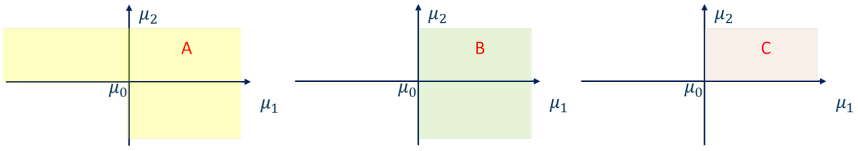


Figure 13 - Graphical presentation for the three mutually exclusive hypotheses

**General Case**

For hypotheses $H_1, \ldots, H_m$, the steps of partitioning principle are:

- For a given parameter space $\theta_L$, choose an appropriate partition $\{\theta_l : l \in L\}$;
- Test $\theta_l$ at unadjusted $\alpha$ level;
- If all $\theta_l$ intersecting $H_i$ which are rejected, we can conclude that $H_i$ is rejected.

Since these hypotheses are mutually exclusive, at most one of them is true. Thus, even though no multiplicity adjustment is made, the resulting multiple test controls the FWER at the $\alpha$ level.

## 2.5 Summary

We use a graph to demonstrate the difference between Union-intersection testing and Intersection-union testing:

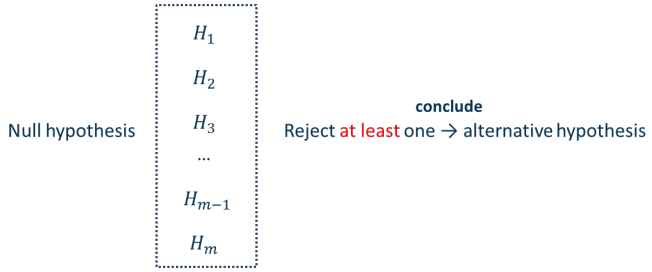



Figure 14 - Union-intersection testing

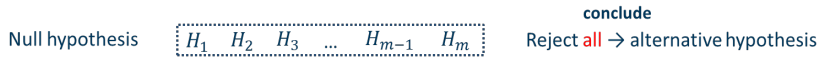



Figure 15 - Intersection-union testing

# 3 Multiple Testing Procedures

## 3.1 Classification of multiple testing procedures

| Category | Procedures | Features |
|---|---|---|
| Based on P-value or nonparametric methods | Bonferroni | Most common in practice, conservative |
| | Holm | Based on Bonferroni, may gain more power than Bonferroni |
| | Shaffer | Commonly used in the situations where there are logical relationships between null hypotheses |
| | Fixed-sequential procedure[1] | No requirement of adjustment |
| | Simes global test | Cannot make inference on elementary hypothesis |
| | Hommel test | Based on Simes, and gain more power than Holm-Bonferroni test |
| | Hochberg test | Based on Simes, and gain more power than Holm test but less than Hommel test |
| Parametric methods | Dunnett and related step-wise procedures | have requirement for data distribution |
| Resampling-based methods | Bootstrap re-sampling | |
| | Permutation test | |

Note:

**1**. Belong to Hierarchical test procedure, along with Fallback procedure; If the hierarchy of hypotheses is specified before data is observed, one can apply a hierarchical test procedure.

### 3.1.1 Single-step and stepwise procedures

- **Single-step procedures**

The decision to reject any hypothesis does not depend on the decision to reject any other hypothesis. In other words, the order in which the hypotheses are tested is not important and one can think of the multiple inferences as being performed simultaneously in a single step. The Bonferroni procedure and Dunnett procedure are examples of single-step procedures.

- **Stepwise procedures**

Stepwise procedures are carried out in a *sequential manner*. Some hypotheses are not tested explicitly and may be retained or rejected by implication. Stepwise procedures provide an attractive alternative to single-step procedures because *they can reject more hypotheses without inflating the overall error rate*.

The stepwise testing approach can be implemented via *step-down* or *step-up* procedures:

- ✓ A **step-down** procedure *starts with the most significant p-value* and continues in a sequentially rejective fashion until a certain hypothesis is retained or all hypotheses are rejected. If a hypothesis is retained, testing stops and the remaining hypotheses are retained by implication. The Holm procedure is an example of a step-down testing procedure.
- ✓ **Step-up** procedures approach the hypothesis testing problem from the opposite direction and carry out individual tests from the least significant one to the most significant one. The final decision rule is reversed compared to step-down procedures; i.e., once a step-up procedure rejects a hypothesis, it rejects the rest of the hypotheses by implication. The Hochberg procedure is an example of a step-up testing procedure.

### 3.1.2 Distributional assumptions

- **Based on P-value or nonparametric methods**

The Procedures that *don't make any assumptions about the joint distribution of the test statistics*. These procedures rely on univariate p-values and thus tend to have a rather straightforward form. They are referred to as p-value based procedures or nonparametric procedures. Examples include many popular procedures such as the Bonferroni and Holm procedures.

- **Parametric methods**

Procedures that *make specific distributional assumptions*, for example, that the test statistics follow a multivariate normal or t-distribution. To contrast this approach with nonparametric procedures based on univariate p-values, they are termed parametric procedures. Examples include the Dunnett and related procedures.

- **Resampling-based methods**

Procedures that do not make specific assumptions and attempt to ap-proximate the true joint distribution of the test statistics. The approximation relies on resampling-based methods (bootstrap or permutation methods) and thus procedures in this class are often referred to as resampling-based procedures.

### 3.1.3 Hierarchical or non-hierarchical structures of testing multiple hypotheses

**Commented [TT9]:** 是不是要 突出 multiple testing procedure

**Commented [YL10R9]:** 添加

Families of null hypotheses are said to be hierarchically ordered or ranked if earlier families serve as gatekeepers in the sense that one tests hypotheses in a given family if the preceding gatekeepers have been successfully passed.

The two commonly used hierarchical families of endpoints in a clinical trial are the family of primary endpoints and the family of secondary endpoints.

These two families are hierarchically ordered with the property that rejections or non-rejections of null hypotheses of secondary endpoints depend on the outcomes of test results of primary endpoints.

The individual endpoints within a family can also have hierarchical ordering, occurring naturally or by design. Hierarchical ordering of multiple endpoints and also of multiple comparisons can considerably reduce the multiplicity burden in controlling the FWER in a trial.

Based on the work by Mohammad Huque, Ph. D (Mohammad Huque, Ph.D & Sirisha Mushti, Ph.D, 2015), hypotheses in confirmatory trials usually follow a hierarchical structure:

- primary endpoint hypotheses are considered more important;
  - ✓ secondary endpoint hypotheses are usually tested for statistical significance after there is a favorable clinically meaningful and statistically significant result involving one or more primary endpoints;
  - ✓ Statistical approaches for clinical trials are therefore tailored to this hierarchical structure, normally optimizing the power for testing the primary endpoint hypotheses.
- For confirmatory trials, the use of standard methods such as Bonferroni, Holm, Hochberg, Dunnett t-tests, etc., on ignoring such hierarchical structures of test hypotheses, are generally considered inefficient.
- In these approaches, for making conclusions at the individual hypotheses levels, strong sense FWER control is needed across both the primary and secondary families of hypotheses.
- Two key statistical approaches for this have been developed that apply to confirmatory clinical trials:
  - ✓ Gatekeeping approaches
  - ✓ (sequentially rejective) graphical methods

Commented [YL11]: 需要过一遍 FDA 的指导文件

A short summary of the testing procedures is provided below (Deli Wang, et al., Overview of multiple testing methodology and recent development in clinical trials, 2015):

**Non-hierarchical hypotheses including but not limited to**

- Non-parametric and semi-parametric procedures
  - ✓ Bonferroni procedure
  - ✓ Simes procedure
  - ✓ Holm step-down procedure
  - ✓ Hochberg step-up procedure
  - ✓ Hommel procedure
- Parametric procedures
  - ✓ Dunnett procedure

**Hierarchical hypotheses including but not limited to**

Commented [TT12]:

- Simple procedures for hierarchical hypotheses
  - ✓ Fixed-sequence procedure

    $\xrightarrow{p_1 \le \alpha, p_2 \le \alpha, p_3 \le \alpha}$

  - ✓ Fallback procedure（**2003/2005**）
- Gatekeeping procedures
  - ✓ Serial gatekeeping procedures
  - ✓ Parallel gatekeeping procedure
  - ✓ Mixture gatekeeping procedure
  - ✓ Other extensions of gatekeeping procedures

**Integrate non-hierarchical and hierarchical hypotheses**

- Graphical approaches

## 3.2 Multiple Testing Procedures

### 3.2.1 Bonferroni method

It uses $\frac{\alpha}{m}$ for all inference; for $i = 1, \dots, m$:

Reject $H_i$ if $p_i \leq \frac{\alpha}{m}$.

With adjusted p-values $q_i = \min(mp_i, 1)$,
Reject $H_i$ if $q_i \leq \alpha$.
Note that $mpi$ > 1 is possible and we thus need to truncate the adjusted p-values at 1, resulting in the minimum expression.

Both rejection rules above lead to the same test decisions.

**Rationale**

The Bonferroni method follows from the **Boole's inequality**:

$$\mathrm{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathrm{P}(A_i),$$

where $A_i = \{p_i \leq \alpha/m\}$ denotes the event of rejecting $H_i$.

For $m = 2$,

According to the definition of Type 1 error, FWER=Prob{reject $H_1$ or reject $H_2$}when H1 and H2 are true, which is to say, $FWER = \mathrm{P}(p_1 \leq \frac{\alpha}{2}$ or $p_2 \leq \frac{\alpha}{2} | H_1, H_2$ are true)

$$= \mathrm{P}(p_1 \leq \tfrac{\alpha}{2} | H_1 is \text{ true}) + \mathrm{P}(p_2 \leq \tfrac{\alpha}{2} | H_2 is \text{ true}) - \mathrm{P}(p_1 \leq \tfrac{\alpha}{2} \text{ and } p_2 \leq \tfrac{\alpha}{2}$$

By Boole's inequality,

$FWER \leq \mathrm{P}(p_1 \leq \frac{\alpha}{2} | H_1 is \text{ true}) + \mathrm{P}(p_2 \leq \frac{\alpha}{2} | H_2 is \text{ true}) = 2 \times \frac{\alpha}{2} = \alpha$

**Properties**

The Bonferroni method is rather conservative if:

- The number of hypotheses is large;
- The test statistics are strongly positively correlated.

The Bonferroni method can be improved:

- Stepwise methods (e.g. **Holm** procedure; see later);
- Accounting for correlations (e.g. **Dunnett** test; see later).

While Bonferroni is rarely used in practice, it is the basis for commonly used advanced multiple test procedures.

## 3.2.2 Holm (Bonferroni based) procedure

Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered unadjusted p-values with associated null hypotheses $H_{(1)} \leq \cdots \leq H_{(m)}$.

Then we have the following stepwise procedure:

Table 1 - Raw p-values.

| | |
|---|---|
| If $\boldsymbol{p_{(1)} \leq \frac{\alpha}{m}}$ | Reject $\boldsymbol{H_{(1)}}$ and continue; else stop |
| If $\boldsymbol{p_{(2)} \leq \frac{\alpha}{m-1}}$ | Reject $H_{(2)}$ and continue; else stop |
| … | |
| If $\boldsymbol{p_{(i)} \leq \frac{\alpha}{m-i+1}}$ | Reject $H_{(i)}$ and continue; else stop |
| … | |
| If $\boldsymbol{p_{(m)} \leq \alpha}$ | Reject $H_{(m)}$ and continue; else stop |

With $p_{(1)} \leq \cdots \leq p_{(m)}$, define adjusted p-values using

- $\tilde{q}_{(1)} = mp_{(1)}$
- $\tilde{q}_{(2)} = \begin{cases} (m-1)p_{(2)}, & \text{if } (m-1)p_{(2)} > q_{(1)} \\ q_{(1)}, & otherwise \end{cases}$
- …
- $\tilde{q}_{(m)} = \begin{cases} p_{(m)}, & \text{if } p_{(m)} > q_{(m-1)} \\ q_{(m-1)}, & otherwise \end{cases}$

Where

$q_{(1)} = \min\{1, mp_{(1)}\}$ ;

$q_{(i)} = \min\{1, \max[(m-i+1)p_{(i)}, q_{(i-1)}]\}$ , $i = 2, \ldots, m$ .

**Properties**

The Holm procedure is a stepwise procedure that is more powerful than the Bonferroni method

- Bonferroni uses the same threshold $\frac{\alpha}{m}$ for all hypotheses;
- Holm uses the larger thresholds $\boldsymbol{\frac{\alpha}{m-i+1}}$.

Sometimes it is called “stepdown Bonferroni” procedure. The Holm procedure can be improved by accounting for correlations (e.g. stepdown Dunnett test; see later).

### 3.2.2.1 Holm's weighted procedure

When FWER control in the strong sense is desired, Holm's (1979) weighted procedure can be used (Yoav Benjamini & Yosef Hochberg, 1997).

Let $p_i^* = \frac{p_i}{w_i}$ and order $p_{(1)}^* \leq p_{(2)}^* \leq \cdots \leq p_{(m)}^*$. The hypothesis $H_{(i)}^*$ and weight $w_{(i)}^*$ correspond to $p_{(i)}^*$. Reject $H_{(i)}^*$ when

$$p_{(j)}^* \leq \frac{\alpha}{\sum_{k=j}^{m} w_{(k)}^*}, j = 1, \ldots, i$$

Holm's unweighted procedure (see Section 3.2.2) is the above procedure with all weights being equal (to 1).

### 3.2.2.2 How closed Bonferroni procedure gives the Holm's procedure

Theorem 1 – $V, n_0, n$ denote the number of incorrectly reject true hypotheses, the number of true hypotheses and the number of hypotheses respectively.

Bonferroni's method controls FWER at level $\alpha$ in a strong sense; that is,

$$FWER \leq \mathbb{E}[V] \leq \frac{n_0}{n}\alpha.$$

*Proof.* Observe that

$$\mathbb{P}(V \geq 1) \leq \mathbb{E}[V],$$

by Markov's inequality. By linearity of expectations, we have

$$\mathbb{E}[V] \leq \sum_{i:H_{0,i}\text{is true}} \mathbb{P}\{p_i \leq \alpha/n\} \leq \frac{n_0}{n}\alpha,$$

as required.

□

**Remark 1.** Note that Bonferroni's method exhibits a stronger notion of error control, in that $\mathbb{E}[V] \leq \alpha$. This is referred to as control of the per familywise error rate.

**Remark 2.** Note that in the proof of Theorem 1, we made no use of the independence or dependence of the $p$-values $p_i$. Thus, Bonferroni's method is valid for dependent tests.

For each non-empty index set $I \subseteq \{1, \ldots, m\}$, consider an intersection hypothesis (global null) defined as

$$H_I = \bigcap_{i \in I} H_i$$

For each index set $I$, consider a valid level-α test $\varphi_I$ for testing $H_I$ (reject if $\varphi_I = 1$ )

$$P(\varphi_I = 1 \mid H_I) \leq \alpha.$$

We use Bonferroni's procedure to construct $\varphi_I$ by Theorem 1:

$$\varphi_I = 1 \Leftrightarrow \min_{i \in I} p_i \leq \alpha/|I|$$

We now show that we can avoid testing $2^m - 1$ number of tests and simply run $m$ tests, where $m$ is the number of hypotheses.

1. Fix any $j$ and consider any $I \subseteq \{1, \dots, m\}$ such that $p_{(j)} = \min\{p_i : i \in I\}$, i.e. the $j$th smallest p-value (among the $m$ p-values) is in fact the minimum p-value among those indexed in $I$.
2. Let $I^+ = \{(j), (j+1), \dots, (m)\}$ denote the index set corresponding to the $m - j + 1$ largest p-values. We claim that

$$\varphi_{I^+_{(j)}} = 1 \implies \varphi_I = 1$$

This is because if $\varphi_{I^+_{(j)}} = 1$ then

$$p_{(j)} = \min\left\{p_i : i \in \varphi_{I^+_{(j)}}\right\} \le \frac{\alpha}{|I^+_{(j)}|} = \frac{\alpha}{m - j + 1} \le \frac{\alpha}{|I|},$$

Note that since $p_{(j)}$ is the smallest p-value in $I$, we must have $|I| \le m - j + 1$.

3. Since by the definition of Bonferroni's procedure[c] which can control type I error in a strong sense

$$\varphi_I = 1 \Leftrightarrow \min_{i \in I} p_i \le \frac{\alpha}{|I|}$$

4. Using the previous claim, we note the following:

$H_{(j)}$ is rejected by closed test

$j = 0;$

**While** $p_{(j+1)} \le \frac{\alpha}{m-j}$ **do**

$j = j + 1;$

Reject $H_{(1)}, \dots, H_{(j)}$.

$\Leftrightarrow \varphi_{I^+_{(1)}} = 1$ and ... and $\varphi_{I^+_{(j)}} = 1$

$\Leftrightarrow \varphi_{I^+_{(1)}} = \frac{\alpha}{m}$ and ... and $\varphi_{I^+_{(j)}} = \frac{\alpha}{m-j+1}$

This gives us a simple procedure to determine which hypotheses to reject based on the closed test: Note that these procedures give us precisely the **Holm's procedure**. This is encouraging in that the closure principle appears to yield reasonable procedures.

## 3.2.3 Fixed-sequence procedure

Recycling of significance levels can also be seen in the fixed-sequence test procedure which is often used for testing multiple hypotheses of clinical trials when the hypotheses are hierarchically ordered in pre-specified testing sequence.

$H_1, \dots, H_m$ is pre-specified (this order normally reflects the clinical importance of the multiple analyses).

**Commented [YL14]:** 这里要注意，p(j)一直是 I 集合里最小的 p-value；
这里**注意逻辑顺序**:
先固定 j，比如 j=3，再去考虑 I，使得 p(3)是 I 里面的最小的 p-value，那么这个 I 集合必然是不包含 p(1),p(2)的，也就是只含有(m-3+1)个元素。

**Commented [YL15]:** P(1)是 I 里最小的 p，这时候 I 集合里有 m 个元素；
P(2)是 I 里最小的 p，这时候 I 集合里有 m-2+1 个元素；

1. [c] Bonferroni test: Reject $\cap_{i=1}^m H_i$ if $\min(P_1, P_2, \dots, P_m) \le \frac{\alpha}{m}$

Testing begins with the first hypothesis, $H_1$, and each test is carried out without a multiplicity adjustment as long as significant results are observed in all preceding tests.

In other words, the hypothesis $H_i, i = 1, \ldots, m$, is rejected at the $i$th step if

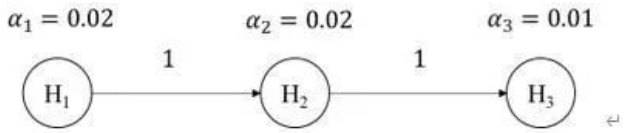
$$p_j \leq \alpha, \quad j = 1,2, \ldots i.$$

The fixed-sequence procedure controls the FWER because, for each hypothesis, testing is conditional upon rejecting all hypotheses earlier in the sequence, which also is the main drawback of fixed-sequence test procedure because it stops testing (i.e., no further recycling allowed) as soon as it fails to reject a hypothesis even if one or more of subsequent hypotheses have extremely small p-values..

Non-inferiority to Superiority(非劣转优效)

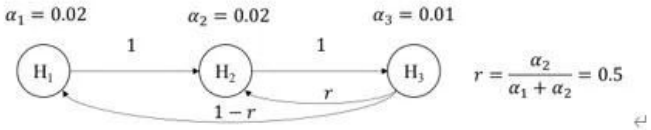
## 3.2.4 Fallback procedure

The fallback method was proposed to address this drawback of the fixed-sequence test strategy. Simple Fallback procedure（单向传递的 fallback）




问题是什么?

Improved Fallback procedure




## 3.2.5 Simes method

Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered unadjusted p-values with associated null hypotheses $H_{(1)} \leq \cdots \leq H_{(m)}$.

Simes method uses all ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ and

Reject global hypothesis global hypothesis of no treatment effect $H_I$ (union-intersection, see Section 2.1) if $p_{(i)} \leq \frac{i\alpha}{m}$ for at least one $i$=1,2,...m;

Where

$\boldsymbol{H_I} = \cap_{i=1}^{m} \boldsymbol{H_i} : \theta_1 = \theta_2 = \cdots = \theta_m = 0$ and $p_1, \ldots, p_m$ are independent.

**Properties**

- Simes' adjusted p-value uses $\min_i \frac{mp_{(i)}}{i}$, which is less than or equal to Bonferroni's $mp_{(1)}$.
- Simes cannot be used to test the individual hypotheses $H_i$.
- Type I error rate is at most $\alpha$ under independence or (certain types of) positive dependence of p-values.

**Simes V.S. Bonferroni**

Suppose $m = 2$ and independence of $p_1$ and $p_2$. The rejecting probability is visualized in Figure 16.

**Bonferroni test**: Reject $\bigcap_{i=1}^{2} H_i$ if $p_{(1)} \leq \frac{\alpha}{2}$; see **Figure 17**:

$$\text{P(reject null hypothesis | Bonferroni)} = 1 - \left(1 - \frac{\alpha}{2}\right)^2 = \alpha - \left(\frac{\alpha}{2}\right)^2 < \alpha$$

**Simes method**: Reject $\bigcap_{i=1}^{2} H_i$ if $p_{(1)} \leq \frac{\alpha}{2}$ or $p_{(2)} \leq \alpha$; see **Figure 18**:

$$\text{P(reject null hypothesis | Simes)} = 1 - \left(1 - \frac{\alpha}{2}\right)^2 + \left(\frac{\alpha}{2}\right)^2 = \alpha$$

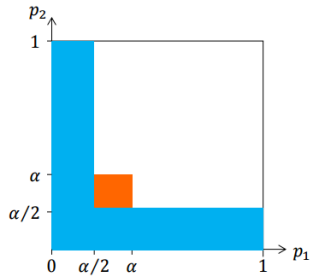



Figure 16 - Bonferroni's versus Simes method; Simes is more powerful than a global test based on Bonferroni, and Simes assumes non-negative correlations between p-values, Bonferroni does not. A decomposition is given below.

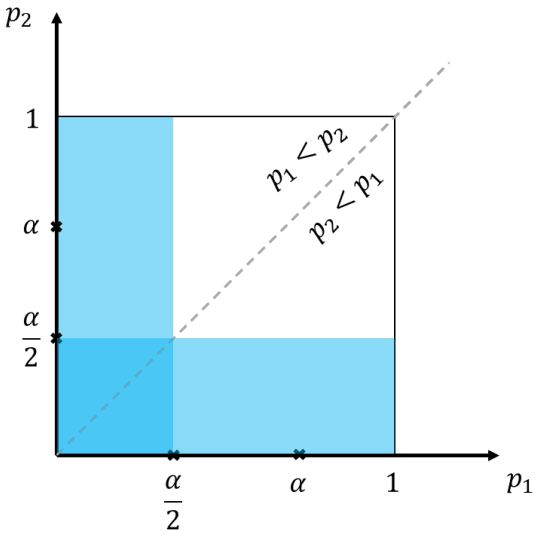



**Figure 17** - Bonferroni's rejection region. The visible area is the rejection region.

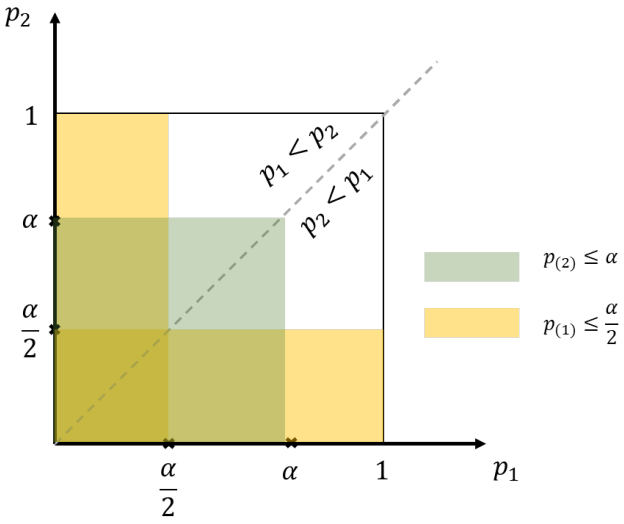



**Figure 18** - Simes' rejection region. The visible area is the rejection region. It is obvious that the Simes' rejection region contains the Bonferroni rejection region.

Since the assumption of independence is unlikely to be met in practice, several authors examined operating characteristics of this test under dependence.

Figure 19 (Alex Dmitrienko, et al., 2010) depicts the relationship between the Type I error rate of

the Simes test, number of comparisons $m = 2,5$ and common correlation coefficient $\rho$ in the case of normally distributed test statistics.

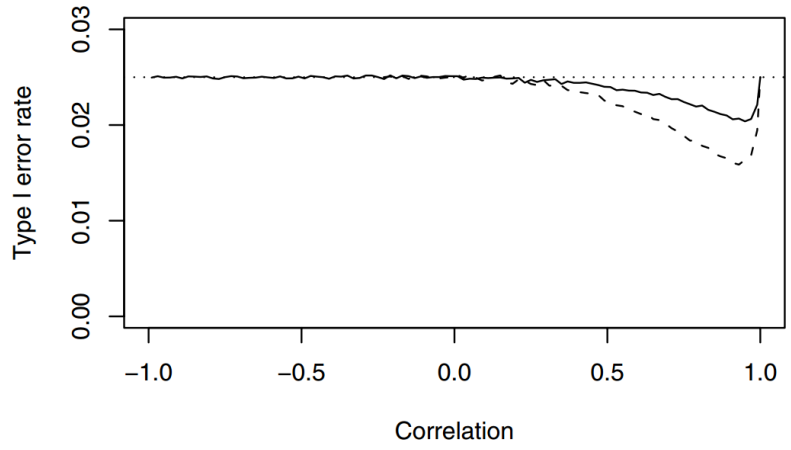



Figure 19 - Type I error rate of the Simes test under the global null hypothesis as a function of the number of comparisons and correlation (solid curve, $m = 2$ comparisons, correlation $> -1$; dashed curve, $m = 5$ comparisons, correlation $> -0.25$). The Simes test is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025.

## 3.2.6 Hochberg Procedure

The Hochberg procedure is another popular procedure based on the Simes method. The Hochberg procedure is an example of a step-up procedure based on univariate p-values. Unlike step-down procedures (e.g., the Holm procedure), this procedure begins with the least significant p-value and examines the other p-values in a sequential manner until it reaches the most significant one.

Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered unadjusted p-values with associated null hypotheses $H_{(1)} \leq \cdots \leq H_{(m)}$.

Beginning with the case of equally weighted hypotheses, the decision rule for the Hochberg procedure is defined as follows:

Table 2 - Raw p-values test algorithm for Hochberg

| | |
|---|---|
| If $\boldsymbol{p_{(m)} \leq \alpha}$ | Reject $\boldsymbol{H_{(1)}, \ldots, H_{(m)}}$ and stop; else retain $\boldsymbol{H_{(m)}}$ and go to next step |
| If $\boldsymbol{p_{(m-1)} \leq \frac{\alpha}{2}}$ | Reject $H_{(1)}, \ldots, H_{(m-1)}$ and stop; else retain $H_{(m-1)}$ and go to next step |
| ... | |
| If $\boldsymbol{p_{(i)} \leq \frac{\alpha}{m-i+1}}$ | Reject $H_{(1)}, \ldots, H_{(i)}$ and stop; else retain $H_{(m-i+1)}$ and go to next step |

...

| | |
|---|---|
| If $\boldsymbol{p_{(1)} \leq \frac{\alpha}{m}}$ | Reject $H_{(1)}$; else retain $H_{(1)}$ |

Adjusted p-values are

$q_{(m)} = p_{(m)}$ ;

$q_{(i)} = \min\{(m-i+1)p_{(i)}, q_{(i+1)}\}$ , $i = m-1, \ldots, 1$ .

**Properties**

- It is more powerful than the Bonferroni-based Holm procedure:
  - ✓ Both procedures use the same thresholds, but Hochberg starts with the largest p-value, whereas Holm starts with the smallest;
- It makes the same assumptions as the Simes test (i.e. independence or positive dependence of p-values);
- The Hochberg procedure can be improved → **Hommel** procedure (see Section 3.2.7) based on the closed test procedure.

**Commented [TT16]:** 建议这下面阐述 truncated holm and Hochberg， 以及相较于 regular holm he Hochberg 的优势

## 3.2.7 Hommel procedure

It was explained in Section 3.2.2 that the Holm procedure results from using a global test based on the Bonferroni for testing intersection hypotheses in a closed procedure. Similarly, the Hommel procedure results from using the Simes method for testing individual intersection hypotheses.

In the case of equally weighted hypotheses, the Hommel procedure can be applied using the following algorithm:

Table 3 - Raw p-values for test algorithm of Hommel procedure.

| | |
|---|---|
| If $\boldsymbol{p_{(m)} > \alpha}$ | Retain $\boldsymbol{H_{(m)}}$ and go to next step; else reject all hypotheses and stop. |
| ... | |
| If $\boldsymbol{p_{(m-j+1)} > \frac{(i-j+1)\alpha}{i}}$ for $\boldsymbol{j = 1, \ldots, i}$ | Retain $H_{(m-i+1)}$ and go to next step; else reject all hypotheses and stop. |
| ... | |
| If $\boldsymbol{p_{(m-j+1)} > \frac{(i-j+1)\alpha}{i}}$ for $\boldsymbol{j = 1, \ldots, m}$ | Retain $H_{(1)}$; else reject $H_{(1)}$. |

It is uniformly more powerful than the Holm procedure because the Simes test is uniformly more powerful than the global test based on the Bonferroni procedure.

For example, the Holm procedure rejects $H_{(1)}$ if and only if $p_{(1)} \leq \frac{\alpha}{m}$ whereas the Hommel

procedure can reject this hypothesis when $p_{(1)} > \frac{\alpha}{m}$, e.g., $H_{(1)}$ is rejected if $p_{(m)} \leq \alpha$.

## 3.2.8 Dunnett Test

When comparing several treatments with a control, the Dunnett test can be used. The methods from Bonferroni, Holm, Simes, and Hochberg can also be used in these situations, but only the Dunnett test exploits the correlation between the p-values.

The following setting will be used throughout this section. Consider a dose-finding clinical trial designed to compare m doses or regimens of a treatment to a placebo. For simplicity, a balanced one-way layout will be assumed; i.e.,

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where

$Y_{ij}$ denotes observation $j = 1, \dots, n_i$ in group $i = 0,1, \dots, m$;

$\mu_i$ the effect of treatment group $i$;

$\varepsilon_{ij}$ are independent and identically normally distributed with mean 0 and variance $\sigma^2$, i.e. $\varepsilon_{ij} \sim N(0, \sigma^2)$.

The ANOVA $F$-test tests the global null $H: \mu_0 = \cdots = \mu_m$.

Here we are interested in comparing $m$ treatments with the control treatment $i$ = 0, i.e. testing the $m$ null hypotheses

$$H_i: \theta_i = \mu_i - \mu_0 \leq 0, \quad i = 1,2, \dots, m$$

Consider the $m$ pairwise t-tests $t_i$ which have the following properties:

- $t_i = \frac{\hat{\mu}_i - \hat{\mu}_0}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_0}}}$, where $\hat{\mu}_i$ and $\hat{\sigma}$ are the ordinary least squares (OLS) of $\mu_i$ and $\sigma$;
- $t_i \sim t_\nu$ under $H_i$, where $t_\nu$ denotes the univariate $t$-distribution with $\nu = \sum_i n_i - m - 1$ degrees of freedom;
- $(t_1, \dots, t_m)$ follows the $m$-variate $t$-distribution with $\nu$ degrees of freedom and correlations

$$\rho_{ij} = \sqrt{\frac{n_i}{n_i + n_0}}\sqrt{\frac{n_j}{n_j + n_0}}, \quad i, j = 1,2, \dots, m$$

**Rejection Rule**

For the $m$ individual null hypotheses, reject $H_i$ if $t_i \geq c_{m,1-\alpha}$ where the quantile $c_{m,1-\alpha}$ is computed such that

$$P\left[(t_1, \dots, t_m) \leq (c_{m,1-\alpha}, \dots, c_{m,1-\alpha})\right] = P\left(\max_i t_i \leq c_{m,1-\alpha}\right) = 1 - \alpha.$$

It should be noted that $(t_1, \dots, t_m)$ follows $m$-variate $t$-distribution (see Section 5.1 for details) with $\nu$ degrees of freedom and correlations $\rho_{ij}$, for $i, j = 1,2, \dots, m$.

In other words, $c_{m,1-\alpha}$ is the $1 - \alpha$ quantile of the distribution of the maximum of $m$ $t$-

distributed random variables.

**Properties**

- Single step test, which is better than Bonferroni as it exploits the known correlations between test statistics;
- Adjusted p-values can be calculated numerically based on the multivariate $t$-distribution;
- The Dunnett test shown here can be extended to any linear and generalized linear model;
- It can be improved by extending it to a stepwise procedure, similar to the Holm procedure (see later);
- Other well-known parametric tests follow the same principle:
  - ✓ For example, the Tukey test compares all treatment groups against each other, also using a multivariate $t$-distribution.

Commented [YL17]: To be added

**Stepwise Dunnett Test**

Let $t_{(1)} \geq \cdots \geq t_{(m)}$ denote the ordered test statistics with associated null hypotheses $H_{(1)}, \ldots, H_{(m)}$.

Table 4 - Algorithm for stepwise Dunnett test procedure.

| | |
|---|---|
| If $\boldsymbol{t_{(1)} \geq c_{m,1-\alpha}}$ | Reject $\boldsymbol{H_{(1)}}$ and continue; else stop. |
| If $\boldsymbol{t_{(2)} \geq c_{m-1,1-\alpha}}$ | Reject $H_{(2)}$ and continue; else stop. |
| … | |
| If $\boldsymbol{t_{(i)} \geq c_{m-i+1,1-\alpha}}$ | Reject $H_{(i)}$ and continue; else stop. |
| … | |
| If $\boldsymbol{t_{(m)} \geq c_{1,1-\alpha}}$ | Reject $H_{(m)}$. |

*Note that $\boldsymbol{c_{m-i+1,1-\alpha}}$ denotes the $1-\alpha$ quantile of the distribution of the maximum of $m-i+1$ $t$-distributed random variables and is computed from the corresponding multivariate $t$-distribution. For example, if $\boldsymbol{H_{(1)}}$ rejected, then the quantile $c_{m-1,1-\alpha}$ is computed from a $(m-1)$-variate $t$-distribution.

Other properties for stepwise Dunnett test:

- It can be shown that $c_{m,1-\alpha} \geq c_{m-1,1-\alpha} \geq \cdots \geq c_{1,1-\alpha}$, where $c_{1,1-\alpha} = t_{\nu,1-\alpha}$ is the quantile from the univariate $t$-distribution with $\nu$ degrees of freedom;
- The stepwise Dunnett test is better than the Holm procedure as it exploits the known correlations between test statistics;
- The stepwise version shown here is sometimes called “stepdown Dunnett” test;
- A “stepup Dunnett” test also exists, like **Hochberg**;

**Summary**

Table 5 - A summary of common multiple testing procedures with/without considering correlations.

| | **Correlation** | | |
|---|---|---|---|
| | Without | | With |
| **Single step** | Bonferroni | Simes | Dunnett |
| **Stepwise** | Holm | Hochberg | Stepdown Dunnett |

**Remarks:**

- Single step methods are less powerful than stepwise methods and not often used in practice;
- Accounting for correlations leads to more powerful procedures, but correlations are not always known;
- Simes-based methods are more powerful than Bonferroni-based methods, but control the FWER only under certain dependence structures;
- *In practice, we select the procedure that is not only powerful from a statistical perspective, but also appropriate from clinical perspective*.

## 3.3 Novel multiple testing techniques

### 3.3.1 Gatekeeping Approaches

Importantly, gatekeeping methods were introduced for addressing advanced multiplicity problems; that is, problems with several families of null hypotheses in which each family serves as gatekeeper for the next one. One moves from one family to the next when a prespecified condition in the gatekeeper family is satisfied. Gatekeeping methods apply to testing of families of hypotheses of primary, secondary, and other objectives of clinical trials.

Because of their usefulness, these approaches have received much attention to-date in the statistical literature. Gatekeeping procedures also satisfy a key regulatory requirement for confirmatory trials that the methods applied must provide strong control of the global FWER of the trial.

The word "global" is used here to indicate that the overall Type I error rate must be protected across all primary and secondary families of hypotheses of the trial. It is not sufficient to provide local FWER control within each family.

There are three types of gatekeeping approaches addressed in the literature that provide global FWER control for the trial and offer investigators powerful methods for addressing advanced multiplicity problems (Mohammad F. Huque, Alex Dmitrienko, & Ralph D'Agostino, 2013):

- the regular gatekeeping procedures;
  - ✓ serial gatekeeping
  - ✓ parallel gatekeeping
  - ✓ tree-structured gatekeeping
- the gatekeeping procedures with retesting;
- the gatekeeping procedures with simultaneous testing.

**Commented [YL18]:** 待补充

The usual strategy is to test all endpoints in the primary family by a method such as Bonferroni and proceed to the secondary family of endpoints only if there has been statistical success in the primary family.

This allows all of the trial $\alpha$ to be used for the primary family. Thus, maximizing the study power for those critical endpoints.

**Commented [TT19]:** 下面我会考虑添加 multistage fallback 和 multistage truncated hochberg

#### 3.3.1.1 Motivation

As a side note, gatekeeping procedures focus on multiplicity adjustments in a single trial.

In the context of registration/marketing authorization packages that normally include two

confirmatory trials with similar sets of primary and secondary analyses, gatekeeping methods can be applied independently to each trial. This approach will ensure Type I error rate control within each confirmatory trial.

To justify the inclusion of secondary findings into the product label, the trial's sponsor can use consistency arguments and demonstrate that multiplicity-adjusted primary and secondary analyses lead to similar conclusions in both trials.

It is worth noting that regulatory guidelines do not currently discuss rules for combining secondary findings across several confirmatory trials (Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;, 2010).

### 3.3.1.2 Clinical trials with serial gatekeepers

**Serial gatekeepers** are often found in clinical trials with multiple ordered endpoints. For example, in a clinical trial with a single primary endpoint and several key secondary endpoints, the endpoints may be arranged in a sequence.

We will begin with a two-family testing problem arising in clinical trials with noninferiority and superiority objectives.

#### 3.3.1.2.1 Two-family testing problem

Consider a trial in patients with ***Type II diabetes*** with three treatment groups

- Group A (a new formulation of an insulin therapy)
- Group B (a standard formulation)
- Group A+B (a combination of the formulations).

The following two scenarios will be examined:

- **Scenario1**. Noninferiority and superiority tests are carried out sequentially for the comparison of A versus B.
- **Scenario2**. A noninferiority test for the comparison of A versus B is carried out first followed by a superiority test for the same comparison and a noninferiority test for the comparison of A+B versus B.

Mathematical notations:

$\delta_1$: the true treatment differences for the comparisons of A versus B

$\delta_2$: the true treatment differences for the comparisons of A+B versus B

$\gamma_1$: positive non-inferiority margin for the comparisons of A versus B

$\gamma_2$: positive non-inferiority margin for the comparisons of A+B versus B

**Sets of hypotheses**

A versus B (noninferiority): $H_1: \delta_1 \leq -\gamma_1$ versus $K_1: \delta_1 > -\gamma_1$

A versus B (superiority): $H_2: \delta_1 \leq 0$ versus $K_2: \delta_1 > 0$
A versus B (noninferiority): $H_3: \delta_2 \leq -\gamma_2$ versus $K_3: \delta_2 > -\gamma_2$

Testing begins with the first family $\mathcal{F}_1$ that includes the test for $H_1$.

First family serves as a **serial gatekeeper** in the sense that all hypotheses of no treatment effect must be rejected in the first family to proceed to the second family.

In Scenario 1, the second family includes the test for $H_2$; in Scenario 2, this family includes the tests for $H_2$ and $H_3$. The decision rules are depicted in the Figure 20.

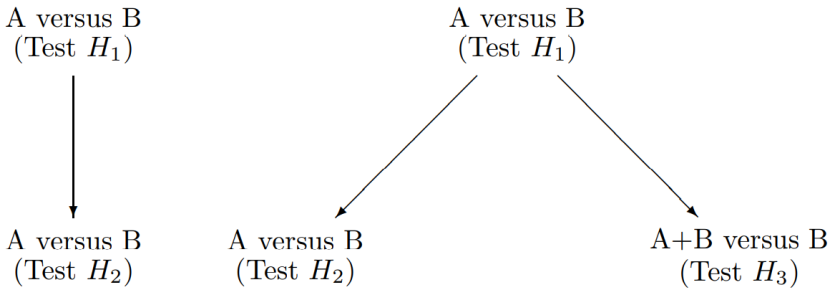



Figure 20 - Scenario 1, left panel; Scenario 2, right panel. Family 1 $\mathcal{F}_1 = \{H_1\}$ serves as a gatekeeper in both scenarios. Note that in Scenario 1, both tests can be carried out at the pre-specific $\alpha$ level since this testing procedure is a special case of the fixed-sequence approach. But multiplicity adjustment should be implemented in Scenario 2.

### 3.3.1.2.2 Serial Gatekeeping procedures

Serial gatekeeping procedures are widely used in clinical trials, mainly due to the fact that they do not require an adjustment for multiplicity.

Consider a clinical trial with multiple, hierarchically ordered objectives/analyses.

To account for the hierarchical ordering, the analyses are grouped into m families denoted by $\mathcal{F}_1, \ldots, \mathcal{F}_m$. Each family ($m$ families in total) includes null hypotheses corresponding to the analyses at the same level in the hierarchy shown in the Table 6.

Table 6 - Families of null hypotheses corresponding to multiple, hierarchically ordered objectives.

| Family | Null hypotheses | Hypothesis weights | Raw $p$-values |
|---|---|---|---|
| $F_1$ | $H_{11}, \ldots, H_{1n_1}$ | $w_{11}, \ldots, w_{1n_1}$ | $p_{11}, \ldots, p_{1n_1}$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $F_i$ | $H_{i1}, \ldots, H_{in_i}$ | $w_{i1}, \ldots, w_{in_i}$ | $p_{i1}, \ldots, p_{in_i}$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $F_m$ | $H_{m1}, \ldots, H_{mn_m}$ | $w_{m1}, \ldots, w_{mn_m}$ | $p_{m1}, \ldots, p_{mn_m}$ |

Note that $w_{in_i}$ is the weight representing the importance of hypotheses within $\mathcal{F}_i$ and $p_{in_i}$ is the associated p-value. Multiplicity adjusted p-values for the hypotheses in $\mathcal{F}_i$ is denoted by $\tilde{p}_{in_i}$. The adjusted p-values are defined with respect to all m families rather than any individual family.

Hierarchical arrangements of endpoints are often used in oncology trials, e.g., overall survival duration, progression-free survival duration, tumor response rate, time to treatment failure and duration of tumor response.

**General serial gatekeeping framework**

If family $\mathcal{F}_i, i = 1,2, \ldots, m-1$ is a **serial gatekeeper**, hypothesis in $\mathcal{F}_{i+1}$ are tested iff

$$\max_{j=1,2,\ldots,n_i} \tilde{p}_{ij} \leq \alpha.$$

**Single decision-making branch**

Within each family $\mathcal{F}_i, i = 1,2, \ldots, m-1$, hypotheses are tested at the nominal $\alpha$ level.

For example, the hypotheses in $F_i$ can be tested using an Intersection-Union test; i.e. all hypotheses are rejected in $\mathcal{F}_i$ if $p_{ij} \leq \alpha, j = 1, \ldots, n_i$ and all hypotheses are retained otherwise. Any FWER-controlling test can be used in $\mathcal{F}_m$, including all popular multiple tests.

**Commented [YL20]:** 引一下 regular multiple tests 的章节

Adjusted p-values for single-branch procedures are easy to compute using the Westfall-Young definition discussed in Section 5.2.

Assume that the Intersection-Union test is used in $\mathcal{F}_i, i = 1,2, \ldots, m-1$:

- $p_i^*$: the largest p-value in family $\mathcal{F}_i$.

$p'_{mj}$: the adjusted p-value for hypothesis $H_{mj}$ produced by the test used in the last family, $j = 1, \ldots, n_m$;

The adjusted p-value for $H_{ij}$ is given by

$$\tilde{p}_{ij} = \begin{cases} \max(p_1^*, \ldots, p_i^*), & \text{if } i = 1, \ldots, m-1 \\ \max(p'_{ij}, p_1^*, \ldots, p_{i-1}^*), & \text{if } i = m \end{cases}$$

**Multiple decision-making branches**

More complicated examples of serial gatekeeping procedures arise in clinical trials with multiple

sequences of hypotheses or multiple decision making branches, e.g., dose-finding studies with ordered endpoints.

In this case, at each fixed dose level, dose-control comparisons for multiple endpoints form a branch within which hypotheses are tested sequentially.

Serial gatekeeping procedures with multiple branches can be constructed based on several multiple tests (see ).

Here we will focus on Bonferroni-based procedures (serial gatekeeping procedures based on other tests are briefly discussed in Section ).

Commented [YL21]: 待补充

Consider a multiple testing problem with $m$ families and assume that each one contains $n$ hypotheses, i.e., $n_1 = \ldots = n_m = n$.

In this case there are $n$ branches and the $j$th branch includes the hypotheses $H_{1j}, \ldots, H_{mj}$. Hypotheses within each branch are tested sequentially as follows:

- Consider the $j$th branch, $j = 1, \ldots, n$. The hypothesis $H_{1j}$ is tested first at an $\alpha/n$ level. If $H_{1j}$ is rejected, the next hypothesis in the sequence, i.e., $H_{2j}$, is tested, otherwise testing within this branch stops.
- In general, the hypothesis $H_{ij}$ is rejected if $p_{kj} \leq \alpha/n$ for all $k = 1, \ldots, i$.

**Type II diabetes clinical trial example**

The Type II diabetes clinical trial as an example is conducted to compare three doses of an experimental treatment (labeled **L**, **M** and **H**) versus **placebo** (labeled **P**).

Each dose-placebo test is carried out with respect to three ordered endpoints:

- hemoglobin A1c (Endpoint E1)
- fasting serum glucose (Endpoint E2)
- HDL cholesterol (Endpoint E3).

The E2 tests are restricted to the doses at which Endpoint E1 is significant and, similarly, the E3 tests are carried out only for the doses at which the E1 and E2 tests are both significant. The testing procedures are depicted in Figure 21. The fixed-sequence approach is applied within each branch. The branches are "connected" using the Bonferroni test as described below. The hypotheses within the three branches are tested sequentially using the Bonferroni-based procedure.

Logical restrictions of this kind facilitate drug labeling and, in addition, improve the power of clinically relevant secondary dose-placebo tests.

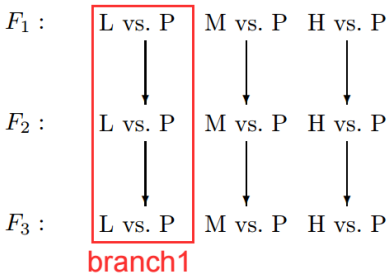



Figure 21 - Three-branch serial gatekeeping procedure with three families of hypotheses in the Type II diabetes clinical trial example (F1, Endpoint E1; F2, Endpoint E2; F3, Endpoint E3). The hypotheses $H_{i1}$ (H-P comparison), $H_{i2}$ (M-P comparison) and $H_{i3}$ (L-P comparison) for the $i$th endpoint are included in $\mathcal{F}_i$, $i = 1,2,3$. The hypotheses are equally weighted within each family and the FWER is set at a two-sided $\alpha = 0.05$.

The two-sided raw and adjusted p-values in this clinical trial example are summarized in Table 7 (Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;, 2010).

Table 7 - Serial gatekeeping procedure in the Type II diabetes clinical trial example. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level. The detailed calculation of adjusted p-value is given in .

| Family | Endpoint | Comparison | Raw $p$-value | Adjusted $p$-value |
|---|---|---|---|---|
| $F_1$ | E1 | L vs. P | 0.0176 | 0.0528 |
| $F_1$ | E1 | M vs. P | 0.0108 | 0.0324* |
| $F_1$ | E1 | H vs. P | 0.0052 | 0.0156* |
| $F_2$ | E2 | L vs. P | 0.0128 | 0.0528 |
| $F_2$ | E2 | M vs. P | 0.0259 | 0.0777 |
| $F_2$ | E2 | H vs. P | 0.0093 | 0.0279* |
| $F_3$ | E3 | L vs. P | 0.0511 | 0.1533 |
| $F_3$ | E3 | M vs. P | 0.0058 | 0.0777 |
| $F_3$ | E3 | H vs. P | 0.0099 | 0.0297* |

**Commented [YL22]:** 加一个计算细节在附录，书上 5.4.2 有算法，但是没有计算细节。

Only Doses M and H are significantly different from placebo for the primary endpoint (Endpoint E1) and thus the remaining branch corresponding to the L-P comparison is eliminated at the first stage of the procedure.

At the second stage, the dose-placebo comparisons for Endpoint E2 are performed only for the dose levels at which Endpoint E1 is significant, i.e., Doses M and H. There is no evidence of a significant effect at Dose M compared to placebo for Endpoint E2 and thus testing within that branch stops.

At the last stage, Dose H is tested against placebo for Endpoint E3. This test is significant and thus

we conclude that Dose H is superior to placebo for all three endpoints whereas Dose M is superior to placebo only for Endpoint E1.

### 3.3.1.3 Clinical trials with parallel gatekeepers

If family $\mathcal{F}_i, i = 1,2,\dots,m-1$ is a **parallel gatekeeper** if at least one significant result must be observed in this family, i.e., one or more hypotheses must be rejected in $\{H_{i1},\dots,H_{in_i}\}$, to proceed to $\mathcal{F}_{i+1}$. Hypothesis in $\mathcal{F}_{i+1}$ are tested at the $\alpha$ level iff

$$\min_{j=1,2,\dots,n_i} \tilde{p}_{ij} \leq \alpha.$$

A sample testing procedure is shown in Figure 22. Examples can be found in clinical trials with **multiple primary endpoints** when each endpoint provides independent proof of efficacy and can lead to a regulatory claim.

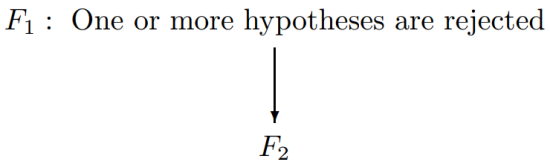



Figure 22 - A problem with a parallel gatekeeper $F_1$.

The parallel gatekeeping methods were introduced in Dmitrienko, Offen and Westfall (Alexei Dmitrienko; Walter W Offen; Peter H Westfall;, 2003) who considered a Bonferroni-based procedure derived using the closure principle (see Section 2.3.2). Since this method relies on a complete enumeration of all intersection hypotheses in the closed family associated with $F_1,\dots,F_m$, the resulting parallel gatekeeping procedures may lack transparency and their implementation can be computationally intensive since it takes order-$2^n$ steps to test $n$ hypotheses.

Further research in this area revealed that a broad class of parallel gatekeeping procedures have a stepwise form. This property streamlines their implementation and interpretation by clinical trial practitioners (US Food and Drug Administration statisticians have repeatedly emphasized the importance of multiple testing procedures that can be understood by clinicians (Center for Drug Evaluation and Research & Center for Biologics Evaluation and Research, 2017)).

#### 3.3.1.3.1 $\alpha$-Propagation

A key concept that has led to the development of novel methods for handling multiplicity problems of clinical trials is the concept of α-propagation (Dmitrienko, Ajit C Tamhane, & Brian L Wiens, 2008).

If a null hypothesis (or endpoint) is tested at a level $\alpha$ (e.g., $\alpha = 0.05$) and its associated p-value

Commented [YL23]: 区别 serial 的 all

Commented [YL24]: Multistage truncated 例子

is significant at this level, then this $\alpha$ is saved and is not lost. This alpha can then be recycled and propagated (or passed) to other null hypotheses or families of null hypotheses. This concept is used in methods such as the fallback, the graphical, and the more general chain methods for handing traditional multiplicity problems of clinical trials.

#### 3.3.1.3.1.1 α-Propagation rules

| Family 1 | Family 2 |
|---|---|
| • **Procedure 1** at $\boldsymbol{\alpha_1 = \alpha}$ level;<br>• $\boldsymbol{A_1 \subseteq N_1}$, index set of null hypotheses accepted in family 1 $\boldsymbol{F_1 = \{H_1, \dots, H_k\}}$ (null hypotheses, $\boldsymbol{N_1 = \{1, \dots, k\}}$ is the index set);<br>• Component procedure (Procedure 1): we should use separable procedure (will be defined later) with local FWER control. | • **Procedure 2** at $\alpha_2$ level;<br>• $\alpha_2 = \alpha_1 - e_1(A_1)$, where $e_1(I)$ is error rate function (will be define later) of procedure 1 and $I \subseteq N_1$;<br>• Family 2 $F_2 = \{H_{k+1}, \dots, H_{2k}\}$ (null hypotheses, $N_2 = \{k+1, \dots, 2k\}$ is the index set);<br>• Component procedure (Procedure 2): we can use any procedure with local FWER control. |

**Error rate function**

Assume that all null hypotheses $H_i,\ i \in I$ are true Error rate function is probability of rejecting at least one true null hypothesis

$$e_1(I) = \sup_{H_I} P\left\{\bigcup_{i \in I} (\text{Reject } H_i) \mid H_I\right\}, \qquad I \subseteq N_1$$

Where "Reject $H_i$" represents the event (the subset of the sample space) that corresponds to rejection of $H_i$ and the supremum of the probability is taken over the entire null space defined by $H_I = \bigcap_{i \in I} H_i$, including any false hypotheses $H_j, j \notin I$. Thus $e_1(I)$ is the maximum probability of making at least one type I error in the sub-family $\{H_i, i \in I\}$.

Based on the definition above, we have:

**(1)**

$A_1 = \emptyset$:

$e_1(\emptyset) = 0$, all null hypotheses are rejected in $F_1$; Here we have $\alpha_2 = \alpha_1 - e_1(\emptyset) = \alpha$.

Null hypotheses in $F_2$ are tested at full $\alpha$ level.

**(2)**

$A_1 = N_1$:

$e_1(N_1) = \alpha$, no null hypotheses are rejected in $F_1$; Here we have $\alpha_2 = \alpha_1 - e_1(N_1) = 0$.

Null hypotheses in $F_2$ are not tested.

For an example, error rate function of Bonferroni procedure is $e_1(I) = \frac{\alpha|I|}{k}$, where $|I|$ is the cardinality of set $I$, i.e., the number of elements in index set $I$.

**Separable procedures**

Procedure 1 is separable if $e_1(I) < \alpha$ provided $I$ is a proper subset[D] of $N_1$.

That is, if a separable procedure is used in $F_1$, a fraction of $\alpha$ can be carried over to $F_2$ if one or more null hypotheses are rejected in $F_1$.

**A **proper subset** of a set A is a subset of A that is not equal to A. In other words, if B is a proper subset of A, then all elements of B are in A but A contains at least one element that is not in B. For example, if A={1,3,5} then B={1,5} is a proper subset of A. The set C={1,3,5} is a subset of A, but it is not a proper subset of A since C=A. The set D={1,4} is not even a subset of A, since 4 is not an element of A.*

[D] A proper subset of a set A is a subset of A that is not equal to A. In other words, if B is a proper subset of A, then all elements of B are in A but A contains at least one element that is not in B. For example, if A={1,3,5} then B={1,5} is a proper subset of A. The set C={1,3,5} is a subset of A, but it is not a proper subset of A since C=A. The set D={1,4} is not even a subset of A, since 4 is not an element of A.

Table 8 - Bonferroni versus Holm with the separable property; Note that $N_1 = \{1,2,3\}$.

<table>
<tr><th>Bonferroni procedure</th><th>Holm procedure</th></tr>
<tr><td>Problem with three null hypotheses<br><table><tr><th>Index set $I$</th><th>Error rate function $e_1(I)$</th></tr><tr><td>$\{1,2,3\}$</td><td>$\alpha$</td></tr><tr><td>$\{1,2\}$, $\{1,3\}$, $\{2,3\}$</td><td>$2\alpha/3$</td></tr><tr><td>$\{1\}$, $\{2\}$, $\{3\}$</td><td>$\alpha/3$</td></tr><tr><td>Empty</td><td>$0$</td></tr></table></td><td>Problem with three null hypotheses<br><table><tr><th>Index set $I$</th><th>Error rate function $e_1(I)$</th></tr><tr><td>$\{1,2,3\}$</td><td>$\alpha$</td></tr><tr><td>$\{1,2\}$, $\{1,3\}$, $\{2,3\}$</td><td>$\alpha$</td></tr><tr><td>$\{1\}$, $\{2\}$, $\{3\}$</td><td>$\alpha$</td></tr><tr><td>Empty</td><td>$0$</td></tr></table></td></tr>
<tr><td>• Error rate function of Bonferroni procedure is $\boldsymbol{e_1(I) = \frac{\alpha|I|}{k}}$;<br>• Bonferroni procedure is separable because $\boldsymbol{e_1(I) < \alpha}$ if $\boldsymbol{I}$ is a proper subset of $\boldsymbol{N_1}$.</td><td>• Error rate function of Holm procedure is $e_1(I) = \alpha^{E}$ unless $I$ is empty;<br>• Holm procedure is not separable.</td></tr>
</table>

**Separability procedures**

Most popular procedures (Holm, fallback, Hochberg and Hommel procedures) do not satisfy the separability condition (Dmitrienko, Ajit C Tamhane, & Brian L Wiens, 2008).

**Truncated procedures**

Truncated procedure is based on a convex combination between a multiple procedure and Bonferroni procedure. Truncated procedure is separable.

- Truncated p-value-based procedures: Truncated Holm, fallback and Hochberg procedures.
- Truncated parametric procedures: Truncated step-down Dunnett procedure.

Refer to the paper ***General Multistage Gatekeeping Procedures*** (Dmitrienko, Ajit C Tamhane, & Brian L Wiens, 2008) for more details of truncated procedures.

**Gatekeeping procedures**

Wide variety of parallel gatekeeping procedures can be built based on these truncated procedures.

### 3.3.1.3.1.2 Algorithm of general multistage gatekeeping procedure

The algorithm is developed based on the Section 3.3.1.3.1.1.

**Proposition**

The 2-stage gatekeeping procedure controls the FWER at the $\alpha$ level. (proof is provided on Section )

The simple two-stage procedure provides useful insights into the nature of gatekeeping inferences.

---

[E] The Holm MTP incorrectly rejects any true hypothesis with probability α and hence is not separable. To see this, consider the problem of testing $H_i : \mu_i = 0$ versus $H_i' : \mu_i > 0$, $(1 \le i \le n)$. Suppose that $\mu_j = 0$ for some $j$ and $\mu_i \to \infty$, $i \ne j$. Then $p_i \to 0$ for $i \ne j$ and $p_j$ will be the largest $p$-value. Therefore, $p_j$ will be compared with α and $H_j$ will be rejected with probability α.

It is important to note that any FWER-controlling MTP can be used at the second stage of the 2-stage gatekeeping procedure. Therefore, one can construct gatekeeping procedures with an arbitrary number of stages by a recursive application of the two-stage procedure.

Since a serial gatekeeper can be expressed as a series of single-hypothesis families, multistage gatekeeping procedures obtained via the recursive algorithm can have a very flexible structure that combines serial gatekeepers and parallel gatekeepers.

**Commented [YL25]:** Need to added in gatekeeping part.

**Characteristics to define the multistage gatekeeping procedure**

$m \geq 2$ families;

$F_i = \{H_{i1}, \dots, H_{im}\}$ for $1 \leq i \leq m$;

$N_i = \{1, \dots, n_i\}$ and $A_i \subseteq N_i$ be the index set corresponding to the accepted hypotheses in $F_i$;

The algorithm for applying the procedure is:

**Start**

Initialize $\alpha_1 = \alpha$

⇓

**Stage 1 ~ Stage $m-1$**

Test $F_i$ at $\alpha_i$ level using any separable procedure with a suitable upper bound on the error rate function $e_i(I)$. Set

$$\alpha_{i+1} \leftarrow \alpha_i - e_i(A_i)$$

---

If $A_i = N_i$, i.e. no hypotheses are rejected in $F_i$, then apply the following procedures, stop testing and accept all hypotheses in $F_{i+1}, \dots, F_m$

$$e_i(A_i) = \alpha_i$$

$$\alpha_{i+1} \leftarrow 0$$

Otherwise, go to next step.

⇓

**Start $m$**

Use any FWER-controlling procedures (e.g., Holm, Hochberg etc.) to test $F_m$ at $\alpha_m$ level.

**Remarks**

- If all hypotheses are rejected at the $i$-th stage ($1 \leq i \leq m-1$), then $A_i = \emptyset$ and $\alpha_{i+1} = \alpha_i$. Thus full $\alpha_i$ is carried over to the next stage.
- At the final stage, any FWER controlling multiple testing procedure may be used, but a truncated multiple testing procedure should not be used since it is less powerful than its untruncated version.

### 3.3.1.3.2 Multistage parallel gatekeeping procedures

Two concepts that play a key role in the framework for constructing **multistage parallel gatekeeping procedures**: the *error rate function* of a multiple test and *separable* multiple tests. Details of these two concepts and the algorithm for constructing general multistage gatekeeping procedure are introduced in Section 3.3.1.3.1.1 and Section 3.3.1.3.1.2 respectively.

Dmitrienko, Tamhane and Wiens (Dmitrienko, Ajit C Tamhane, & Brian L Wiens, 2008) proposed truncated versions of popular tests by taking a convex combination of their critical values with the critical values of the Bonferroni test.

As a result, a truncated test is uniformly more powerful than the Bonferroni test but uniformly less powerful than the original test.

Some truncated versions of some tests are define in Appendix 5.5; e.g., the Holm, the Hochberg, the fallback and Dunnett tests, and their error rate functions.

### 3.3.1.3.3 Computation of adjusted p-values for a given null hypotheses and a multiple testing procedures

**Commented [YL26]:** 待补充，与例子一起理解更好

The Westfall-Young definition of an multiplicity adjusted p-value is: the adjusted p-value for a given null hypothesis and an MTP is defined as the significance level at which the procedure rejects the hypothesis.

**Commented [YL27]:** 定义为可以拒绝 hypothesis 的显著水平

It is given in Appendix 5.2 can be applied to calculate adjusted p-values for multistage gatekeeping procedures using the following direct calculation algorithm.

This algorithm loops through a grid of significance levels between 0 and 1 to find the lowest level at which each hypothesis is rejected:

Let $\alpha = \frac{k}{K}\ (0 < k < K)$ for some sufficiently large value of $K$.

The adjusted p-value, $\tilde{p}_{ij}$, for hypotheses $H_{ij}$ is the smallest $\alpha$ (corresponding to the smallest $k$) for which $H_{ij}$ is rejected.

Since multistage gatekeeping procedures have a simple stepwise form, this direct-calculation algorithm is quite fast even when the number of hypotheses is large.

### 3.3.1.3.4 Example: EPHESUS trial

This trial (Bertram Pitt, et al., 2001) was conducted to assess the effects of eplerenone on morbidity and mortality in patients with severe heart failure. In this clinical trial example, we will consider

two families of endpoints:

- Two primary endpoints:
  - ✓ all-cause mortality (Endpoint P1, with hypothesis[F] $H_{11}$)
  - ✓ cardiovascular mortality + cardiovascular hospitalization (Endpoint P2).
- Two major secondary endpoints:
  - ✓ cardiovascular mortality (Endpoint S1)
  - ✓ all-cause mortality + all-cause hospitalization (Endpoint S2).

The family of primary endpoints serves as a **parallel gatekeeper** for the family of secondary endpoints. The hypotheses are equally weighted within each family and the pre-specified FWER is $\alpha = 0.05$. Table 9 displays two sets of two-sided p-values for the four endpoints that will be used in this example (note that these p-values are used here for illustration only).

Table 9 - Two-sided p-values in the cardiovascular clinical trial example.

| Family | Hypothesis | Endpoint | Raw $p$-value | |
|---|---|---|---|---|
| | | | Scenario 1 | Scenario 2 |
| $F_1$ | $H_{11}$ | P1 | 0.0121 | 0.0121 |
| $F_1$ | $H_{12}$ | P2 | 0.0337 | 0.0872 |
| $F_2$ | $H_{21}$ | S1 | 0.0084 | 0.0084 |
| $F_2$ | $H_{22}$ | S2 | 0.0160 | 0.0160 |

A two-stage parallel gatekeeping procedure will be set up as follows:

- The hypotheses in $\mathcal{F}_1$ and $\mathcal{F}_2$ will be tested using the truncated and regular Holm tests, respectively.

The truncated Holm test is carried out using four values of the truncation parameter ($\gamma = 0,\ 0.25,\ 0.5$ and $0.75$) to evaluate the impact of this parameter on the outcomes of the four analyses.

**Scenario 1**

Let $\gamma = 0.25$. The hypotheses $H_{11}$ and $H_{12}$ are tested using the truncated Holm test at $\alpha_1 = \alpha = 0.05$. The smaller p-value, $p_{11} = 0.0121$, is less than

$$\left[\frac{\gamma}{2} + \frac{1-\gamma}{2}\right]\alpha = \frac{\alpha}{2} = 0.025$$

And thus $H_{11}$ is rejected.

Further, $p_{12} = 0.0337$, is compared to

$$\left[\frac{\gamma}{2} + \frac{1-\gamma}{2}\right]\alpha = \frac{5\alpha}{8} = 0.03125.$$

The corresponding hypothesis cannot be rejected.

To find the fraction of $\alpha$ that can be carried over to the hypotheses in $\mathcal{F}_2$, note that the set of retained hypotheses in $\mathcal{F}_1$ includes only one hypothesis. Thus,

[F] Assume no treatment effect between treatment groups.

$$\alpha_2 = \alpha_1 - e_1(A_1) = \alpha - \left[\gamma + \frac{(1-\gamma) \mid A_1 \mid}{n}\right]\alpha = \frac{3\alpha}{8} = 0.01875,$$

Where $\mid A_1 \mid = 1$ and $n = 2$.

Applying the regular Holm test in $\mathcal{F}_2$ at $\alpha_2$, it is easy to verify that $H_{21}$ and $H_{22}$ are rejected at level $\alpha_2$.

### 3.3.1.3.5 Example: LDL-C trial

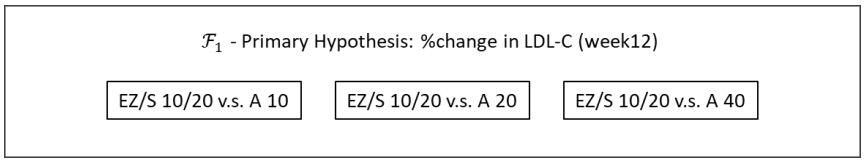
There are 3 primary hypotheses in the study:

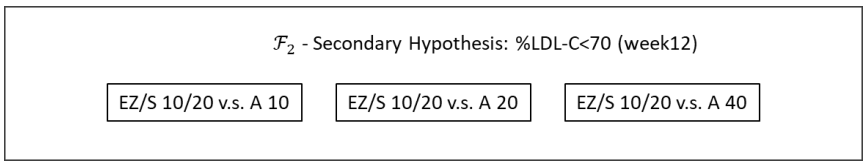
- The ezetimibe/simvastatin combination tablet at 10/20 mg will lower LDL-C more than atorvastatin 10 mg after 12 weeks of treatment;
- The ezetimibe/simvastatin combination tablet at 10/20 mg will lower LDL-C more than atorvastatin 20 mg after 12 weeks of treatment;
- The ezetimibe/simvastatin combination tablet at 10/40 mg will lower LDL-C more than atorvastatin 40 mg after 12 weeks of treatment.

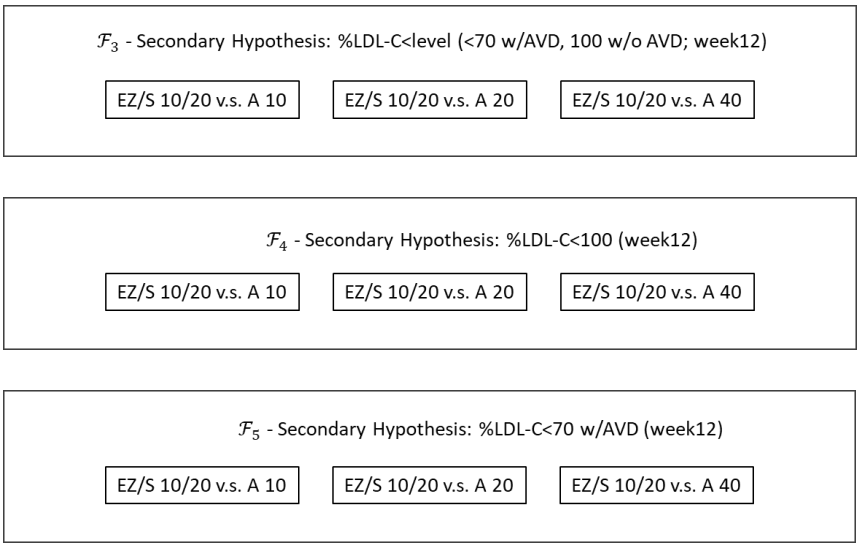
There are 4 sets of secondary hypotheses, which relate to the following endpoints:

1. Percent of patients with LDL-C < 70 mg/dL (1.81 mmol/L);
2. Percent of patients with LDL-C <100 mg/dL (2.59 mmol/L) (for patients with moderately high or high risk for CHD without atherosclerotic vascular disease) and LDL-C <70 mg/dL (1.81 mmol/L) (for patients with high risk for CHD with atherosclerotic vascular disease);
3. Percent of patients with LDL-C <100 mg/dL (2.59 mmol/L);
4. Percent of patients with LDL-C <70 mg/dL (1.81 mmol/L) among patients with high risk for CHD with atherosclerotic vascular disease.

**Each of the secondary hypotheses involves the same 3 treatment comparisons as the primary hypotheses.** Families of hypotheses to be tested are constructed as:






To control for multiplicity across the primary and secondary hypotheses, **multistage parallel gatekeeping procedures (Truncated Hochberg's procedure and regular Hochberg's procedure)** will be used in $\mathcal{F}_1$ to $\mathcal{F}_4$ and $\mathcal{F}_5$, respectively, at an $\alpha$-level of $0.05$ (see Figure 23).

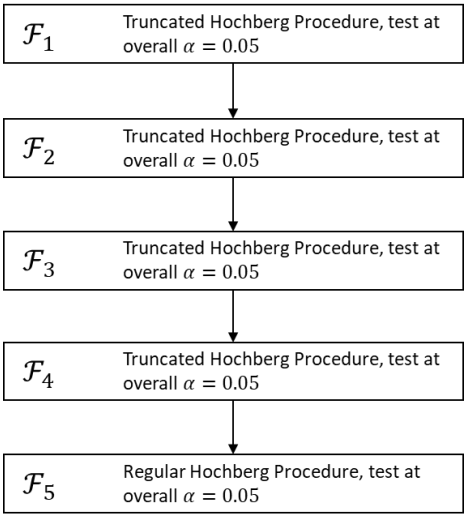



Figure 23 - A decision tree associated with parallel gatekeeping testing strategy.

This strategy provides reasonable control for the experiment-wise error rate, in that the set of primary comparisons, each set of secondary comparisons, and each unique treatment comparison related to the primary and secondary hypotheses are controlled at the 0.05 level. The nominal p-

values for all comparisons will be reported.

The truncated Hochberg test utilizes the same set of critical values but it is set up as a step-up test. For $\gamma = 0$ and $\gamma = 1$, this truncated test reduces to the Bonferroni and regular Hochberg tests, respectively. Based on Section 5.6.2, for the LDL-C trial, the multiple testing procedures are performed as follows:

1. Order all 15 p-values associated with 15 hypotheses $p_{(1)} < \cdots < p_{(15)}$ and check if $p_{(n-i+1)} \leq \left(\frac{\gamma}{i} + \frac{1-\gamma}{n}\right)\alpha$;
2. If $p_{(15)} \leq \left(\frac{\gamma}{15} + \frac{1-\gamma}{15}\right)\alpha \approx (0.067) \times 0.05$, reject all hypotheses and stop. Else continue;
3. If $p_{(14)} \leq \left(\frac{\gamma}{14} + \frac{1-\gamma}{15}\right)\alpha \approx (0.067 + 0.0048\gamma) \times 0.05$, reject all hypotheses $H_{(1)}, \ldots, H_{(14)}$ and stop. Else continue;
4. …
5. If $p_{(n-i+1)} \leq \left(\frac{\gamma}{i} + \frac{1-\gamma}{n}\right)\alpha$, reject all hypotheses $H_{(1)}, \ldots, H_{(i)}$ and stop. Else continue;
6. …

### 3.3.1.4 Clinical trials with tree-structured gatekeepers

The **tree gatekeeping methods** serve as a unified framework that includes serial and parallel methods as well as a combination of serial and parallel methods with logical restrictions.

This framework is quite general and can be used to address multiplicity issues in a wide variety of clinical trial applications.

#### 3.3.1.4.1 General framework

Within the tree gatekeeping framework, gatekeepers are defined at the hypothesis rather than family level, i.e., a hypothesis in a certain family may be testable whereas another hypothesis in the same family may not.

Mathematical notations

- $H_{ij}$ is a hypothesis in a family $\mathcal{F}_i$, where $i = 2, \ldots, m; j = 1, \ldots, n_i$;
  - ✓ $R_{ij}^S$ denotes serial rejection set includes hypotheses from $\mathcal{F}_1$ to $\mathcal{F}_{m-1}$;
  - ✓ $R_{ij}^P$ denotes parallel rejection set includes hypotheses from $\mathcal{F}_1$ to $\mathcal{F}_{m-1}$ (do not overlap with $R_{ij}^S$).

The $H_{ij}$ is testable if:

- All hypotheses are rejected in $R_{ij}^S$ and at least one hypothesis is rejected in $R_{ij}^P$.

The following two conditions hold:

$$\max_{k,l \in R_{ij}^S} \tilde{p}_{kl} \leq \alpha \quad \text{and} \quad \min_{k,l \in R_{ij}^P} \tilde{p}_{kl} \leq \alpha.$$

Tree gatekeeping procedures simplifies to serial gatekeeping procedures if $R_{ij}^S = \mathcal{F}_{i-1}$ and $R_{ij}^P$ is empty for all hypotheses $H_{ij},\ i = 2, \ldots, m$, and to parallel gatekeeping procedures if $R_{ij}^S$ is empty and $R_{ij}^P = \mathcal{F}_{i-1}$ for all hypotheses $H_{ij},\ i = 2, \ldots, m$.

The tree gatekeeping methodology was motivated by multiple testing problems that arise in trials when decision trees include multiple branches and/or logical restrictions, e.g.,

**Commented [YL29]:** 实际应用的例子?

- Clinical trials with complex hierarchically ordered hypotheses, e.g., hypotheses associated with multiple endpoints (primary, secondary and tertiary) and multiple test types (noninferiority and superiority), e.g., a hypertension clinical trial with multiple endpoints and noninferiority/superiority tests;
- Dose-finding studies with multiple endpoints and logical restrictions, e.g., a Type II diabetes clinical trial with a primary and two secondary endpoints and the metformin-rosiglitazone combination therapy trial that included a comparison of two metformin-rosiglitazone regimens to metformin on several endpoints.

**Example**

$\mathcal{F}_1 = \{H_{11}, H_{12}, H_{13}\}$;
$\mathcal{F}_2 = \{H_{21}\}$;
$R_{21}^S = \{H_{11}\}$;
$R_{21}^P = \{H_{12}, H_{13}\}$.

The hypothesis $H_{21}$ can be tested only if there is a significant result in $R_{21}^S$ and at least one significant result in $R_{21}^P$ as shown in Figure 23.

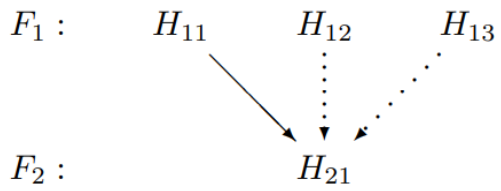



Figure 24 - Tree gatekeeping procedure in a two-family problem. A solid line is used to define a "serial" connection and dotted lines are used for "parallel" connections.

### 3.3.1.4.2 Closure-based tree gatekeeping procedures

Unlike parallel gatekeeping procedure, Bonferroni tree gatekeeping procedures do not, in general, have a straightforward stepwise form. To define a tree gatekeeping procedure, one needs to utilize the closure principle and use a weighted Bonferroni test for each intersection hypothesis in the closed family associated with the $m$ families of interest.

Dmitrienko, Tamhane and Liu and Kordzakhia et al. (Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;, 2010) derived a weight assignment algorithm that satisfies the monotonicity condition. This algorithm is given in the Appendix 5.6.

Dmitrienko, Tamhane and Liu (Alex Dmitrienko; Ajit C Tamhane; Lingyun Liu; Brian L Wiens;, 2008) defined a general approach to defining a broad family of tree gatekeeping procedures that includes Bonferroni tree gatekeeping procedures as a special case. This approach is based on combining multiple tests across families of hypotheses and enables clinical trial sponsors to set up powerful procedures that take into account complex logical restrictions. Examples include tree gatekeeping procedures based on the Hochberg or Dunnett tests.

**Example**

This example involves six hierarchically ordered null hypotheses grouped into four families.

- Family $\mathcal{F}_1$ includes $H_{11}$ (noninferiority hypothesis for A versus B).
- Family $\mathcal{F}_2$ includes $H_{21}$ (superiority hypothesis for A versus B) and $H_{22}$ (noninferiority hypothesis for A+B versus B).
- Family $\mathcal{F}_3$ includes $H_{31}$ (superiority hypothesis for A+B versus B) and $H_{32}$ (noninferiority hypothesis for A+B versus A).
- Family $\mathcal{F}_4$ includes $H_{41}$ (superiority hypothesis for A+B versus A).

A decision tree associated with this testing strategy is displayed in Figure 24.

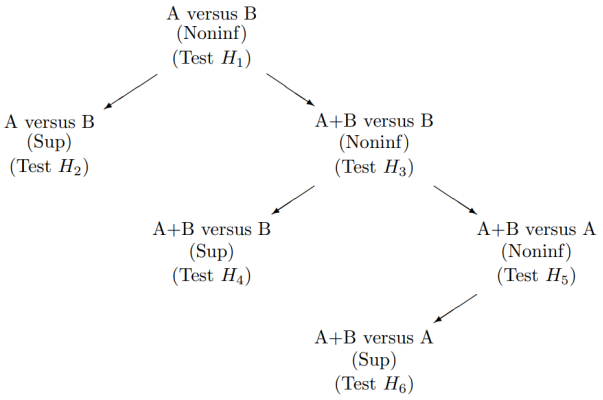



Figure 25 - Decision tree in the combination-therapy clinical trial example (Noninf, Noninferiority; Sup, Superiority).

Now, to account for the logical restrictions among the six hypotheses (the restrictions are displayed in), the serial rejection sets are given by

$$\begin{aligned} R_{21}^S &= R_{22}^S = \{H_{11}\}, \\ R_{31}^S &= R_{32}^S = \{H_{22}\}, \\ R_{41}^S &= \{H_{32}\}. \end{aligned}$$

and the parallel rejections sets are empty. For an example, $H_{31}$ becomes testable if all hypotheses

(which is actually $H_{22}$) in $R_{31}^{S}$ are rejected and at least one hypothesis in $R_{31}^{P}$ is rejected.

A Bonferroni tree gatekeeping procedure based on the algorithm defined in the Appendix 5.6 will be used to control the FWER at the two-sided 0.05 level. The adjusted p-values produced by this tree gatekeeping procedure are listed in Table 10.

Table 10 - Bonferroni tree gatekeeping procedure in the combination-therapy clinical trial example. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level.

| Family | Hypothesis | Raw $p$-value | Adjusted $p$-value |
|---|---|---|---|
| $F_1$ | $H_{11}$ | 0.011 | $0.011^*$ |
| $F_2$ | $H_{21}$ | 0.023 | $0.046^*$ |
| $F_2$ | $H_{22}$ | 0.006 | $0.012^*$ |
| $F_3$ | $H_{31}$ | 0.018 | $0.046^*$ |
| $F_3$ | $H_{32}$ | 0.042 | 0.084 |
| $F_4$ | $H_{41}$ | 0.088 | 0.088 |

The table shows that:

- $H_{11}$, is rejected at the two-sided 0.05 level and thus the hypotheses $H_{21}$ and $H_{22}$ become testable.

⇓

- Both of $H_{21}$ and $H_{22}$ are also rejected and, since $H_{22}$ is included in the serial rejection sets of the hypotheses in $\mathcal{F}_3$, the tree gatekeeping procedure tests $H_{31}$ and $H_{32}$ at the next step.

⇓

- The adjusted p-value for $H_{31}$ is significant but the adjusted p-value for $H_{32}$ is not. Since the hypothesis $H_{41}$ depends on $H_{32}$, the former is retained without testing.

## 3.3.2 Graphical approaches

For clinical trials with multiple treatment arms or endpoints a variety of sequentially rejective, weighted Bonferroni-type tests have been proposed, such as gatekeeping procedures, fixed sequence tests, and fallback procedures.

Since these procedures rely on the ***closed test principle***, they usually require the explicit specification of a large number of intersection hypotheses tests. The underlying test strategy may therefore be difficult to communicate.

Frank Bretz et al. proposed a simple iterative **graphical approach** to construct and perform such Bonferroni-type tests (Bretz, Frank, Maurer, Willi, Brannath, Werner, & Posch, Martin, 2009). The resulting multiple test procedures are represented by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses.

### 3.3.2.1 The definition of graphical approach

The figure defines both:

- a test for the global intersection hypothesis in the full closure through the initial allocation of the significance level $\alpha_1$, $\alpha_2$ to the individual hypotheses;
- a sequentially rejective multiple test procedure (since after rejecting, for example, $H_1$, only $H_2$ remains to be tested).

In this sense the Figure 25 defines an iterative graph for the weighted Bonferroni–Holm procedure.

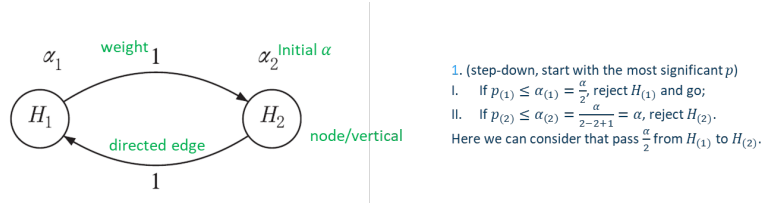
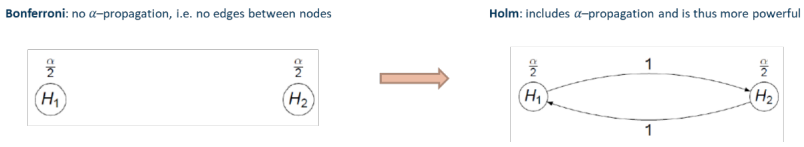
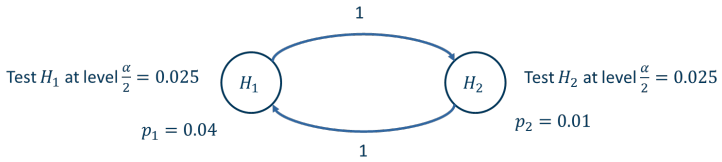


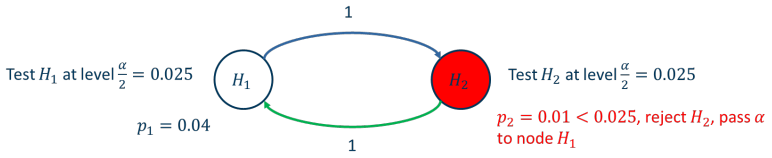

Figure 26 - Graphical illustration of the weighted Bonferroni–Holm procedure with two hypotheses.

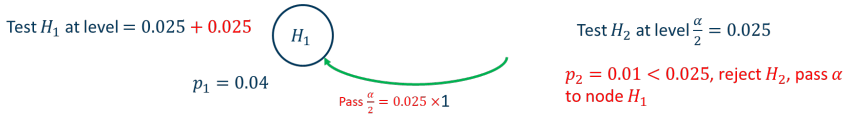
A vivid demonstration is provided below **(m=2, $\alpha$=0.05)**:

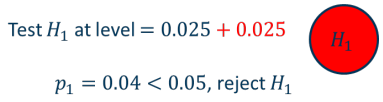





Since Bonferroni-Holm is a step-down procedure, it will start with the most significant p-value which is $p_2$:




**Formal Definition**

- Let initial levels $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ with $\sum_{i=1}^{m} \alpha_i = \alpha \in (0,1)$
- $m \times m$ transition matrix $\boldsymbol{G} = (g_{ij})$ where $g_{ij}$ is the fraction of the level of $H_i$ that is propagated to $H_j$ with $0 \le g_{ij} \le 1$, $g_{ii} = 0$, and $\sum_{j=1}^{m} g_{ij} \le 1, \forall i = 1, \dots, m$
- $(\boldsymbol{G}, \boldsymbol{\alpha})$ defines a directed graph with an associated multiple test

To illustrate the proposed graphs, consider the following Figure 26 for an example involving $m = 3$ hypotheses. For the graph we have $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and transition matrix

$$\boldsymbol{G} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}$$

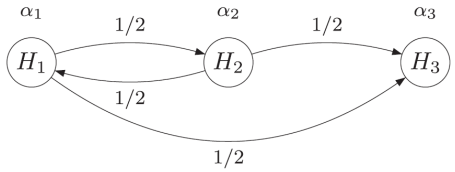



Figure 27 - Example multiple test procedures to illustrate

### 3.3.2.2 Algorithm to update the graph

Let $M = \{1,2,\dots,m\}$, a graph $\mathcal{G} = (\boldsymbol{\alpha}, \boldsymbol{G})$

1. Set $I = M$.
2. Let $j = \arg\min_{i \in I} \frac{p_i}{\alpha_i}$.
3. If $p_j \leq \alpha_j$, reject $H_j$; otherwise stop.
4. Update the graph:

   $I \Leftarrow I \backslash \{j\}$

   $\alpha_\ell \Leftarrow \begin{cases} \alpha_\ell + \alpha_j g_{j\ell}, & \ell \in I \\ 0, & \text{otherwise} \end{cases}$

   $g_{\ell k} \Leftarrow \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & k, \ell \in I; \ell \neq k \\ 0, & \text{otherwise} \end{cases}$
5. If $|I| \geq 1$, go to step 1; otherwise stop.

### 3.3.2.3 Example 1 - Fixed sequence test

Consider first fixed sequence tests, where the test sequence of the hypotheses is fully specified in advance. Each hypothesis is tested at level $\alpha$, where non-rejection at any step renders further testing unnecessary.

The Figure 27 and Figure 28 illustrates the fixed sequence test with three hypotheses, where $H_1$ precedes $H_2$, which in turn precedes $H_3$. Note the initial allocation of the overall significance level $\alpha$ to the individual hypotheses. If, for example, $H_1$ is rejected, the initially allocated significance level (at the vertex $H_1$) is passed on fully to $H_2$ (as indicated by the directed edge with associated weight 1). Accordingly, $\alpha = (\alpha, 0, 0)$ and

$$\mathrm{G} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

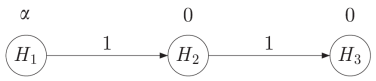



Figure 28 - Graphical illustration of the fixed sequence test with three hypotheses

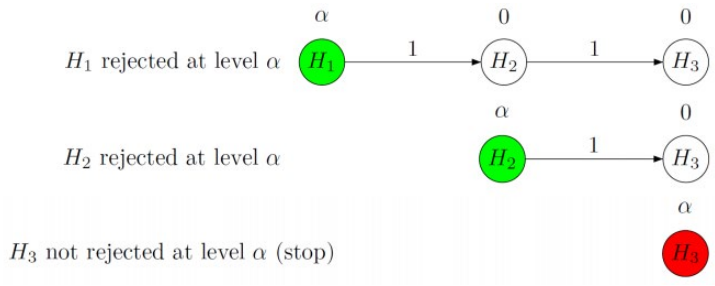



Figure 29 - A break down of graphical illustration of the fixed sequence test; Note that Green=rejection and Red=no rejection (and stop) in this figure.

### 3.3.2.4 Example 2 - Fallback procedure

Commented [YL30]: 和 fixed sequence 去比一下

*Wiens BL.* proposed a modification of the fixed sequence test, which over-comes the dependence on the order of the hypotheses (while sacrificing some power for the individual tests, since they are performed at local significance levels less than $\alpha$). In the notation from Algorithm, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and

$$\boldsymbol{G} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1-r & r & 0 \end{pmatrix}$$

**Simple demonstration**

1. Test $H_1$ at $\alpha_1$. Suppose $p_1 \geq \alpha_1$, not reject $H_1$, retain it;
2. Test $H_2$ at $\alpha_2$. Suppose $p_2 \geq \alpha_2$, not reject $H_2$, retain it;
3. Test $H_3$ at $\alpha_3$. Suppose $p_3 < \alpha_3$, reject $H_3$, pass $r\alpha_3$ to $H_2$ and $(1-r)\alpha_3$ to $H_1$ ;

*Fallback*

4. Test $H_1$ at $\alpha_1 + (1-r)\alpha_3$...
5. Test $H_2$ at $\alpha_2 + r\alpha_3$...

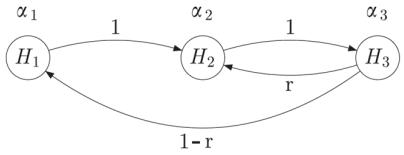



Figure 30 - Improvement of the fallback procedure by Wiens BL, Dmitrienko A. with $r = \alpha_2/(\alpha_1 + \alpha_2)$.

### 3.3.2.5 Example 3 - Bonferroni–Holm procedure

As seen from the Figure 30, $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ where $\alpha_1 = \alpha_2 = \alpha_3 = \frac{\alpha}{3}$ and

$$G = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

fully specify the weighted Bonferroni–Holm procedure. Note that weights other than 0.5 could be used as entries for G, thus generalizing the Bonferroni–Holm procedure.

Assume Bonferroni–Holm procedure with observed p-values $p = (0.02, 0.055, 0.012)$ and overall significance level $\alpha = 0.05$. Figure below displays the resulting sequentially rejective test procedure.

The hypothesis $H_3$ is rejected at the first step, since $p_3 < 0.01667 = \alpha/3$. The associated local significance level $\alpha/3$ is split equally and passed on to the remaining (not yet rejected) hypotheses $H_1$ and $H_2$, as indicated by the directed edges in the left graph in Figure 31.

Based on the algorithm proposed before, we have:

$I \Leftarrow \{1,2,3\}\backslash\{3\}$ where $j = 3$ since $H_3$ is rejected.

$$\alpha_1 \Leftarrow \begin{cases} \alpha_1 + \alpha_1 g_{31} = \alpha_1 + \alpha_1 \times 0.5 = 0.025, \ell = 1 \in I = \{1,2\} \\ 0, \quad \text{otherwise} \end{cases}$$

$$g_{12} \Leftarrow \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}} = \frac{0.5 + 0.5 \times 0.5}{1 - 0.5 \times 0.5} = 1, \quad k = 2 \\ 0, \quad \text{otherwise} \end{cases}$$

Similarly, $g_{21}$ will be updated to 1 and $\alpha_2$ will be updated to 0.025, which is the middle graph.

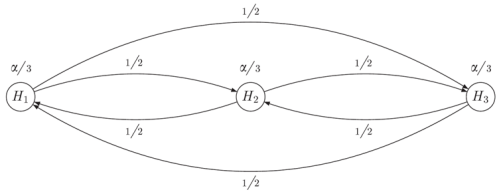



Figure 31 - Graphical illustration of the Bonferroni–Holm procedure with $m$=3 hypotheses and initial allocation

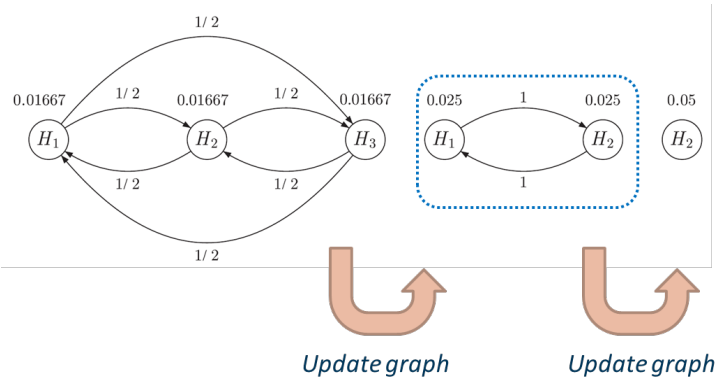



Figure 32 - A vivid demonstration

### 3.3.2.6 Shifting significance levels between families of hypotheses

Consider a situation where families of hypotheses are given and where the rejection of hypotheses in one family is of interest only if all the hypotheses from another family were rejected.

In such cases a multiple test procedure can be applied that allows for a reallocation of the significance level between families of hypotheses.

Such a test strategy can be implemented with graphs that include edges with infinitesimally small weights. Along the vertices with an infinitesimally small weight $\varepsilon$ no significance level is passed.

However, if during the iterative procedure for a vertex only infinitesimal outgoing edges remain, they become non-infinitesimal edges after normalization (such that the sum of outgoing weights becomes one) and can pass the level to other hypotheses.

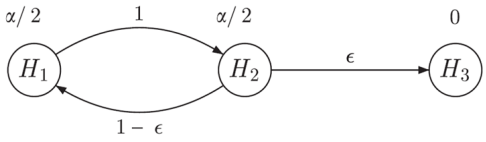



Figure 33 - The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values $p_1 = 0.04$, $p_2 = 0.01$, and $p_3 = 0.03$

#### 3.3.2.6.1 Example 1

As an example consider the test of three hypotheses $H_1$, $H_2$,and $H_3$, where $H_1$, $H_2$ are of primary interest and $H_3$ is of interest only if $H_1$ and $H_2$ can be both rejected.
The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values $p_1 = 0.04$, $p_2 = 0.01$, and $p_3 = 0.03$.
If both $H_1$ and $H_2$ can be rejected, then $H_3$ is tested at level $\alpha = 0.05$.
Additionally, to achieve weights that sum to one, the weight of the edge $H_2 \longrightarrow H_1$ is set to $1 - \varepsilon$ (instead of 1 in the Bonferroni–Holm procedure).

**Simple demonstration**

As $p_2 = 0.01 < \frac{\alpha}{2} = 0.025$, $H_2$ is rejected in the first step and its level $\frac{\alpha}{2} = 0.025$ is shuffled to hypothesis $H_1$, since by the above calculation rules $\lim_{\varepsilon \to 0} \frac{(1-\varepsilon)\alpha}{2} = \frac{\alpha}{2}$ and $\lim_{\varepsilon \to 0} \varepsilon\alpha = 0$.

Now, the node $H_2$ is dropped from the graph and the edges of $H_1$ are updated as shown in the middle graph in Figure 33.

In particular, the edge from $H_1$ to $H_3$ gets the weight $\frac{g_{13}+g_{12}g_{23}}{1-g_{12}g_{21}} = \frac{0+1\times\varepsilon}{1-1\times(1-\varepsilon)} = 1$.

In the second step, $H_1$ is rejected and its level is passed on to $H_3$ that is rejected in the last step.

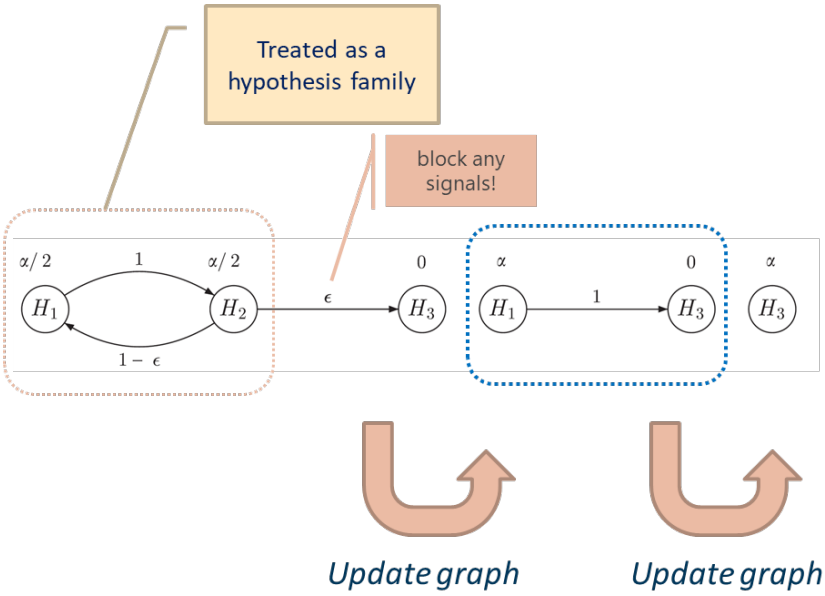



Figure 34 - A vivid presentation of updating a graph with $\epsilon$-edge; The testing starts from hypothesis $H_3$.

### 3.3.2.6.2 Example 2

Assume that $H_3$ and $H_4$ are of interest only if both $H_1$ and $H_2$ are rejected. We wish to perform the Bonferroni–Holm procedure at level for the two hypotheses $H_1$ and $H_2$ of primary interest. The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with parameters

- $\alpha = 0.05,\ r_1 = 0.8,\ r_2 = 0.2$
- the observed p-values $p_1 = 0.04,\ p_2 = 0.01,\ p_3 = 0.03,\ p_4 = 0.04.$

If both hypotheses $H_1$ and $H_2$ are rejected, the significance level $\alpha$ is shuffled to $H_3$ and $H_4$ according to the weights $r_1$ and $r_2$:

- $H_3$ receives $r_1\alpha$ and $H_4$ receives $r_2\alpha$.

If both $H_1$ and $H_2$ can be rejected, then $H_3$ is tested at level $\alpha = 0.05$.

**Simple demonstration**

As $p_2 = 0.01 < \frac{\alpha}{2} = 0.025,\ H_2$ is rejected in the first step and its level $\frac{\alpha}{2} = 0.025$ is shuffled to hypothesis $H_1$, since by the above calculation rules $\lim_{\varepsilon\to 0}\frac{(1-\varepsilon)\alpha}{2} = \frac{\alpha}{2}$ and $\lim_{\varepsilon\to 0}\varepsilon\alpha = 0$. The $\alpha$ level for $H_1$ is updated to $\alpha$.

Node $H_2$ is dropped and the graph will be updated to

- Weight for edge $H_1$ to $H_3$: $g_{13} = \frac{g_{13}+g_{12}g_{23}}{1-g_{12}g_{21}} = \lim_{\varepsilon\to 0}\frac{0+1\times r_1\varepsilon}{1-1\times(1-\varepsilon)} = r_1$
- Weight for edge $H_1$ to $H_4$: $g_{14} = \frac{g_{14}+g_{12}g_{24}}{1-g_{12}g_{21}} = \lim_{\varepsilon\to 0}\frac{0+1\times r_2\varepsilon}{1-1\times(1-\varepsilon)} = r_2$

The edges of $H_1$ are updated as shown in the middle graph in Figure 34. Since $H_1$ is rejected at $\alpha = 0.05$, Now $H_3$ will receive the level $r_1\alpha = 0.8 \times 0.5 = 0.4$, $H_3$ will receive the level $r_1\alpha = 0.2 \times 0.5 = 0.1$.

Then the tests within 'family' consists of $H_3$ and $H_4$ will be perform as the right graph illustrating.

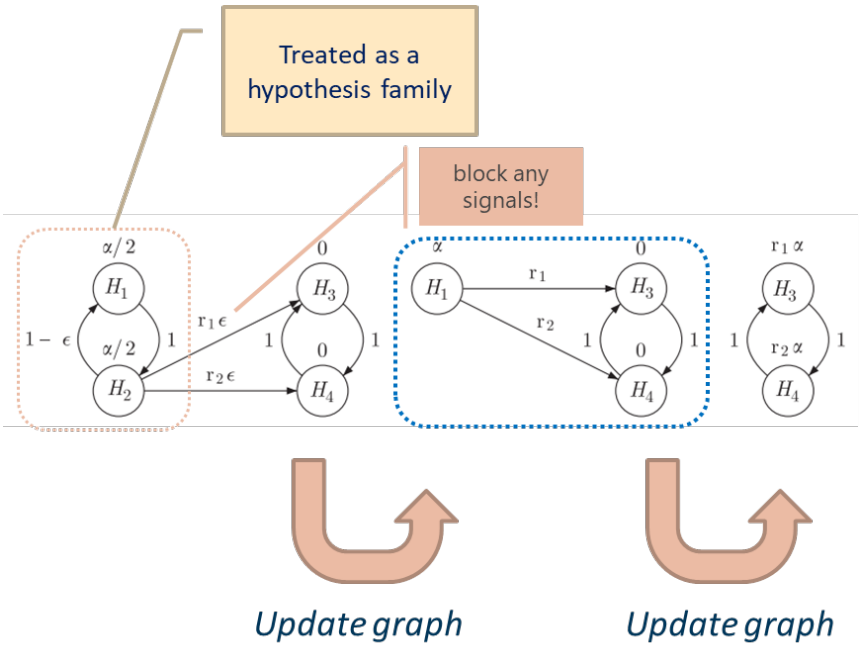



Figure 35 - A vivid presentation of updating a graph with ϵ-edge.

### 3.3.2.6.3 Example 3

The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with parameters Infinitesimal weights can also be used to uniformly improve the gatekeeping procedure (as shown in Figure 36).

- $\alpha = 0.05$
- the observed p-values $p_1 = 0.02,\ p_2 = 0.04,\ p_3 = 0.01,\ p_4 = 0.015$.

***Simple demonstration***

As $p_1 = 0.02 < \frac{\alpha}{2} = 0.025$, $H_1$ is rejected in the first step and its level $\frac{\alpha}{2} = 0.025$ is shuffled to hypotheses $H_3$ and $H_4$, which are both now assigned level $\frac{\alpha}{4} = 0.0125$.

Next, $H_3$ is rejected and passes the level on to $H_4$, which then is rejected. If we use the above gatekeeping procedure, the testing will stop. Although $H_3$ and $H_4$ are both rejected, the significance level cannot be shuffled to $H_2$ (which has not been rejected yet) since a corresponding edge is missing.

*Thus, the gatekeeping procedure can be improved by adding $\varepsilon$-edges from $H_3$ to $H_1$ and from $H_4$ to $H_2$ (see the below figure).*

The only outgoing edge from $H_4$ after the rejection of $H_1$ and $H_3$ is the $\varepsilon$-edge to $H_2$ and is thus assigned the weight 1.

After rejecting $H_4$ the level is passed to $H_2$, which then can also be rejected. In this numerical example we can therefore **reject all 4 hypotheses** with the improved procedure.

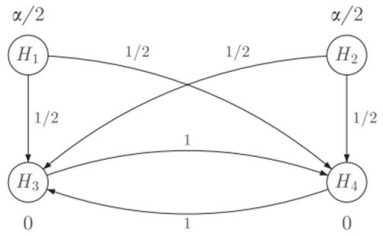



Figure 36 - Graphical illustration of the gatekeeping procedure with four hypotheses.

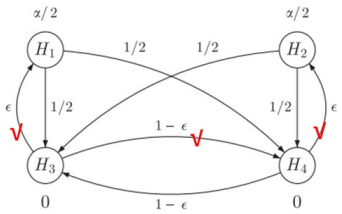



Figure 37 - Graphical illustration of the improved gatekeeping procedure with four hypotheses by adding $\varepsilon$-edge. The tests start from hypothesis $H_1$ and the $\varepsilon$-edge can be treated as a "blocker".

### 3.3.2.7 Late phase development of a new drug for the indication of multiple sclerosis

**The primary objective of the study**

To compare two dose levels of the new drug with a control treatment for three:

- annualized relapse rate
- number of lesions in the brain
- disability progression

We have therefore six elementary hypotheses:

$H_{ij}: \theta_{ij} \leq 0$ where

$\theta$ denotes the mean difference of treatment with different dose level and control
$i = 1,2$ (1-high dose; 2-low dose)
$j = 1,2,3$ for 3 hierarchically-ordered endpoints
In the following we describe several test strategies and use the graphical tools developed in this article to visualize them.
It is NOT the purpose to recommend one strategy, since each has its advantages and disadvantages. The following discussion is rather meant to demonstrate the flexibility of Bonferroni-based closed test procedures and the need to understand the study objectives well in order to propose a reasonable test strategy with good operational characteristics (i.e. high probability of success for the study).

### 3.3.2.7.1 Strategy 1

Consider a fixed sequence test to the six hypotheses being and to test each hypothesis sequentially at level $\alpha$.
The sequence H11→ H21→ H12→ H22→H13→ H23 is a reasonable possibility, see Figure 37.


Figure 38 - A fixed sequence test to the six hypotheses.

In practice such a strategy is often not recommended because of the inherent risk to stop too early. If, for example, the observed p-value for H11 is larger than $\alpha$, none of the subsequent hypotheses can formally be rejected, even if their p-values are very small.

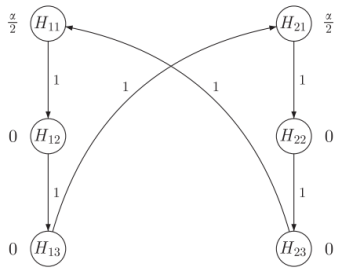
### 3.3.2.7.2 Strategy 2

Consider an alternative approach that avoids stopping too early if the hypotheses corresponding to the first dose cannot be rejected is to group the six elementary hypotheses according to the dose into the two families

- $\mathcal{F}_1$= {H11, H12, H13} – high dose level
- $\mathcal{F}_2$ = {H21, H22, H33} – low dose level

Assuming that there is the wish to reject the secondary (tertiary) endpoint for dose i=1,2 iff the associated primary (primary and secondary) endpoints were rejected before.

Within each family the endpoints are tested in a fixed sequence at 'Bonferronized' Level $\alpha/2$. If for any dose level the three related null hypotheses can be rejected, the fixed sequence for the other dose level can be conducted at level $\alpha$.




The Figure 38 shows a modification of this test strategy that puts more weight on the hypotheses corresponding to the endpoints in the primary positions of the hierarchy. After each rejection, the level is split between the two families and allocated to the first endpoint in each family that has not been rejected so far. If all the hypotheses are rejected in a family, the total level is allocated to the other family.

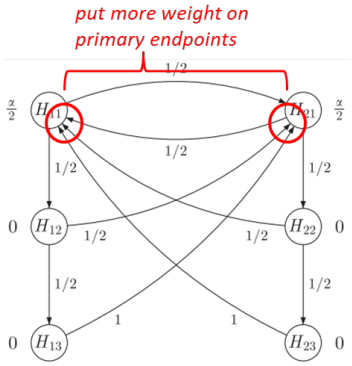



Figure 39 - Visualization of different implementations for Strategy 2.

### 3.3.2.7.3 Strategy 3

Consider In some situations it may be reasonable to order the dose levels, for example, because of safety concerns or because the higher dose level i=1 is expected to have a larger treatment effect than the lower dose level i=2.

Such assumptions then may lead to different families of hypotheses than considered previously. If one is indeed willing to assume $\theta_{1j} > \theta_{2j}$ for all $j$, it seems natural to start testing the high dose for the primary endpoint at level.

**If H11 was rejected, the question is then whether one can argue that H12 is more important than H21 (or vice-versa)?** It will lead to a fixed sequence as discussed in Strategy 1, or whether both hypotheses H12 and H21 are equally important.

In the latter case, this would lead naturally to the set of families:

- F1 = {H11}

- F2 = {H12, H21}
- F3 = {H13, H22}
- F4 = {H23}

where Fi precedes Fj for $1 \leq i < j \leq 4$.

E.g. one could test F3 only if at least one hypothesis from F2 was rejected. An alternative approach is to test F3 only, if both hypotheses from F2 were rejected, as visualized in Figure 39.

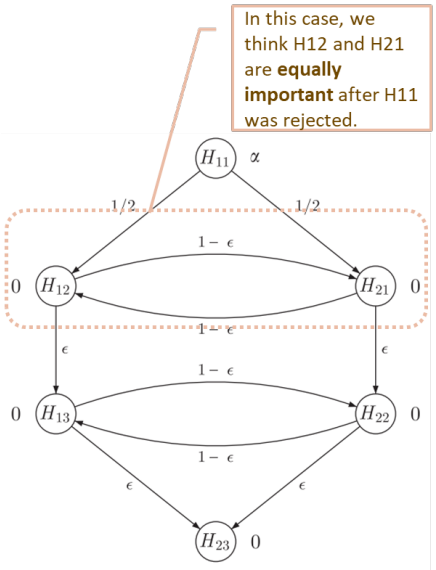



Figure 40 - Visualization of implementations for Strategy 3.

## 3.3.3 Gatekeeping approach versus Graphical approach

The paper *Overview Of Multiple Testing Methodology And Recent Development In Clinical Trials* (Deli Wang, et al., Overview of multiple testing methodology and recent development in clinical trials, 2015) provides an application oriented and comprehensive overview of commonly used multiple testing procedures and recent developments in statistical methodology in multiple testing in clinical trials.

Commonly used multiple testing procedures are applied to test non-hierarchical hypotheses[G] and gatekeeping procedures can be used to test hierarchically ordered hypotheses while controlling the overall type I error rate.

The recently developed graphical approach has the flexibility to integrate hierarchical[H] and non-hierarchical procedures into one framework. A graphical multiple testing procedure with "no-dead-end" provides an opportunity to fully recycle $\alpha$ across hypothesis families.

**Non-hierarchical hypotheses**

- Non-parametric and semi-parametric procedures
  - ✓ Bonferroni procedure
  - ✓ Simes procedure
  - ✓ Holm step-down procedure
  - ✓ Hochberg step-up procedure
  - ✓ Hommel procedure
- Parametric procedures
  - ✓ Dunnett procedure

**Hierarchical hypotheses**

- Simple procedures for hierarchical hypotheses
  - ✓ Fixed-sequence procedure
  - ✓ Fallback procedure
- Gatekeeping procedures
  - ✓ Serial gatekeeping procedures
  - ✓ Parallel gatekeeping procedure

[G] We categorize multiple testing procedures for non-hierarchical hypotheses as "non-hierarchical" multiple testing procedures which include commonly used procedures such as the Bonferroni procedure, the Holm procedure, the Simes based procedures (Hochberg and Hommel procedures) and the Dunnett procedure.

[H] The hypotheses are grouped into ordered families. The hypotheses in higher ordered families are tested first. The hypotheses in lower ordered families are tested only if at least one null hypothesis is rejected in higher ordered families. In other words, the higher ordered families serve as gatekeepers for lower ordered ones. Suppose there is a trial with multiple endpoints (primary, secondary, tertiary, etc.) and multiple doses. It may be meaningful to test all doses on the primary endpoint and carry on to secondary and/or tertiary endpoints only for the doses that are significant for the primary endpoint. In this case, the endpoint serves as the group factor for the families. Gatekeeping procedures are designed to handle multiple hierarchical families of hypotheses.

- ✓ Other extensions of gatekeeping procedures

**Integrate non-hierarchical and hierarchical hypotheses**

- Graphical approaches

## 3.3.4 Adaptive Designs and Confirmatory Hypothesis Testing

There are obvious reasons for inspecting accumulating information while a clinical trial is in progress. Ethical considerations in studies with human subjects and economic issues, measured in terms of time, money and the number of patients available for future studies, are the most prominent ones.

Many of them lead to the classical question of a sequential design: at what point during the course of a study does sufficient evidence accumulate, in favor of or against the test treatment, for discontinuation to be justified?

Statisticians participating in designing clinical trials often are also confronted with questions from their clinical team members like:

- Why can't we do an interim analysis and, depending on the results, not only stop a trial for proven efficacy or evident futility, but
  - ✓ stop or delete one or more of the treatment regimens?
  - ✓ change treatment regimens?
  - ✓ change inclusion or exclusion criteria?
  - ✓ change the primary endpoint of efficacy?
  - ✓ recalculate the sample size?

Such questions arise naturally and should be carefully considered at the planning stage. In the not-so-distant past the common answer to them was: by such interventions the Type I error rate will be altered, estimates will be biased and no valid statistical methods to deal with them appropriately exist. However, in the last decade much progress has been made to devise statistical methods for adaptive designs.

Statistical inferences based on this novel methodology for adaptive designs allows implementation of design adaptations without inflating the Type I error rate. These adaptations may be based on the unblinded data collected up to an interim analysis as well as external information and the adaptation rules do not need to be specified in advance — an indispensable prerequisite to cope with the unexpected.

Whereas in early phases of drug development control of the Type I error rate may not be a high priority, it always helps in the interpretation of data. Control of Type I error rate is, however, of utmost importance if adaptive methods are to be applied, for example in confirmatory drug development or when combining Phase II and III trials in a combination Phase II/III trial (also known as adaptive seamless trials or confirmatory stagewise adaptive trials).

### 3.3.4.1 Causes of multiplicity and bias in adaptive designs

The principal differentiation of adaptive designs compared to traditional fixed designs is the ability to perform interim analyses in order to take decisions affecting the further conduct of the trial.

This leads to repeatedly testing of one or multiple hypotheses and the possibility to change design features based on interim data. Since the same interim data is subsequently used for hypothesis testing and estimation such approaches may cause bias in estimation and inflation of the Type I error rate if not adequately controlled.

The different sources of bias and basic methods for respective adjustments are listed in Table 11.

Table 11 - Sources and control of Type I error rate inflation.

| Sources of potential error rate inflation | Techniques for error rate control |
|---|---|
| Repeated hypothesis testing with early rejection of null hypotheses at an interim analysis. | Classical group sequential designs, e.g., designs based on the $\alpha$-spending approach. |
| Adaptation of design and analysis features with combination of information across trial stages, e.g., sample size re-assessment based on the treatment effect, data-driven changes in the timing of the next interim analysis or testing strategy. | Combination of $p$-values, e.g., the inverse normal method, Fisher's combination test, conditional rejection probability, adjustment for known adaptation rule. |
| Multiple hypothesis testing, e.g., adaptive selection among initial hypotheses at an interim analysis. | Classical multiple testing methods, e.g., appropriate tests for intersection hypotheses together with closed procedures. |

### 3.3.4.2 Repeated hypothesis testing at interim analyses in group sequential designs

Repeated testing of hypotheses occurs in trial designs that foresee interim analyses of accrued data together with formal testing of one or several hypotheses together with the possibility of early rejection or retention of the hypotheses.

It is well known that repeated testing of a particular hypothesis (without adjusting the significance level) inflates the Type I error rate. It may also be deflated if interim analyses allow for retention of the null hypothesis and the significance levels of the respective tests are not adjusted.

Commented [YL31]: ?

The design of a trial is called "**group sequential**" when stopping is only foreseen after having accrued additional data of groups of patients and not just single patients. There exists a vast literature on technical and operational aspects of such trials.

We will introduce here only the basic concepts and notations that are needed in the context of the

generalization of group sequential designs to adaptive designs, i.e., designs allowing also interim decisions other than stopping or continuing the trial with an otherwise unaltered design.

### 3.3.4.2.1 Basic concepts and notations for group sequential designs

**Assumptions**

- a parallel group 2-arm trial with $k > 0$ planned interim analyses, including the final analysis;
- the treatment effect in comparison to a control denoted by the single parameter of interest, $\theta$, that can take on any real value;
- there is only one null hypothesis $H: \theta \leq 0$ versus alternative hypothesis $K: \theta > 0$;
- the amount of statistical information available at interim analysis $t, t = 1, \dots, k$ is $I_t$;
- the respective test statistics, e.g., for comparing a test treatment to a control, taking into account all data up to analysis t is denoted by $Z_t, t = 1, \dots, k$;
- in the most common case, $Z_t$'s follow (asymptotically) a multivariate normal distribution (MVN) with $E[Z_t] = \theta\sqrt{I_t}$ and $\text{Cov}(Z_t, Z_{t'}) = \sqrt{I_t/I_{t'}}, 1 \leq t \leq t' \leq k$.

**Concepts**

**stopping boundaries**: a threshold for summary statistics of the accruing data determining whether the trial should be stopped (and $H$ be either rejected or retained) or whether the trial should continue;

A pair of hypotheses ($H$ and $K$) and respective stopping boundaries can always be translated (at least asymptotically) into probabilities of errors "spent" up to a certain interim or the final analysis.

### 3.3.4.2.2 Stopping[I] boundaries

**At the first interim analysis $t$ ($t = 1$)**

the test statistic $Z_1$ is compared to lower and upper "stopping boundaries" $l_1$ and $u_1$, respectively. If $Z_1 \leq l_1$, the trial stops and the null hypothesis $H$ is retained (or equivalently "futility" is declared). If $Z_1 \geq u_1$, $H$ is rejected in favor of $K$ and the trial is also stopped. If $Z_1 \in (l_1, u_1)$ the trial continues to the next planned interim or final analysis.

**At the interim analysis $t$ ($t < k$)**

If $Z_t \leq l_t$, the trial stops and the null hypothesis $H$ is retained (or equivalently "futility" is declared). If $Z_t \geq u_t$, $H$ is rejected in favor of $K$ and the trial is also stopped. If $Z_t \in (l_t, u_t)$ the trial continues to the next planned interim or final analysis.

**At the final analysis $t$ ($t = k$)**

If the trial is not stopped at any of the interim analyses, a final test is done with $Z_k$ being

[I] Though stopping for futility is foreseen as an option in almost all group sequential trials, it is not necessarily formally dependent on stopping boundaries, but can be decided upon by an independent data monitoring committee (IDMC).

compared to the decision thresholds $l_k = u_k$. In this case $H$ is either retained or rejected.

A futility assessment may make use of conditional or predictive probabilities of success and/or emerging trends in data besides of the primary parameters of efficacy, e.g., with regards to safety.

In settings where no stopping rule for futility is pre-specified, such stopping cannot inflate the Type I error rate but decreases power.

In any case, if the decision rules (boundaries) are pre-specified they need to be defined such that the overall Type I error rate, i.e., the probability to reject $H$ at any of the interim or at the final analysis, is guaranteed not to exceed a predefined level $\alpha$, e.g., one-sided $\alpha = 0.025$.

The probability to reject the null hypothesis at interim analysis $t, t = 1, \dots, k$, given a true treatment $\theta$ is

$$\alpha_t(\theta) = P_\theta\left(\{Z_t \geq u_t\} \cap \bigcap_{s=1}^{t-1}\{l_s < Z_s < u_s\}\right).$$

The probability to stop exactly at analysis $t$ and retain $H$ can be similarly expressed by

$$\beta_t(\theta) = P_\theta\left(\{Z_t \leq l_t\} \cap \bigcap_{s=1}^{t-1}\{l_s < Z_s < u_s\}\right).$$

When $\theta = 0$, $\alpha_t(0)$ is called

- the $\alpha$ level spent at interim analysis $t$.

Under the alternative hypothesis $K$ with $\theta = \Delta > 0$, $\beta_t(\Delta)$ is called

- Type II error rate spent at interim decision $t$.

**Overall Type I error rate $\boldsymbol{\alpha(0)}$** is given by

$$\alpha(0) = \sum_{t=1}^{k} \alpha_t(0),$$

**Overall Type II error rate $\boldsymbol{\beta(\Delta)}$** is given by

$$\beta(\Delta) = \sum_{t=1}^{k} \beta_t(\Delta).$$

It should be noted that, if at the final analysis a decision is taken with respect to rejection or retention of $H$, i.e., if $l_k = u_k$, then $\alpha(0) = 1 - \beta(0)$. These spent levels are not to be confused with the "nominal" decision levels.

For a given number of interim analyses, there is a large choice of "standard" types of boundaries, ranging from those that make an early rejection relatively difficult (**O'Brien-Fleming-type boundaries**) to those with equal rejection levels at equally spaced interim analyses (**Pocock-type boundaries**).

**Commented [YL32]:** 不预先设置停止条件，则不会导致 type I error 增长，但会导致 power 下降

**Commented [YL33]:** 如果预先设置了决策边界 for futility，则需要预先定义一个 type I error rate。任何中期分析或者最终分析都不能超过这个值

**Commented [YL35]:** 待补充 O'Brien 等方法在 appendix

### 3.3.4.3 Group sequential Holm procedure with multiple primary endpoints

Yining Ye et al. (Yining Ye; Ai Li; Lingyun Liu; Bin Yao;, 2013) proposed a **group sequential Holm procedure** when there are multiple primary endpoints. This method addresses multiplicities arising from multiple primary endpoints and from multiple analyses in a group sequential design. **The group sequential Holm procedure is shown to be a closed testing procedure and controls the FWER in a strong sense when multiplicities arise from both multiple analyses over time and from multiple endpoints in a group sequential setting.**

We consider multiple primary endpoints in the context of group sequential designs where the objective is to seek regulatory approvals on **at least one of the primary endpoints**.

It is worth noting that

- the method is not expected to have a power advantage for rejecting at least one hypothesis;
- The proposed method is more powerful than the parallel group sequential method and avoids the need to prespecify a test order as in the fixed sequence approach;
- 在 IA 中也可以回收 alpha 传递给 B(or A) IA，unblinding consideration），GSD 的 boundary 是需要动态调整的。
- The GSHv procedure recycles $\alpha$ from a rejected hypothesis to all stages of the group sequential boundaries of the unrejected hypotheses, thus modifying their entire boundaries. The GSHf procedure recycles $\alpha$ from rejected hypotheses only to the final stages of the group sequential boundaries of the unrejected hypotheses, thus modifying only their final stage critical constants.
- for the group sequential Holm procedures, the study can only stop at the interim analysis when all primary hypotheses are rejected;
- the proposed method ignores the correlation among the endpoints. In most clinical settings, it is difficult to justify that a certain degree of correlation can be obtained reliably among endpoints. When it is possible to quantify the correlation (e.g., the biomarker subpopulation case) or when a bound on the correlation can be estimated, further gains on efficiency may be achieved. In these situations, additional research is needed to incorporate the correlation into the proposed procedure.
- 可适用于两种情形：
    - it may be desirable to conduct a global oncology trial with both OS and PFS as primary endpoints [20] so testing of OS does not depend on the outcome of PFS. I
    - a trial to investigate as primary objectives the treatment effects in both the overall population and the biomarker subpopulation [21,22].

**Commented [YL36]:** 没有在同一次 IA 中的多个终点之间 alpha recycle 的过程; $\alpha$ is split between the two endpoints each with independent group sequential procedures and no $\alpha$ reallocation

**Commented [TT37]:** 这个也很重要

#### 3.3.4.3.1 Methodology

Consider a clinical trial to assess the treatment effect on either A or B or both:

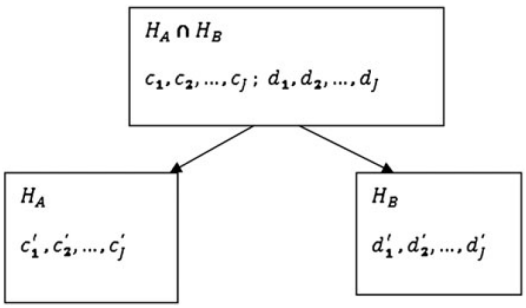
- Two primary endpoints denoted by A and B;
- $J$ times of interim analyses and $j\ (j = 1, \dots, J)$ is for the specific analysis;
- **Wald-statistics $X_j$** for A and $Y_j$ for B based on cumulative data;
- Null hypothesis $H_A$ $(H_B)$ of no treatment effect on the endpoint A (B);
- $\alpha$ allocation: $\alpha_A = w_A\alpha$ and $\alpha_B = w_B\alpha$, where $w_A + w_B = 1$;
- Group sequential boundaries:
  i. $c_j$ for A at significance level $\alpha_A$, and $c_j'$ for A at significance level $\alpha$. Based on consonant property, there is $c_j \geq c_j'$.
  ii. $d_j$ for B at significance level $\alpha_B$, and $d_j'$ for B at significance level $\alpha$; Based on consonant property, there is $d_j \geq d_j'$.

  The boundaries satisfy the following equations (see the definition of union-intersection testing in Section 2.1):

$$\begin{aligned} P\left(\cup_{j=1}^{J}\{X_j > c_j\}\right) &= \alpha_{\mathrm{A}} \\ P\left(\cup_{j=1}^{J}\{X_j > c_j'\}\right) &= \alpha \\ P\left(\cup_{j=1}^{J}\{Y_j > d_j\}\right) &= \alpha_{\mathrm{B}} \\ P\left(\cup_{j=1}^{J}\{Y_j > d_j'\}\right) &= \alpha \end{aligned}$$

- Decisions making:
  - ✓ If either of the two endpoints crossed its corresponding boundary $c_j$ or $d_j$, then the other endpoint can be tested using the full level boundary.
    - ◆ Example: if endpoint A crossed its **level $\alpha_A$ boundary $c_{j^*}$ at some look $j^*$**, then efficacy with respect to endpoint A can be claimed and **its type I error $\alpha_A$ can be reallocated to endpoint B** so that endpoint B can be tested using full level $\alpha$ boundary $d_j'$ **where $j \geq j^*$**.
- The rationale of $\alpha$-reallocation in this procedure
  - ✓ The group sequential procedure strongly controls the type I error rate at level $\alpha$ because it is a closed test. If the intersection hypothesis $H_A \cap H_B$ is rejected, then at least one of the individual hypotheses $H_A$ and $H_B$ will be rejected by the closed test which is illustrated by the following diagram **Figure 40**.
-

**Figure 41 - Closed test.**




**Boundary values for endpoints**

**Commented [TT40]:** 这个需要强调

**Commented [YL41R40]:** 更正：应该是>=不是>，同次 IA 间是可以 alpha reallocation 的

**Commented [YL42R40]:**

**Commented [YL43]:** @Tan Tao 咱们把 CTP 这句话加在这里?

With regard to the boundary values for each endpoint, we can use different methods.

*For example*, we can use O'Brien–Fleming boundaries for endpoint A and Pocock boundaries for endpoint B.

a) After one hypothesis is rejected, one may continue using the predefined interim boundaries with $c_j' = c_j$ or $d_j' = d_j$ for $j < J$. In other words, the boundaries can be left unchanged at the interim analyses except at the final analysis $J$.

b) Alternatively, $c_j'$ or $d_j'$ may be updated with completely different values after the $\alpha$ reallocation.

We term boundary a) as **group sequential Holm fixed (GSHf)** and b) as **group sequential Holm variable (GSHv)**.

It has been shown to be a closed testing procedure and preserves the familywise error rate in the strong sense (see Appendix 5.5 for detailed explanation). The group sequential procedure strongly controls the type I error rate at $\alpha$ level because it is a closed test. To see this, consider the closed family

$$\{H_A \cap H_B, H_A, H_B\}.$$

An $\alpha$ level for global test for $H_A \cap H_B$ is as follows:

- Reject $H_A \cap H_B$ if endpoint A cross its level $\alpha_A$ group sequential boundary $c_j$ $(j = 1, \dots, J)$ at any interim look $j$, or if endpoint B cross its level $\alpha_B$ group sequential boundary $d_j$ $(j = 1, \dots, J)$ at any interim look $j$:

$$P(reject\ H_A \cap H_B) = P\left(\bigcup_{j=1}^{J}\{X_j > c_j\}\ or\ \bigcup_{j=1}^{J}\{Y_j > d_j\}\right)$$
$$\leq P\left(\bigcup_{j=1}^{J}\{X_j > c_j\}\right) + P\left(\bigcup_{j=1}^{J}\{Y_j > d_j\}\right)$$
$$= \alpha_A + \alpha_B = \alpha$$

When multiple endpoints are the only concern without an interim analysis, the method simplifies to the weighted Holm procedure (see Section 3.2.2.1).

The proposed method is more powerful than the parallel group sequential method and avoids the need to anticipate the testing order as in the fixed sequence testing scheme.

We will compare both methods (**GSHf** and **GSHv**) with the naïve approach where $\alpha$ is split between the two endpoints each with independent group sequential procedures and no $\alpha$ reallocation. For ease of reference, we label the naïve approach as **group sequential Bonferroni (GSB)**.

**Low-dimensional**

Scenarios are:

**Commented [YL44]:** 一种情况：某次（j-th）IA 中拒了一个 endpoint，另一个 endpoint 的 boundary 不变。注意，这里 j<J，即：仅 IA 过程（非 final analysis）中保持不变

**Commented [YL45R44]:** 这里的 j<J 仅是 example

**Commented [YL46]:** 另一种情况：某次（j-th）IA 中拒了一个 endpoint，另一个 endpoint 的 boundary 在 alpha 重分配后相应要变化。

**Commented [TT48]:** 能不能稍微展开解释一下？

**Commented [YL49R48]:** 好的，我加在 holm procedure 下

**Commented [YL50R48]:** 添加回复在 appendix

- one interim analysis and one final analysis are planned $J = 2$;
- the nondecreasing functions defined over $t \in [0,1]$ are $\alpha(t), \alpha_A(t)$ and $\alpha_B(t)$; Such that $\alpha(0) = \alpha_A(0) = \alpha_B(0) = 0, \alpha_A(1) = \alpha_A, \alpha_B(1) = \alpha_B, \alpha(1) = \alpha$;
- the information fraction for endpoint A at the interim analysis is $t_{1A}$;
- the information fraction for endpoint B at the interim analysis is $t_{1B}$;
- the information fraction for endpoint A at the final analysis is $t_{2A}$, where $t_{2A} = 1$;
- the information fraction for endpoint B at the final analysis is $t_{2B}$, where $t_{2B} = 1$;
- boundary values $c_1, c_1', c_2, c_2'$ and they are calculated by using Lan-DeMets error spending function.

**Commented [YL51]:** 用 error function 去 approximate 不同的 boundary 值

We can calculate boundary values for endpoint A from the following two equations under the null hypotheses:

$$P(X_1 > c_1) = \alpha_A(t_{1A})$$
$$P(X_1 > c_1') = \alpha(t_{1A})$$
$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2) = \alpha_A(t_{2A}) = \alpha_A$$
$$P(X_1 > c_1') + P(X_1 \le c_1', X_2 > c_2') = \alpha(t_{2A}) = \alpha.$$

- An error spending function that approximates the O'Brien–Fleming boundary is given by

$$\alpha_{A_{OF}}(t_{1A}) = 2\Phi\left(\frac{Z_{1-\frac{\alpha_A}{2}}}{\sqrt{t_{1A}}}\right),$$

Where $\Phi(\cdot)$ is the standard normal cdf and $Z_{1-\frac{\alpha_A}{2}}$ is the $(1-\frac{\alpha_A}{2})$ quantile of the standard normal distribution. Note that $\alpha_A$ is two-sided level.

**Commented [YL53]:** 好像不对啊?

- An error spending function that approximates the Pocock boundary is given by

$$\alpha_{A_{PO}}(t_{1A}) = \alpha_A \ln\{1 + (e-1)t_{1A}\}.$$

When $c_1' = c_1$, that is, the GSHf procedure is preferred, we can calculate the critical boundaries $c_1, c_2, c_2'$ from the following equations under the null hypotheses:

$$P(X_1 > c_1) = \alpha_A(t_{1A})$$
$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2) = \alpha_A(t_{2A}) = \alpha_A$$
$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2') = \alpha(t_{2A}) = \alpha.$$

For both the GSHv and GSHf procedures, similar calculations as mentioned previously can be performed to obtain boundaries $d_1, d_1', d_2, d_2'$.

**Example: MONET1 Study**

We apply the group sequential Holm methods to an actual clinical trial. The MOtesanib Non-Small Cell Lung Cancer Efficacy and Tolerability (MONET1) study was a phase 3, placebo-controlled randomized oncology clinical trial (ClinicalTrials.gov Identifier: NCT00460317).

The primary objectives of this study were

- to determine if motesanib in combination with chemotherapy would improve survival

- ✓ in the overall study population and
- ✓ in subjects with adenocarcinoma histology (adenocarcinoma subpopulation).

The type I error (1-sided 2.5%) was split between

- the overall population (1.5%, one sided)
- the adenocarcinoma subpopulation (1%, one sided).

The study had 80% power requiring 742 deaths in the overall population to detect a hazard ratio of 0.80 (12.5 months v 10 months) for OS with two-sided $\alpha = 0.03(\alpha_A)$ in the patients with non-squamous histology and 80% power (13 months v 10 months) for OS with two-sided $\alpha = 0.02(\alpha_B)$ in the requiring 593 deaths in the adenocarcinoma subpopulation to detect a hazard ratio of 0.77.

A total of 1060 subjects were enrolled including 70% with the adenocarcinoma histology.

An interim analysis was planned when 50% (370 events) of the total deaths occurred in the overall population. The number of deaths for patients with adenocarcinoma histology was also close to the 50% target in the subpopulation at the interim analysis. A negligible amount of type I error (0.00005, one sided) was assigned at the interim for each hypothesis in the original design.

To apply the **GSHv** method, we use the **O'Brien–Fleming spending function**. The critical boundaries can be obtained by solving the following equations (Note that $X_1$ and $X_2$ are the log-rank statistics at interim and final analysis for the overall population):

$$P(X_1 > c_1) = \alpha_A(t_{1A}) = \alpha_A(0.5) = 2\Phi\left(-\frac{Z_{\frac{\alpha_A}{2}}}{\sqrt{t_{1A}}}\right) = 2\Phi\left(-\frac{Z_{0.015}}{\sqrt{0.5}}\right)$$

$$P(X_1 > c_1') = \alpha(t_{1A}) = \alpha(0.5) = 2\Phi\left(-\frac{Z_{\frac{\alpha}{2}}}{\sqrt{t_{1A}}}\right) = 2\Phi\left(-\frac{Z_{0.025}}{\sqrt{0.5}}\right)$$

$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2) = \alpha_A(t_{2A}) = \alpha_A = 0.015$$
$$P(X_1 > c_1') + P(X_1 \le c_1', X_2 > c_2') = \alpha(t_{2A}) = \alpha = 0.025.$$

It can be shown that $c_1 = 3.25$; $c_1' = 2.96$; $c_2 = 2.18$; $c_2' = 1.97$ for the overall population. Similarly, $d_1 = 3.46$; $d_1' = 2.96$; $d_2 = 2.33$; $d_2' = 1.97$ in the adenocarcinoma subpopulation. Rejection region at the interim analysis for this method GSHv is shown in Figure 41 where $\alpha$ is split between the overall population and the adenocarcinoma subpopulation (0.015 and 0.01, one-sided respectively).

if the **GSHf** is used with the same **O'Brien–Fleming spending function**, then critical boundaries can be obtained by solving

$$P(X_1 > c_1) = \alpha_A(t_{1A}) = \alpha_A(0.5) = 2\Phi\left(-\frac{Z_{\frac{\alpha_A}{2}}}{\sqrt{t_{1A}}}\right) = 2\Phi\left(-\frac{Z_{0.015}}{\sqrt{0.5}}\right)$$

**Commented [TT54]:** Simulation needed

**Commented [TT55]:** 1/1.25= 0.8?

**Commented [TT56]:** 在这里表明

$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2) = \alpha_A(t_{2A}) = \alpha_A = 0.015$$
$$P(X_1 > c_1) + P(X_1 \le c_1, X_2 > c_2') = \alpha(t_{2A}) = \alpha = 0.025.$$

It can be shown that $c_1 = 3.25;\ c_2 = 2.18;\ c_2' = 1.96$ for the overall population. Similarly, $d_1 = 3.46;\ d_2 = 2.33;\ d_2' = 1.96$ in the adenocarcinoma subpopulation. Table 12 summarizes the boundary values of the various methods.

Table 12 - Boundary values for the MONET1 trial.

| | Overall population ($\alpha_A = 0.015$) | | Adenocarcinoma subpopulation ($\alpha_B = 0.01$) | |
|---|---|---|---|---|
| Approach[a] | Interim | Final | Interim | Final |
| Original design | 3.89 | 2.17 | 3.89 | 2.32 |
| GSHv | 3.25 (2.96[b]) | 2.18 (1.97[b]) | 3.46 (2.96[b]) | 2.33 (1.97[b]) |
| GSHf | 3.25 | 2.18 (1.96[b]) | 3.46 | 2.33 (1.96[b]) |
| GSB | 3.25 | 2.18 | 3.46 | 2.33 |

[a]Spending function approximating the O'Brien–Fleming boundary is used.
[b]Numbers in parentheses are the boundary values when $\alpha$ is reallocated.
GSHv, group sequential Holm variable method; GSHf, group sequential Holm fixed method; GSB, group sequential Bonferroni.

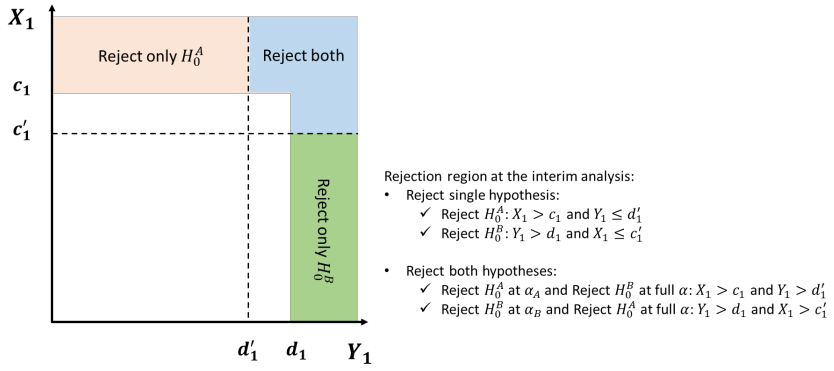



Figure 42 - Rejection region at the interim analysis for MONET1 (GSHv).

### 3.3.4.3.2 Extension of methodology

We can extend the proposed method to the situation with more than two primary endpoints.

**Mathematical notations**

$K$: the number of endpoints of interest, which are also the number of associated hypotheses to be tested $H_1, \dots, H_K$;

$J$: the total number of interim analyses;

$F^{(0)} = \{1,2, \dots, K\}$: the index set of all endpoints;

$w_k^{(0)}$: weight for endpoint $k, k = 1, \dots, K$ where $\sum_k^K w_k^{(0)} = 1$;

$\alpha_k^{(0)}$: type I error rate allocated for $H_k, k = 1, \dots, K$ at each interim analysis, where $\alpha_k^{(0)} =$

$w_k^{(0)}\alpha$;

$Z_{k1}, Z_{k2}, \dots, Z_{kJ}$: Wald statistics to test $H_k, k = 1, \dots, K$ at each interim look;

$c_{k1}^{(0)}, c_{k2}^{(0)}, \dots, c_{kJ}^{(0)}$: group sequential boundary for $H_k, k = 1, \dots, K$ at significance level $\alpha_k^{(0)}$ at each interim look such that

$$P\left(\bigcup_{j=1}^{J}\{Z_{kj} > c_{kj}\}\right) = \alpha_k^{(0)}, \qquad k = 1,2, \dots, K.$$

**Algorithm**

Step 1. Test single hypothesis $H_k, k = 1, \dots, K$

Scenario 1. None of the $K$ endpoints crossed its boundary at any of the $J$ looks, retain all $H_k$ and stop;

Scenario 2. Any one of the $K$ endpoints crossed its boundary at any of the $J$ looks, then efficacy with respect to this endpoint can be claimed.

**Details**

Test $H_k$ with $c_{k1}^{(0)}, c_{k2}^{(0)}, \dots, c_{kJ}^{(0)}$ at significance level $\alpha_k^{(0)}$:

$j^{(1)}$: earliest interim look where at least one endpoint can be rejected;

$k^{(1)}$: set of the $m_1$ endpoints that crossed their boundaries at interim look $j^{(1)}$;

$F^{(1)}$: $F^{(0)}\backslash\{k^{(1)}\}$ which the set of remaining $(K - m_1)$ endpoints.

Commented [YL58]: 也就是被拒掉的 hypothesis 的 index 集合

There are two steps to update significance level and boundaries sequentially.

***α updates***

For $F^{(1)}$, the significance level assigned to all the individual hypotheses $H_q$ $(q \in F^{(1)})$ will be updated to

$$\alpha_k^{(1)} = \alpha_k^{(0)} + \sum_{i \in k^{(1)}} \alpha_i^{(0)} \times \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}} = \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}}\alpha, \quad \left(k \in F^{(1)}\right)$$

Where $w_k^{(1)} \equiv \frac{w_k^{(0)}}{1 - \sum_{i \in k^{(1)}} w_i^{(0)}}$ and $\alpha_k^{(1)} = w_k^{(1)}\alpha$.

⇓

***Boundaries updates***

The boundaries for $H_k$ $(k \in F^{(1)})$ at significance level $\alpha_k^{(1)}$ will be updated using the error spending approach that satisfies the following equation:

$$P\left(\cup_{j=1}^{J}\left\{Z_{kj} > c_{kj}^{(1)}\right\}\right) = \alpha_k^{(1)}, \quad \left(k \in F^{(1)}\right)$$

Step 2. Repeatedly to test single hypothesis $H_k$ $(k \in F^{(i-1)})$ for any interim look $j$ $(j \geq j^{(i-1)})$ using updated boundary values $c_{k1}^{(i-1)}, c_{k2}^{(i-1)}, \dots, c_{kJ}^{(i-1)}$.

Scenario 1. None of the $(K-m_1-\cdots-m_{i-1})$ endpoints crossed its boundary at any of the $j$ look, retain all $H_k\left(k\in F^{(i-1)}\right)$ and stop;

Scenario 2. Any one of the $(K-m_1-\cdots-m_{i-1})$ endpoints crossed its boundary at some interim look $j^{(i)}$ $(j^{(i)}\geq j^{(i-1)})$, then efficacy with respect to this endpoint can be claimed:

**Details**

$k^{(i)}$: set for $m_i$ cross boundaries at interim look $j^{(i)}$;

$F^{(i)}$: $F^{(i-1)}\backslash\{k^{(i)}\}$ which the set of remaining $(K-m_1-\cdots-m_{i-1})$ endpoints.

***α updates***

$$\alpha_k^{(i)}=\alpha_k^{(i-1)}+\sum_{i\in k^{(i)}}\alpha_i^{(i-1)}\frac{w_k^{(i-1)}}{1-\sum_{i\in k^{(i)}}w_i^{(i-1)}}=\frac{w_k^{(i-1)}}{1-\sum_{i\in k^{(i)}}w_i^{(i-1)}}\alpha,\quad\left(k\in F^{(i)}\right)$$

***Boundaries updates***

$$P\left(\cup_{j=1}^{J}\left\{Z_{kj}>c_{kj}^{(i)}\right\}\right)=\alpha_k^{(i)}\quad\left(k\in F^{(i)}\right)$$

For the above algorithm, repeat Step 1 and Step 2 recursively until all endpoints are rejected or complete final analysis. Note that after the allocation of $\alpha$, the boundaries should satisfy

$$c_{kj}^{(i_2)}\leq c_{kj}^{(i_1)},$$

where $i_1<i_2$ and $j=1,2,\ldots,J$.

For GSHf $c_{kj}^{(i)}$ is fixed at $c_{kj}^{(0)}$ for $j<J$, otherwise for GSHv.

**A simple illustration**

Let $K=3,\ J=4$ (3 interim analyses (IA) + 1 final analysis).

A dynamic illustration for the extension of algorithm is shown in Table 13.

Commented [YL60]: 需 double confirm 及 review 正确性

Table 13 - An illustration of the extension of algorithm with three endpoints and three interim analyses.

| | $H_1$ | $H_2$ | $H_3$ | Variables |
|---|---|---|---|---|
| IA1 | Test with $c_{11}^{(0)}$ (calculated based on significance level $\alpha_1^{(0)}$) and cross the boundary value, the endpoint 1 is rejected, efficacy with respect to this endpoint can be claimed. | the endpoint 2 is not rejected | the endpoint 3 is not rejected | $j^{(1)}\leftarrow 1$<br>$k^{(1)}\leftarrow\{1\}$<br>$F^{(1)}\leftarrow\{2,3\}$<br>$\alpha_2^{(0)}$ and $\alpha_3^{(0)}$ are updated to $\alpha_2^{(1)}$ and $\alpha_3^{(1)}$ ;<br><br>Calculate $c_{2j}^{(1)}, c_{3j}^{(1)}$ based on $\alpha_2^{(1)}$ and $\alpha_3^{(1)}$ |

| | | | | |
|---|---|---|---|---|
| IA2 | | Test with $c_{22}^{(1)}$ ($j = 2$ and $j > j^{(1)}$) and cross the boundary value, the endpoint 2 is rejected, efficacy with respect to this endpoint can be claimed. | the endpoint 3 is not rejected | $j^{(2)} \leftarrow 1$<br>$k^{(2)} \leftarrow \{1,2\}$<br>$F^{(2)} \leftarrow \{3\}$<br>$\alpha_3^{(1)}$ is updated to $\alpha_3^{(2)}$ ;<br><br>Calculate $c_{3j}^{(2)}$ based on $\alpha_3^{(2)}$ |
| IA3 | | | the endpoint 3 is not rejected | None of the $(3-1-1=1)$ endpoint crossed its boundary at 3rd look, retain all $H_k\left(k \in F^{(2)} = \{3\}\right)$ and stop testing; |
| Final | | | Test with $c_{33}^{(3)}$ ($j = 4$ and $i = 3$ and $j > j^{(3)} = 2$) and cross the boundary value, the endpoint 3 is rejected, efficacy with respect to this endpoint can be claimed. | All the endpoints are rejected and stop testing. |

### 3.3.4.3.3 Repeated Hypothesis Testing Using Sequentially Rejective Graphical Procedures

The graphical approach by Bretz et al. (Bretz, Frank, Maurer, Willi, Brannath, Werner, & Posch, Martin, 2009) allows one to implicitly define a weighting strategy on all inter-section hypotheses $H_J = \bigcap_{i \in J} H_i$. Note that all the mathematical symbols are following those define in Section 3.3.2.

The elementary hypotheses are represented by nodes with associated weights $w_i$ representing the local significance levels $\alpha w_i$. The transition weight $g_{ij}$ associated with a directed edge between any two vertices $H_i$ and $H_j$ indicates the fraction of the (local) significance level at the initial node (head) that is added to the significance level at the terminal node (tail) if the hypothesis $H_i$ at the head is rejected.

The sum of the transition weights with tail on node $H_i$ is restricted by $1$ for $i \in I = \{1, \ldots, h\}$ and there are no elementary loops (edges where head and tail coincide); that is, $g_{ii} = 0$ for $i \in I$.

We can extend the algorithm from Maurer et al. (Willi Maurer & Ekkehard Glimm &Frank Bretz, 2011) to group sequential designs. More specifically, the following algorithm determines sequentially rejective graphical testing procedures based on consonant closed weighted Bonferroni tests using group sequential boundaries.

0. Set $t = 1$, $I = \{1, 2, \ldots, h\}$.
1. At interim analysis $t$ compute unadjusted $p$-values $p_{i,t}$ and nominal significance levels $\alpha^*_{i,t}(\alpha w_i(I))$ for $i \in I$.
2. Select a $j \in I$ such that $p_{j,t} \le \alpha^*_{j,t}(\alpha w_j(I))$ and reject $H_j$; go to Step 3.
   If no such $j$ exists and $t < k$, the trial can be continued with $t \to t + 1$; go to Step 1 in this case, otherwise stop.
3. Update the graph:

$$I \to I \setminus \{j\}$$

$$w_\ell(I) \to \begin{cases} w_\ell(I) + w_j(I) g_{j\ell}, & \ell \in I, \\ 0, & \text{otherwise}, \end{cases}$$

$$g_{\ell k} \to \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ell \ne k, \\ & g_{\ell j} g_{j\ell} < 1 \\ 0, & \text{otherwise}. \end{cases}$$

4. If $|I| \ge 1$, go to Step 1; otherwise stop.

Figure 43 – Algorithm determines sequentially rejective graphical testing procedures based on consonant closed weighted Bonferroni tests using group sequential boundaries.

During the revision of the article (Willi Maurer & Frank Bretz, 2013) introducing the algorithm in Figure 42, the group sequential Holm procedure for multiple primary endpoints defined in this section was

published. The graphical approach algorithm also **includes their group sequential version of the weighted (and unweighted) Holm procedure as a special case**. For example, with $h$ hypotheses and weights $w_i,\ 0 < w_i < 1, i = 1,\ldots,h,\ \sum_{i=1}^{h} w_i = 1$, it is equivalent to the group sequential graphical procedure by setting $I = \{1,2,\ldots,\},\ w_i(I) = w_i$, and $g_{ij} = w_j/(1 - w_i),\ i,j = 1,\ldots,h, i \neq j$.

**Example: Diabetes trial with two preplanned interim analyses comparing two doses (low and high) against placebo**

- Two pre-planned interim analyses;
- Comparing two doses (low and high) against placebo;
- Two hierarchically ordered endpoints (HbA1c level and body weight).

We have in total $m = 4$ hypotheses, grouped into

- Two primary hypotheses $H_1,\ H_2$ comparing low and high doses, respectively, against placebo for HbA1c;
- Two secondary hypotheses $H_3,\ H_4$ (same dose-control comparisons for body weight).

Pairs of parent-descendant hypotheses to reflect the hierarchy among the two end-points within a same dose:

- $\{H_1, H_3\}$ and $\{H_2, H_4\}$

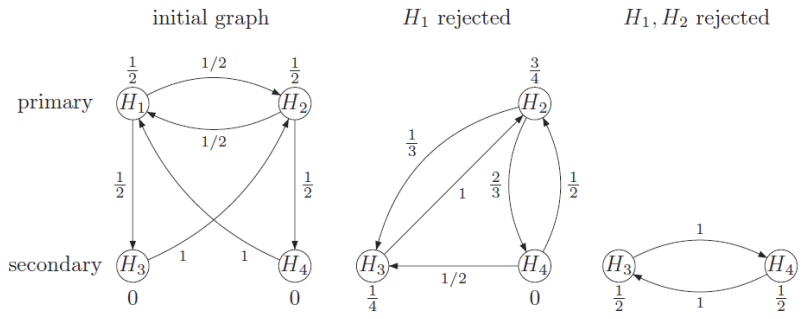



Figure 44 - Graph for a successive sequentially rejective testing procedure with a rejection sequence example.

The left panel in Figure 43 displays the graphical testing strategy employed in this case study. The four elementary hypotheses $H_1,\ldots,H_4$ are represented by nodes with associated weights $w_1, w_2, w_3, w_4$ representing the local significance levels $\alpha_i = \alpha w_i\ ,\ i = 1,\ldots,4$.

According to the successiveness principle, we do not want to reject a descendant secondary hypothesis until its parent primary hypothesis is rejected. Thus, we set the initial local significance levels $\alpha_3 = \alpha_4 = 0$.

Assume that both doses are considered to be equally important and thus let $\alpha_1 = \alpha_2 = \alpha/2$. Further-

more, assume that if one of two primary hypotheses can be rejected at local significance level $\alpha/2$, this level is halved and propagated to the descendant secondary hypothesis (within the same dose) as well as the other primary hypothesis (for the other dose). If both primary hypotheses can be rejected in sequence, both secondary hypotheses are tested at updated levels $\alpha_3 = \alpha_4 = \alpha/2$, with the possibility to further propagate the levels between each other. Alternatively, if both parent-descendant hypotheses within a same dose can be rejected, the remaining hypotheses for the other doses are tested hierarchically at level $\alpha$.

### 3.3.4.3.4 Example: HER2CLIMB

**Protocol Title**

Phase 2 Randomized, Double-Blinded, Controlled Study of Tucatinib vs. Placebo in Combination with Capecitabine and Trastuzumab in Patients with Pretreated Unresectable Locally Advanced or Metastatic HER2+ Breast Carcinoma (HER2CLIMB)

**Study Objectives**

***Primary Objective***

- To assess the effect of tucatinib vs. placebo in combination with capecitabine and trastuzumab on progression-free survival (PFS) per Response Evaluation Criteria In Solid Tumors (RECIST) 1.1 based on blinded independent central review (BICR)

***Secondary Objectives***

- To assess the effect of tucatinib vs. placebo in combination with capecitabine and trastuzumab in patients with brain metastases at baseline, defined as patients with a history of brain metastases, current brain metastases, or equivocal brain lesions at baseline, using RECIST 1.1 based on BICR
- To assess the effects of tucatinib vs. placebo in combination with capecitabine and trastuzumab on overall survival (OS)
- To assess the effect of tucatinib vs. placebo in combination with capecitabine and trastuzumab on PFS per RECIST 1.1 based on investigator assessment
- To assess the effects of tucatinib vs. placebo in combination with capecitabine and trastuzumab on objective response rate (ORR) per RECIST 1.1 based on BICR and by the investigator
- To assess the duration of response (DOR) of tucatinib in combination with capecitabine and trastuzumab per RECIST 1.1 based on BICR and by the investigator
- To assess the clinical benefit rate (CBR) [stable disease (SD) or non-complete response (CR)/non-progressive disease (PD) for ≥ 6 months, or best response of CR or partial response (PR)] of tucatinib vs. placebo in combination with capecitabine and trastuzumab per RECIST 1.1 based on BICR
- To assess the health-related quality of life and health economics associated with tucatinib vs. placebo in combination with capecitabine and trastuzumab based on patient health status collected using the EQ-5D-5L instrument and health resources utilized in patient care

**Endpoints**

***Primary Endpoint***

PFS, defined as the time from randomization to documented disease progression (as determined by BICR per RECIST 1.1), or death from any cause, whichever occurs first

***Secondary Endpoints***

*Key Secondary Endpoints*

- PFS in patients with brain metastases at baseline using RECIST 1.1 as determined by BICR
- OS

*Other Secondary Endpoints*

- PFS, defined as the time from randomization to investigator-assessed documented disease
- progression (per RECIST 1.1), or death from any cause, whichever occurs first
- ORR (RECIST 1.1) as determined by BICR as well as the investigator
- DOR (RECIST 1.1) as determined by BICR as well as the investigator
- CBR (RECIST 1.1) as determined by BICR as well as the investigator

**Study Design**

This is a Phase 2, randomized, international, multi-center, double-blinded study of tucatinib or placebo in combination with capecitabine and trastuzumab in patients with pretreated unresectable locally advanced or metastatic HER2+ breast cancer who have had prior treatment with trastuzumab, pertuzumab, and T-DM1.

***Randomization***

After signing informed consent and meeting all eligibility criteria, patients will be randomized in a 2:1 ratio using a dynamic hierarchical randomization scheme to receive tucatinib or placebo in combination with capecitabine and trastuzumab.

Stratification factors will include presence or history of treated or untreated brain metastases or brain lesions of equivocal significance (yes/no), ECOG PS (0 vs. 1), and region of world (US vs. Canada vs. Rest of World). Stratification for presence of brain metastases will be based upon medical history and investigator assessment of screening contrast brain MRI.

***Dose administration***

Treatment will be administered in cycles of 21 days each.

- Tucatinib 300 mg or placebo will be given orally twice daily (PO[J] BID[K]).
- Capecitabine will be given at 1000 mg/m$^2$ PO BID on Days 1–14 of each 21-day cycle.
- Trastuzumab will be given as a loading dose of 8 mg/kg intravenously (IV) followed by 6 mg/kg once every 21 days, except in specific circumstances where it may be given weekly to compensate for modifications in treatment schedule. In instances of subcutaneous trastuzumab use, a fixed dose of 600 mg is administered without a loading dose. Following an IV loading dose of trastuzumab, 6 mg/kg of trastuzumab is administered once every 21 days, except in specific circumstances where it may be given

**Commented [YL61]:** 基线脑转移

**Commented [YL62]:** 不可切除的局部晚期或转移性HER2+乳腺癌

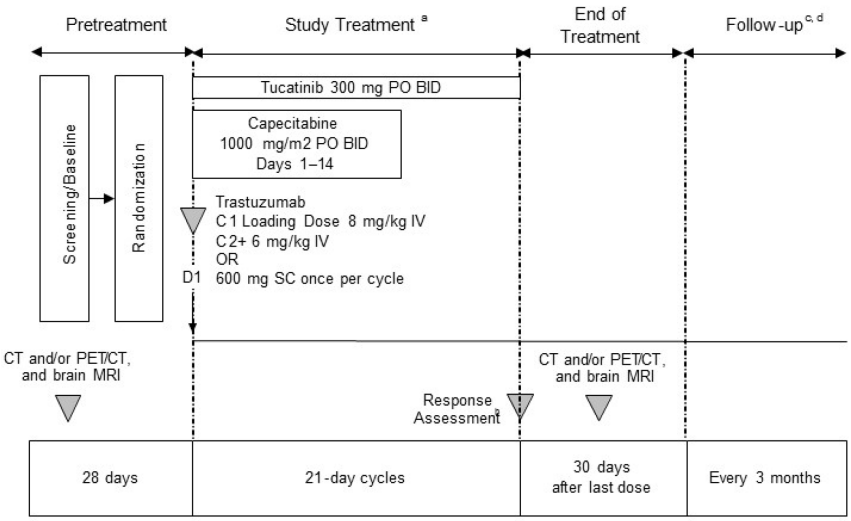
[J] Oral administration
[K] Twice daily

weekly to compensate for modifications in treatment schedule. Subcutaneous trastuzumab is given only once every three weeks as there is no allowance for weekly dosing. There is no ability to modify the trastuzumab dose when administered subcutaneously. Dose modifications of tucatinib or placebo and capecitabine will be allowed.

Commented [YL63]: 皮下

Treatment will continue until unacceptable toxicity, disease progression, withdrawal of consent, or study closure. In the absence of clear evidence of disease progression (per RECIST 1.1), development of CNS symptoms, or radiographic changes thought to pose potential immediate risk to the patient, all efforts should be made to continue treatment until unequivocal evidence of radiologic progression occurs. No crossover from placebo to tucatinib will be allowed. However, patients assessed as having isolated progression in the brain, may be eligible to continue on study treatment for clinical benefit after undergoing local therapy to CNS disease, with approval from the medical monitor.

Safety monitoring will be performed by the sponsor throughout the study on a blinded basis. An independent Data Monitoring Committee (DMC) will regularly review all relevant safety data (blinded and unblinded) as outlined in a separate DMC charter. Ad hoc meetings of the DMC may be held upon the request of the sponsor or DMC.




**a.** Treatment will continue until unacceptable toxicity, disease progression, withdrawal of consent, or study closure. Patients with CNS progression may undergo local therapy to CNS lesions and continue on study treatment with approval from the medical monitor for clinical benefit.

**b.** Contrast CT, PET/CT (CT must be of diagnostic quality), and/or MRI, and brain contrast MRI scan at baseline, every 6 weeks for the first 24 weeks, and then every 9 weeks thereafter until PD, initiation of a new therapy, withdrawal of consent, or study closure. Patients without brain metastases at baseline do not require brain contrast MRIs while on treatment. A brain contrast MRI is required at the 30-Day Follow-up Visit for all patients.

**c.** Assessment of overall survival and/or disease recurrence, as well as collection of information regarding any additional anti-cancer therapies administered after completion of study treatment.

**d**. If study treatment is discontinued for reasons other than disease progression (per RECIST 1.1) or death, every reasonable effort will be made to obtain contrast CT, PET/CT and/or MRI, and contrast brain MRI (only in patients with known brain metastases) approximately every 9 weeks until disease progression (per RECIST 1.1), death, withdrawal of consent, or study closure.

Statistical Methods

**Interim Analyses**

One formal interim analysis for superiority is planned for PFS for the subgroup of patients with brain metastases at baseline ($PFS_{BM}$) and two formal interim analyses for superiority are planned for OS if the primary analysis for PFS is statistically significant. The interim analyses and final analysis will be conducted at the timing described in the Figure 42 and Table 14 below:

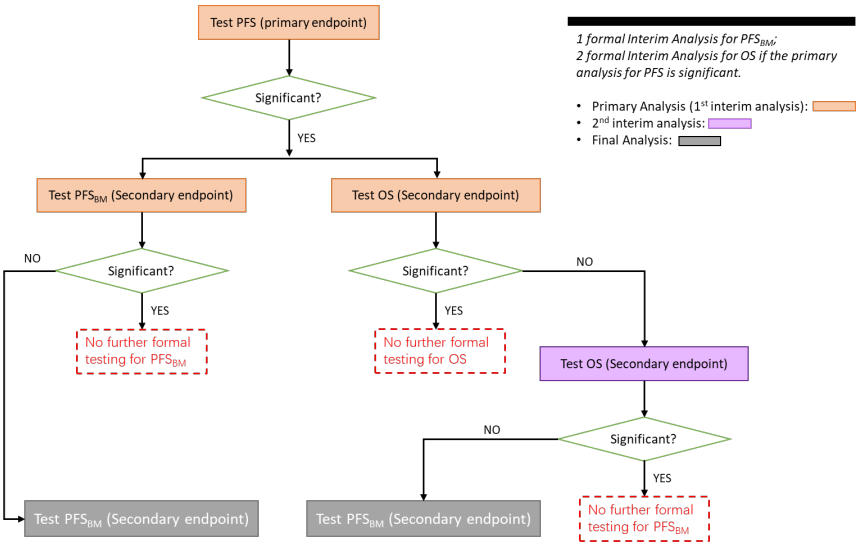



Figure 45 - A flowchart of multiple testing strategy.

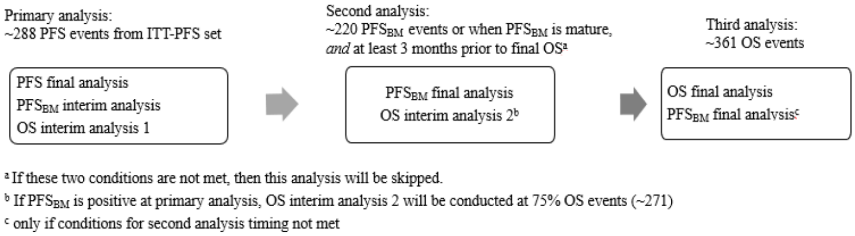
Table 14 - A summary of timing of Analyses.

| INTERIM ANALYSIS | DESCRIPTION |
|---|---|
| **1ST ANALYSIS (PRIMARY ANALTSIS)** | The primary analysis of PFS will occur when: at least 288 PFS events determined by BICR have occurred in the first 480 randomized patients in the ITT population; and enrollment has been completed for the study.<br><br>An interim analysis for the key secondary endpoints $PFS_{BM}$ and OS in the ITT population will also be performed at this time if PFS is statistically significant. |
| **2ND ANALYSIS** | • If $PFS_{BM}$ is NOT statistically significant at the time of the primary analysis of PFS, the final analysis of $PFS_{BM}$ will be performed when approximately 220 PFS |

**Commented [YL64]:** 与 2nd IA 有关

| | |
|---|---|
| | events based on BICR have occurred in the subgroup of subjects with a history of brain metastases and/or brain metastases or brain lesions of equivocal significance at baseline. If OS is NOT statistically significant at the time of the primary analysis of PFS, a second interim analysis for OS will be performed at this time. Update of PFS will also be provided at the time of final analysis for $PFS_{BM}$.<br>***or***<br>• If $PFS_{BM}$ is statistically significant at the time of the primary analysis of PFS, no further formal testing of $PFS_{BM}$ will be conducted. A second interim analysis for OS will be performed when approximately 75% (271) of the total required 361 OS events have occurred in the ITT population, if OS is not statistically significant at the time of the primary analysis of PFS. |
| **3RD ANALYSIS (FINAL ANALYSIS)** | If OS is not statistically significant at the first or second analysis, the final analysis of OS will be performed after 361 OS events have occurred in the ITT population. |

The timing of analyses for the primary and key secondary endpoints are illustrated in Figure 43.




[a] If these two conditions are not met, then this analysis will be skipped.
[b] If $PFS_{BM}$ is positive at primary analysis, OS interim analysis 2 will be conducted at 75% OS events (~271)
[c] only if conditions for second analysis timing not met

Figure 46 - Timing of Primary and Key Secondary Endpoints Analyses.

**Stopping Boundaries**

The stopping boundaries will be determined using Lan-DeMets spending functions for the O'Brien and Fleming boundaries. See details in Section **Multiplicity**.

**Multiplicity**

The sequence of testing will begin with the evaluation of PFS. To maintain strong control of the family-wise type I error rate at 0.05, the PFS will be tested using the first 480 randomized patients in the ITT-PFS set at 0.05 level first.

As illustrated in Figure 44, If it is significant, then the key secondary endpoints will be tested using the group sequential Holm variable (GSHv) procedure (Yining Ye; Ai Li; Lingyun Liu; Bin Yao;, 2013).

**Commented [YL65]:** 在第一次期中分析的时候，如果 Primary endpoint 统计学显著，才会对 key secondary endpoint 使用 GSHv.

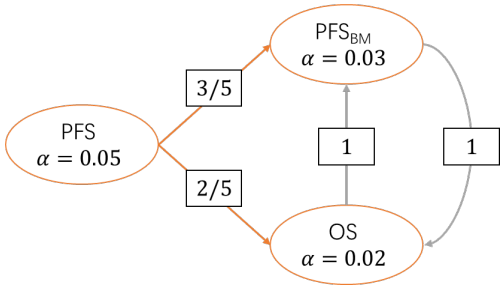



Figure 47 - Type I Error Reallocation Strategy Following Closed Testing Principle.

If PFS is statistically significant, then the key secondary endpoints $PFS_{BM}$ and OS will be tested using group sequential boundaries:

- The $\alpha = 0.05$ split between $PFS_{BM}$ and OS is $\alpha = 0.03$ and $\alpha = 0.02$, respectively. If only one of the two key secondary endpoints is statistical significant, the unused alpha can be passed to the other one following the GSHv procedure. Each one will be tested at the interim analysis(s) and again at the final analysis, if not rejected at the interim analysis.
- The information fraction $t$ is the ratio between number of events at interim analysis and number of events at final analysis. For illustration purpose, we assume $t = 0.812$ for $PFS_{BM}$ and $t_1 = 0.626$, $t_2 = 0.779$ for the two interim analyses for OS as illustrated in Table 15.

Table 15 - A summary of the information fraction at interim analysis and final analysis.

| | Information fraction $t$ | |
|---|---|---|
| | **$PFS_{BM}$** | **OS** |
| 1 (primary analysis) | 0.812 | 0.626 |
| 2 | Not Applicable | 0.779 |
| 3 (final analysis) | 1 | 1 |

A Lan-DeMets O'Brien-Fleming approximation spending function (see Appendix 5.1) will be used for the calculation of efficacy boundaries for $PFS_{BM}$ and OS. The boundary at interim analysis is determined according to the Lan-DeMets O'Brien-Fleming (LD(OF)) approximation spending function

$$\alpha_{LD(OF)}(t) = 4\left(1 - \Phi\left(-\frac{Z_{\frac{\alpha}{4}}}{\sqrt{t}}\right)\right) = 4\left(1 - \Phi\left(\frac{Z_{1-\frac{\alpha}{4}}}{\sqrt{t}}\right)\right)$$

for two-sided tests, where $Z_{\frac{\alpha}{4}}$ is the upper $\frac{\alpha}{4}$ critical point (quantile) of the standard normal distribution.

The GSHv procedure operates as follows:

Scenario 1. Begin with a 0.03-level group sequential boundary for $PFS_{BM}$ and a 0.02-level group sequential boundary for OS. The corresponding boundaries for the two endpoints are given in Table 16. If both of the endpoints are found significant at analysis 1 (primary analysis), then no more

**Commented [TT66]:** 思考一下为什么这么分配?

**Commented [YL67]:** 这个表达式有点奇怪? 为什么是alpha/4?

**Commented [TT68]:** For 2-sided

**Commented [YL69]:** Code for scenario1 & 2:

```
alphaOF <- function(t,alphainit){
  a <- 4*(1-pnorm((qnorm(1-alphainit/4))/sqrt(t)))
  return(a)
}
#scen1
pfsbm_alpha1 <- alphaOF(0.812,0.03)
pfsbm_alpha2 <- alphaOF(1,0.03)
os_alpha1 <- alphaOF(0.626,0.02)
os_alpha2 <- alphaOF(0.779,0.02)
os_alpha3 <- alphaOF(1,0.02)

#scen2: 0.05 level
pfsbm_full_alpha1 <- alphaOF(0.812,0.05)
pfsbm_full_alpha2 <- alphaOF(1,0.05)
os_full_alpha1 <- alphaOF(0.626,0.05)
os_full_alpha2 <- alphaOF(0.779,0.05)
os_full_alpha3 <- alphaOF(1,0.05)

使用 gsDesign 包:
library(gsDesign)
aaa <- sfLDOF(alpha=0.03/2, t=0.812)
print(round(aaa$spend*2,digits = 4))
```

formal statistical testing for $PFS_{BM}$ and OS will be conducted.

Table 16 - **Initial** LD(OF) nominal P-value boundaries for $PFS_{BM}$ (2 analyses) and OS (3 analyses).

| Analysis | $PFS_{BM}$ ($\alpha = 0.03, t = 0.812$) | OS ($\alpha = 0.02, t_1 = 0.626, t_2 = 0.779$) |
|---|---|---|
| **1** | 0.0139 | 0.0023 |
| **2** | 0.0259 | 0.0069 |
| **3** | | 0.0176 |

Scenario 2. If only one endpoint is found significant at the analysis 1 (primary analysis) then the $\alpha$ can be recycled to the other endpoint:

- If $PFS_{BM}$ is significant at interim but OS is not then the $\alpha$ will be recycled from $PFS_{BM}$ to OS and use **a 0.05-level LD(OF) boundary** for OS.
- If OS is significant at analysis 1 (primary analysis) but $PFS_{BM}$ is not then the $\alpha$ will be recycled from OS to $PFS_{BM}$ and use **a 0.05-level LD(OF) boundary** for $PFS_{BM}$.

**The corresponding 0.05-level LD(OF) boundaries** are given in Table 17. The unrejected hypothesis can be re-tested at the current and future analysis using the modified boundaries.

Table 17 - LD(OF) boundaries for $PFS_{BM}$ (2 lo analyses) and OS (3 analyses) at $\alpha = 0.05$ level.

| Analysis | $PFS_{BM}$ ($\alpha = 0.05, t = 0.812$) | OS ($\alpha = 0.05, t_1 = 0.626, t_2 = 0.779$) |
|---|---|---|
| **1** | 0.0258 | 0.0092 |
| **2** | 0.0425 | 0.0194 |
| **3** | | 0.0429 |

Scenario 3. If neither of the endpoints is found significant at analysis 1 (primary analysis), then both endpoints will be tested again at analysis 2. The initial boundaries for final analysis follow Table 16 (analysis 2). If only one endpoint is found significant by these initial boundaries, then the other one can be tested again using the modified boundary as shown in Table 17 (analysis 2). For example, if $PFS_{BM}$ was found significant at final analysis at $\alpha = 0.0259$ level, but OS was not significant at $\alpha = 0.0069$ level, then OS can be **tested again** at the $\alpha = 0.0194$ level.

Scenario 4. If $PFS_{BM}$ is significant at analysis 1 or 2, the boundary of OS analysis at analysis 3 is 0.0429; otherwise, the boundary for OS analysis at analysis 3 is 0.0176.

*Note that the boundaries presented in the tables will be adjusted with the actual information fraction.*

***When the second interim analysis is not conducted***

The second interim analysis for OS may not be conducted, which means both $PFS_{BM}$ and OS will have at most 2 analyses. In that case, LD(OF) boundaries at each analysis will be modified as illustrated in Table 18 and Table 19. Similar to Table 16 and Table 17, the information fraction $(t)$ in Table 18 and Table 19 are for illustration purpose only.

If both $PFS_{BM}$ and OS are statistically significant, the secondary endpoint of ORR by BICR in the ITT-OS set will be formally tested between two treatment arms at the two-sided $\alpha = 0.05$ level.

**Commented [YL70]:** 约为 t=0.967，不是 t=1，为什么在最后一次分析时候信息量不是 t=1？是因为 data cut 的时候 os 事件数没达到预设的~361？

**Commented [YL71]:** 假如在 1st IA 的时候 PFSBM 显著，OS 不显著，则依据 sequential Holm 算法，会（$j \geq j^{(1)}$）去使用更新后的$\alpha$来检验另一个不显著的终点，比如此处为 OS。

**Commented [TT72]:** In this case(recycling occurs at stage s > 1), GSHv wastes the portion of the recycled significance level allocated to stages 1, . . . , s-1 since those stages can't be revisited.

GSHf does not waste any recycled significance level, but the trial has to continue to the final stage to benefit from recycling.

**Commented [YL73]:** 这一段非常重要：
1.如果 PFSBM 和 OS 在 1st IA 都不显著，则 stop testing，等 2nd IA; 1st IA 用来与检验统计量比较的 boundary 如表 16 analysis1 那行所示。
2.另一种情况： 如果在 1st IA 的时候，其中一个终点显著，那么另一个可以在 1st IA 同时期，用 table 17 analysis 1 时候的 boundary 去检验，这里这个 boundary 的 alpha 是 0.05（"完整"的）；

**Commented [YL74]:** 如果 PFSBM 在前两次 IA 都会显著，那么 final analysis 也就是第三次分析时候的 OS 是可以基于 0.05-level（full alpha） boundary 来检验的

**Commented [YL75]:** Lan-DeMets error spending function 的特性在于，不用事先定义我们需要执行几次 IA，这样会带来很大的便利性，我们可以在分析的过程中去调整 IA 的次数。比如此处，我不打算执行第二次 IA 了。

**Commented [YL76]:** 此处指的是 IA 的信息比例

Table 18 - Initial LD (OF) boundaries for $PFS_{BM}$ (2 analyses) and OS (2 analyses).

| Analysis | $PFS_{BM}$ $(\alpha = 0.03, t = 0.812)$ | OS $(\alpha = 0.02, t = 0.626)$ |
|---|---|---|
| **1** | 0.0139 | 0.0023 |
| **2** | 0.0259 | 0.0193 |

Table 19 - LD (OF) boundaries for $PFS_{BM}$ (2 analyses) and OS (2 analyses) at $\alpha = 0.05$ level.

| Analysis | $PFS_{BM}$ $(\alpha = 0.05, t = 0.812)$ | OS $(\alpha = 0.05, t = 0.626)$ |
|---|---|---|
| **1** | 0.0258 | 0.0092 |
| **2** | 0.0425 | 0.0471 |

### 3.3.4.4 Weighted parametric group sequential design (WPGSD)

**Group sequential design (GSD)** is widely used in clinical trials in which correlated tests of multiple hypotheses are used. Multiple primary objectives resulting in tests with known correlations include evaluating

1. multiple experimental treatment arms,
2. multiple populations,
3. the combination of multiple arms and multiple populations, or
4. any asymptotically multivariate normal tests.

In the paper by Keaven M. Anderson et al. (Keaven M. Anderson, Zifang Guo, Jing Zhao, & Linda Z. Sun, 2021), they focused on the first 3 of these and extend the framework of the weighted parametric multiple test procedure from fixed designs with a single analysis per objective to a GSD setting where different objectives may be assessed at the same or different times, each in a group sequential fashion.

Pragmatic methods for design and analysis of weighted parametric group sequential design (WPGSD) under closed testing procedures are proposed to maintain the strong control of familywise Type I error rate (FWER) when correlations between tests are incorporated.

This results in the ability to relax testing bounds compared to designs not fully adjusting for known correlations, increasing power or allowing decreased sample size.

**Commented [YL77]:** 这种方法与不完全 include 已知相关性的 design 相比，可以放松边界值，增加 power 或者降低样本量；这些是这个方法的优势

The proposed unified framework of weighted parametric group sequential design (WPGSD) focuses on closed testing procedures for GSD with multiple endpoints by Tang and Geller (Dei-In Tang & Nancy L Geller, 1999) and the graphical approach in GSD of Maurer and Bretz (Willi Maurer & Frank Bretz, 2013).

While it is not obvious how to use spending functions to calculate boundaries for the test statistics in Tang and Geller (Dei-In Tang & Nancy L Geller, 1999), detailed algorithms to compute boundaries in WPGSD are described in this section. The proposed framework comprehensively covers many procedures in the previous literature as special cases: for example, multi-arm multi-stage designs, multiple population GSD, or general GSD with multiple correlated endpoints.

#### 3.3.4.4.1 Motivating examples

**Scenario 1**

First consider a 2-arm controlled clinical trial example with one primary endpoint E and 3 patient populations defined by the status of two biomarkers (KEYNOTE-181 trial), evaluating pembrolizumab vs. investigator's choice of chemotherapy as second-line therapies for patients with advanced or metastatic squamous cell carcinoma and adenocarcinoma of the esophagus or Siewert type I adenocarcinoma of the esophagogastric junction.

Assume an **interim analysis (IA)** and a **final analysis (FA)** are planned for the study.

The 3 primary hypotheses of the trial are:

1) $H_1$ to test that the OS in experimental treatment is superior to the control in the squamous cell carcinoma subgroup (Population 1);
2) $H_2$ to test the superiority in the subgroup with PD-L1 CPS ≥10 (Population 2);
3) $H_3$ to test the superiority in the intent-to-treat population (Population 3).

Tests of these null hypotheses were inherently correlated due to the overlapping populations as shown in Figure 48.

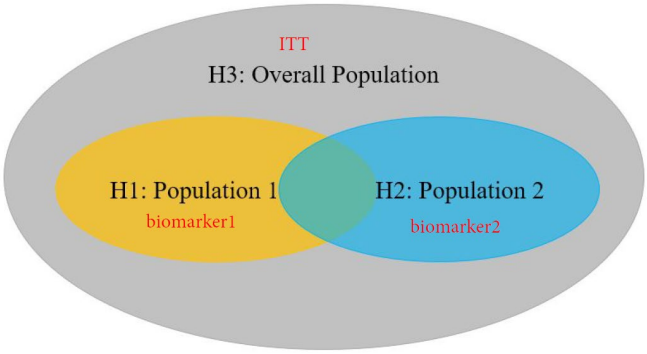



Figure 48 - The 3 populations of Scenario 1.

In current practice, the closed weighted Bonferroni approach is often used to split $\alpha$ among hypotheses, leaving room for improvement using methods that account for correlation among tests.

**Scenario 2**

Consider a common example where correlation among primary hypotheses could be considered is a group sequential design with multiple experimental arms versus a common control.

Assume subjects are randomized 1:1:1:1 among 4 treatment arms:

- experimental arms and
- 1 single control arm.

The 3 experimental arms could be different dose levels of the same drug (e.g., low-dose, mid-dose, and high-dose), or different combinations of multiple drugs (e.g., Drug A, Drug B, and Drug A + Drug B).

Suppose the primary endpoint of the trial is a single time-to-event endpoint, with a hypothesis for each experimental arm vs. control. With one planned interim and one final analysis, the trial has a total of 6 test statistics that are inherently correlated.

Correlations among test statistics from the same hypothesis are the usual temporal correlation in a group sequential design. Correlations among hypotheses arise from the fact that the control arm events are utilized in comparisons for multiple experimental arms.

## 3.3.4.4.2 Methodology

In this section, the notations for graphical approaches are exactly the same as those in Section 3.3.2.

### 3.3.4.4.2.1 Correlated hypothesis

Among the tests of the $m$ individual hypotheses, some of them can have known correlations.

One example is to test the treatment effect of the same endpoint but in nested or overlapping populations. Another example is to test the treatment effect of different treatment arms or doses versus a shared control arm.

For simplicity, we assume the plan is for all hypotheses to be tested at each of the $K$ planned analyses if the trial continues to the end for all hypotheses. We assume further that the distribution of the $m \times K$ tests of $m$ individual hypotheses at all $K$ analyses is multivariate normal with a completely known correlation matrix. Neither of these assumptions is necessary, but they are common and enable a more straightforward presentation of the methods.

Let $Z_{ik}$ be the standardized normal test statistic for hypothesis $i \in I$, analysis $1 \leq k \leq K$.
Let $n_{ik}$ be the the number of observations (or number of events for time-to-event endpoints) collected cumulatively through stage $k$ for hypothesis $i$.
We use the wedge operator in $n_{i\wedge i',k\wedge k'}$ to denote he number of observations (or events) included in both $Z_{ik}$ and $Z_{i'k'}$.

The key of the parametric tests is to utilize the correlation among the test statistics. The correlation between $Z_{ik}$ and $Z_{i'k'}$ is

$$\mathrm{Corr}(Z_{ik}, Z_{i'k'}) = \frac{n_{i\wedge i',k\wedge k'}}{\sqrt{n_{ik}n_{i'k'}}}.$$

The full correlation matrix of $Z_{ik}$ and $Z_{i'k'}$ $(mK \times mK)$ can be derived this way and is referred to as the complete correlation structure (CCS).

### 3.3.4.4.2.2 Well-ordered family of group sequential bounds

Commented [YL78]: 良序，一种数学理论

A restriction of spending functions to those that produce well-ordered bounds as $\alpha$-levels change is required to apply the graphical and closed testing procedures for group sequential design (Willi Maurer & Frank Bretz, 2013).

We fix statistical information for each analysis at $0 \leq \mathfrak{T}_{i1} \leq \cdots \leq \mathfrak{T}_{iK}$. For simplicity of notation, we just assume that bounds $c_{ik}(\gamma)$ will be defined for $H_i$ at all significance levels of $\gamma$ satisfying

$0 < \gamma < 1$,

$$\gamma = 1 - P\left(\bigcup_{k=1}^{K} \{Z_{ik} \geq c_{ik}(\gamma)\}\right),$$

For a well-ordered family Maurer and Bretz (Willi Maurer & Frank Bretz, 2013), we require for any

$$0 < \gamma_1 < \gamma_2 < 1,$$

that

$$c_{ik}(\gamma_1) \geq c_{ik}(\gamma_2).$$

Maurer and Bretz (Willi Maurer & Frank Bretz, 2013) generate such bounds using a well-ordered spending function family.

It should be noted that the O'Brien-Fleming-like and Pocock-like spending functions of Lan and DeMets, the power spending functions, and the Hwang-Shih-DeCani spending function all produce well-ordered boundary families.

Where we use spending function families for boundary setting, we will assume they are well-ordered. A completely-ordered family Liu and Anderson (Qing Liu & David L. DeMets, 1989) adds the requirement that $c_k(\gamma)$ is strictly increasing in $\gamma$ and as $\gamma \swarrow 0$, we have $c_{ik}(\gamma) \nearrow \infty$.

# 4 Bibliography


Alex Dmitrienko, Frank Bretz, Peter H. Westfall, James Troendle, Brian L. Wiens, Ajit C. Tamhane, & Jason C. Hsu. (2010). Multiple Testing Methodology. In A. Dmitrienko, F. Bretz, P. H. Westfall, J. Troendle, B. L. Wiens, A. C. Tamhane, & J. C. Hsu, *Multiple testing problems in pharmaceutical statistics* (p. 65). Chapman & Hall/CRC.

Alex Dmitrienko; Ajit C Tamhane; Lingyun Liu; Brian L Wiens;. (2008). A note on tree gatekeeping procedures in clinical trials. *Statistics in medicine*, 3446–3451.

Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;. (2010). *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton: Chapman and Hall/CRC.

Alexei Dmitrienko; Walter W Offen; Peter H Westfall;. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in medicine*, 2387–2400.

Bertram Pitt, Gordon Williams, Willem Remme, Felipe Martinez, Jose Lopez-Sendon, Faiez Zannad, . . . Richard Bittman & Jay Kleiman. (2001). The EPHESUS Trial: Eplerenone in Patients with Heart Failure Due to Systolic Dysfunction Complicating Acute Myocardial Infarction. *Cardiovascular Drugs and Therapy*, 79-87.

Bretz, Frank, Maurer, Willi, Brannath, Werner, & Posch, Martin. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28. doi:https://doi.org/10.1002/sim.3495

Center for Drug Evaluation and Research, & Center for Biologics Evaluation and Research. (2017). *Multiple Endpoints in Clinical Trials Guidance for Industry.* Rockville, MD: U.S. Department of Health and Human Services, Food and Drug Administration.

Deli Wang, Yihan Li, Xin Wang, Xuan Liu, Bo Fu, Yunzhi Lin, . . . Walter Offen. (2015). Overview of multiple testing methodology and recent development in clinical trials. *Contemporary Clinical Trials*, 8.

Deli Wang, Yihan Li, Xin Wang, Xuan Liu, Bo Fu, Yunzhi Lin, . . . Walter Offen. (2015). Overview of multiple testing methodology and recent development in clinical trials. *Contemporary Clinical Trials*, 13–20.

Dmitrienko, A., Ajit C Tamhane, & Brian L Wiens. (2008). General multistage gatekeeping procedures. *Biometrical Journal*, 667-677.

Eugene Grechanovsky, & Yosef Hochberg. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*, 79-91.

Keaven M. Anderson, Zifang Guo, Jing Zhao, & Linda Z. Sun. (2021). A unified framework for weighted parametric group sequential design (WPGSD). *stat.ME*, 2103.10537,arXiv.

Mohammad F. Huque, Alex Dmitrienko, & Ralph D'Agostino. (2013). Multiplicity Issues in Clinical Trials With Multiple Objectives. *Statistics in Biopharmaceutical Research, 5*, 321-337. doi:10.1080/19466315.2013.807749

Mohammad Huque, Ph.D, & Sirisha Mushti, Ph.D. (2015). *Alpha-recycling for the analyses of primary and secondary endpoints of clinical trials.* FDA/CDER/OTS/ Office of Biostatistics. Rockville, Maryland: BASS Conference.

Willi Maurer & Frank Bretz. (2013). Multiple Testing in Group Sequential Trials Using Graphical Approaches. *Statistics in Biopharmaceutical Research*, 311-320.

Willi Maurer, & Ekkehard Glimm &Frank Bretz. (2011). Multiple and Repeated Testing of Primary, Coprimary, and Secondary Hypotheses. *Statistics in Biopharmaceutical Research*, 336-352.

Yining Ye; Ai Li; Lingyun Liu; Bin Yao;. (2013). A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine*, 1112-1124.

Yoav Benjamini, & Yosef Hochberg. (1997). Multiple Hypotheses Testing with Weights. *Scandinavian Journal of Statistics*, 407-418.

# 5 Appendix

## 5.1 Error Spending Methods

From https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/statug/statug_seqdesign_details37.htm

For each sequential design, the $\alpha$ and $\beta$ errors spent at each stage can be computed from the boundary values. For example, for a $K$-stage design with an upper alternative hypothesis $H_1 : \theta = \theta_1$ and early stopping to reject the null hypothesis $H_0 : \theta = 0$, the boundary values in a standardized $Z$ scale are the upper $\alpha$ critical values $a_k$, $k = 1, 2, \ldots, K$. At each interim stage, the null hypothesis $H_0$ is rejected if the observed standardized $Z$ statistic $z_k \geq a_k$. Otherwise, the process continues to the next stage. At the final stage, the hypothesis is rejected if $z_K \geq a_K$. Otherwise, the null hypothesis is accepted.

The boundary values $a_k$ are derived such that the overall Type I error probability

$$\alpha = \sum_{k=1}^{K} \alpha_k$$

where $\alpha_k$ is the $\alpha$ spending at stage $k$. That is, at stage 1,

$$\alpha_1 = P_{\theta=0}\left( z_1 \geq a_1 \right)$$

At a subsequent stage $k$,

$$\alpha_k = P_{\theta=0}\left( z_j < a_j, j = 1, 2, \ldots, k-1,\ z_k \geq a_k \right)$$

Since each design can be uniquely identified by the $\alpha$ and $\beta$ errors spent at each stage, a design can then be derived by specifying the $\alpha$ and $\beta$ errors to be used at each stage. The error spending method (Lan and DeMets 1983) distributes the error to be used at each stage and then derives the boundary values. Numerous forms of the error spending function are available. Kim and DeMets (1987) examine the functions $f(t) = t$, $f(t) = t^{\frac{3}{2}}$, and $f(t) = t^2$, where $t$ is the information fraction. Jennison and Turnbull (2000, p. 148) generalize these functions to the power functions $f(t; \rho) = t^\rho, \rho > 0$.

The ERRFUNCPOC option uses the cumulative error spending function (Lan and DeMets 1983)

$$E(t) = \begin{cases} 1 & \text{if } t \geq 1 \\ \log(1 + (e-1)t) & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

With a specified error of $\alpha$ or $\beta$, the cumulative error spending at stage $k$ is $\alpha\, E(\Pi_k)$ or $\beta\, E(\Pi_k)$, where $\Pi_k = I_k / I_X$ is the information fraction at stage $k$. The method produces boundaries similar to those produced with Pocock's method.

The ERRFUNCOBF option uses the cumulative error spending function (Lan and DeMets 1983)

$$E(t; a) = \begin{cases} 1 & \text{if } t \geq 1 \\ \frac{1}{a}\, 2\left(1 - \Phi\left(\frac{z_{(1-a/2)}}{\sqrt{t}}\right)\right) & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $a$ is either $\alpha$ for the $\alpha$ spending function or $\beta$ for the $\beta$ spending function. That is, with a specified error of $\alpha$ or $\beta$, the cumulative error spending at stage $k$ is $\alpha\, E(\Pi_k; \alpha)$ or $\beta\, E(\Pi_k; \beta)$. The method produces boundaries similar to those produced with the O'Brien-Fleming method.

The ERRFUNCGAMMA option uses the gamma cumulative error spending function (Hwang, Shih, and DeCani 1990)

$$E(t; \gamma) = \begin{cases} 1 & \text{if } t \geq 1 \\ \frac{1 - e^{-\gamma t}}{1 - e^{-\gamma}} & \text{if } 0 < t < 1, \gamma \neq 0 \\ t & \text{if } 0 < t < 1, \gamma = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma$ is the parameter $\gamma$ specified in the GAMMA= option. That is, with a specified error of $\alpha$ or $\beta$, the cumulative error spending at stage $k$ is $\alpha\, E(\Pi_k; \gamma)$ or $\beta\, E(\Pi_k; \gamma)$.

The ERRFUNCPOW option uses the cumulative error spending function (Jennison and Turnbull 2000, p. 148)

$$E(t;\rho) = \begin{cases} 1 & \text{if } t \geq 1 \\ t^{\rho} & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\rho$ is the power parameter specified in the RHO= option. That is, with a specified error of $\alpha$ or $\beta$, the cumulative error spending at stage $k$ is $\alpha\, E(\Pi_k; \rho)$ or $\beta\, E(\Pi_k; \rho)$.

Error spending methods derive boundary values at each stage sequentially and require much more computation than other types of methods for group sequential trials with a large number of stages, especially for a two-sided asymmetric design with early stopping to accept $H_0$, or to reject or accept $H_0$.

Note that for a two-sided design with the STOP=BOTH or STOP=ACCEPT option, at each interim stage, the SEQDESIGN procedure first produces the lower and upper $\beta$ boundary values based on the one-sided $\beta$ spending. If the lower $\beta$ boundary value is greater than or equal to its corresponding upper $\beta$ boundary value, there is no early stopping to accept the null hypothesis at this stage, and the corresponding $\beta$ spending is distributed proportionally to the remaining stages.

For the error spending functions not available in the SEQDESIGN procedure, you can first compute the corresponding error spending at each stage explicitly, then use the SEQDESIGN procedure with the ERRSPEND= option to specify these errors directly.

For example, if the information levels are equally spaced in a five-stage design, the option ERRFUNCPOW (RHO=2) produces relative cumulative errors of $(1/5)^2$, $(2/5)^2$, $(3/5)^2$, $(4/5)^2$, and 1. This is equivalent to using the option ERRSPEND (1 4 9 16 25).

## 5.2 Adjusted significance levels and p-values

In most simple cases, a multiplicity adjustment can be performed by computing a reduced significance level for each individual hypothesis.

In general, adjusted significance levels are used less frequently than adjusted p-values, mainly because adjusted significance levels depend on the $\alpha$ level. However, there are cases when the use of adjusted significance levels simplifies multiplicity adjustments.

Unlike adjusted significance levels, adjusted p-values capture the degree of multiplicity adjustment without reference to the pre-specified error rate and thus one can choose different $\alpha$ levels for different sets of hypotheses. Another advantage of adjusted p-values is that they incorporate the structure of the underlying decision rule which can be quite complex.

A general definition of an adjusted p-value is given in **Westfall and Young**:

The **adjusted p-value** for a hypothesis is the smallest significance level (the smallest FWER level) at which one would reject the hypothesis using the given multiple testing procedure. This definition can be illustrated by applying it to closed testing procedures.

**Commented [YL79]:** 最小的 significance level

For example, $H_i$, if all intersection hypotheses containing $H_i$ are rejected. If $p_I$, $I \subseteq \{1, \ldots, m\}$, denotes the p-value for testing the intersection hypothesis $H_I$, the adjusted p-value for $H_i$ is the largest p-value associated with the index sets including $i$:

$$\tilde{p}_i = \max_{I: i \in I} p_I$$

The hypothesis $H_i$ is rejected if the adjusted p-value does not exceed the pre-specified $\alpha$ level, i.e., $p_i \leq \alpha$. This general approach will be utilized to derive adjusted p-values for multiple testing procedures commonly used in pharmaceutical applications (all of which can be formulated as closed testing procedures).

## 5.3 Proof of 2-stage gatekeeping procedure controls the FWER at the α level

**Proposition 4.1** *The two-stage gatekeeping procedure controls the FWER at the* α *level.*

**Proof of Proposition 4.1** Define the following events:

$$B_1 = \{\text{One or more true null hypotheses are rejected in } F_1\},$$
$$B_2(x) = \{\text{One or more true null hypotheses are rejected at level } x \text{ in } F_2\}.$$

Note that, due to α-consistency of the second-stage MTP, $B_2(x) \subseteq B_2(y)$ if $x \leq y$. Further, since the MTP controls the FWER within $F_2$, $P(B_2(x)) \leq x$. Also, let $e_1^*(I)$ be an upper bound on the error rate function of the first-stage MTP, $I \subseteq N_1$, $\alpha_2$ be the random level at which the second-stage MTP is carried out within the two-stage procedure and $\bar{E}$ be the complement of the event $E$.

The FWER of the two-stage gatekeeping procedure can be written as

$$P(B_1 \cup B_2(\alpha_2)) = P(B_1) + P(\bar{B}_1 \cap B_2(\alpha_2)).$$

Let $T_1 \subseteq N_1$ denote the set of indices corresponding to the true null hypotheses in $F_1$. By the definition of the error rate function, $P(B_1) \leq \mathrm{e}_1^*(T_1)$.

Next consider $\bar{B}_1 \cap B_2(\alpha_2)$. Since

$$\bar{B}_1 = \{\text{No true null hypotheses are rejected in } F_1\},$$

we have $T_1 \subseteq A_1$ and thus, due to the monotonicity condition (4),

$$\alpha_2 = \alpha - \mathrm{e}_1^*(A_1) \leq \alpha - \mathrm{e}_1^*(T_1)$$

when $\bar{B}_1$ is true. Therefore,

$$\bar{B}_1 \cap B_2(\alpha_2) \subseteq \bar{B}_1 \cap B_2(\alpha - \mathrm{e}_1^*(T_1))$$

and

$$P(\bar{B}_1 \cap B_2(\alpha_2)) \leq P(\bar{B}_1 \cap B_2(\alpha - \mathrm{e}_1^*(T_1))) \leq P(B_2(\alpha - \mathrm{e}_1^*(T_1))) \leq \alpha - \mathrm{e}_1^*(T_1).$$

Therefore $P(B_1 \cup B_2(\alpha_2)) \leq \mathrm{e}_1^*(T_1) + \alpha - \mathrm{e}_1^*(T_1) = \alpha$ and thus the two-stage procedure controls the FWER at the α level. The proof of Proposition 4 is complete.

The asterisk identifies the adjusted $p$-values that are significant at the 0.05 level.

## 5.4 A detailed example of sequentially rejective Bonferroni-based closed testing procedures

Commented [YL81]: 其实也是 Bonferroni-Holm weighting strategy in the graphical approach

For Figure 9, Figure 10 and Figure 11, the details regarding the use of closure principle are

| Hypotheses | H1 | H2 | H3 |
|---|---|---|---|
| **H123** | $w_1$ | $w_2$ | 0 |
| H12 | $w_1$ | $w_2$ | 0 |
| **H13** | $w_1+\delta_2 w_2$ | - | $(1-\delta_2)w_2$ |
| H1 | 1 | - | - |
| **H23** | - | $w_2+\delta_1 w_1$ | $(1-\delta_1)w_1$ |
| H2 | - | 1 | - |
| H3 | - | - | 1 |

The parts in red are assumptions.

A. We assume that the unadjusted p-value for hypothesis $H_1$ satisfies $\frac{p_1}{w_1} \leq \alpha$ and $\frac{p_1}{w_1} < \frac{p_2}{w_2} < 1$.

B. We assume that the unadjusted p-value for hypothesis $H_2$ satisfies $\frac{p_2}{w_2+\delta w_1} \leq \alpha$ and $\frac{p_2}{w_2+\delta w_1} < \frac{p_3}{(1-\delta)w_1} < 1$.

1. Let $J = \{1,2,3\}$ and the vector of weights for $\{H_1, H_2, H_3\}$ is $\vec{W}_{step-1} = (w_{step-1,1}(J), w_{step-1,2}(J), w_{step-1,3}(J))$.
   Since $H_3$ is tested only when at least one primary hypothesis is rejected, the weights for the three hypotheses are set to

$$\begin{cases} w_{step-1,1}(J) = w_1 \\ w_{step-1,2}(J) = w_2 \\ w_{step-1,3}(J) = 0 \end{cases};$$

   Where $w_1 + w_2 + w_3 = w_1 + w_2 = 1$.

   We test $H_J$ by comparing $p_J = \min\{\frac{p_1}{w_1}, \frac{p_2}{w_2}, 1\}$. Based on the **assumption A**, $H_J$ can be rejected and the corresponding elementary $H_1$ with the minimum adjusted p-value $\frac{p_1}{w_1} < \alpha$

can also be rejected at level $\alpha$. (See Section 2.3.2.1)

2. Since $H_1$ is rejected in **Step 1**, the index set $J$ is updated to $J' = \{1,2,3\}\backslash\{1\} = \{2,3\}$. Based on the Section 2.3.2.2, we have

$$J' \subseteq J$$

And the procedure should meet the condition

$$w_j(J') \geq w_j(J) \quad \text{for all } J' \subseteq J \subseteq I \text{ and } j \in J'.$$

The vector of weights for $\{H_2, H_3\}$ is $\vec{W}_{step-2} = (w_{step-2,2}(J), w_{step-2,3}(J))$ should satisfy

$$w_{step-2,2}(J') \geq w_{step-2,2}(J)$$
$$w_{step-2,3}(J') \geq w_{step-2,3}(J)$$

Therefore we can set the updated

$$\vec{W}_{step-2} = (w_{step-2,2}(J), w_{step-2,3}(J)) = (w_2 + \delta_1 w_1, 0 + (1 - \delta_1 w_1))$$

It is obvious that $w_2 + \delta_1 w_1 + 0 + (1 - \delta_1)w_1 = 1$ and

$$w_{step-2,2}(J') = w_2 + \delta_1 w_1 > w_{step-1,2}(J) = w_2$$
$$w_{step-2,3}(J') = (1 - \delta_1)w_1 > w_{step-1,3}(J) = 0$$

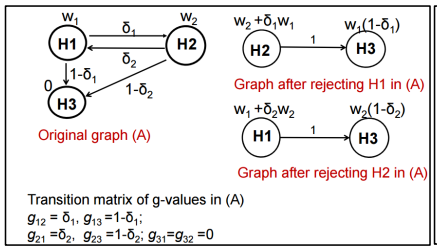
We test $H_J$, by comparing $p_J = \min\{\frac{p_2}{w_2+\delta_1 w_1}, \frac{p_3}{(1-\delta_1)w_1}\}$.

Based on the **assumption B**, $H_J$, can be rejected and the corresponding elementary $H_2$ with the minimum adjusted p-value $p2w_2 + \delta_1 w_1 < \alpha$ can also be rejected at level $\alpha$. (See Section 2.3.2.1)

**Commented [YL82]:** @Tan Tao

3. Since both $H_1$ and $H_2$ are rejected, then one has to simply test $H_3$ at level $\alpha$.

The above complicated steps can be simplified to the following graph:




Transition matrix g-values (original graph):
$g_{12} = \delta_1$, $g_{13} = 1-\delta_1$;
$g_{21} = \delta_2$, $g_{23} = 1-\delta_2$; $g_{31} = g_{32} = 0$

Method for calculating new g-values after rejection of the hypothesis $H_j$:

$$g_{ik}(\text{new}) = (g_{ik}(\text{old}) + g_{ij} * g_{jk})/(1 - g_{ij} * g_{ji})$$

Example after rejecting $H_1$:

$g_{23}$ = {$g_{23}$(old) = $(1-\delta_2)$ + $g_{23}$ (going through $H_1$) = $\delta_2(1-\delta_1)$}/{$(1- g_{21} x\ g_{21}) = (1-\delta_1\delta_2)$} =1

## 5.5 Closure testing principle in group sequential Holm procedure

The group sequential procedure strongly controls the type I error rate at level ¸ because it is a closed test.

考虑整个试验有两个需要检验的终点 A, B，对应假设为$H_A, H_B$；有$(J-1)$次期中分析（IA）及一次最终分析（FA），总共$J$次分析。

根据闭合原理（CTP），想要在$\alpha$水平上拒绝假设$H_A$，需要在显著水平$\alpha$上拒绝$H_A \cap H_B$以及$H_A$。

对于$H_A \cap H_B$，在成组序贯设计中，如果在所有$J$次分析中，终点 A 对应的检验统计量$X_A$在任意一次分析 $j$（$j = 1,2,\dots,J$）中超过了其对应的拒绝域边界值$c_j$（这个边界值是基于显著水平$\alpha_A(\leq \alpha)$计算的），则可以拒绝$H_A \cap H_B$。同样的，以上叙述对终点 B 适用。

基于$\alpha$水平对$H_A \cap H_B$做假设检验的数学表达式为

$$P(reject\ H_A \cap H_B) = P\left(\left(\bigcup_{j=1}^{J}\{X_j > c_j\}\right) \cup \left(\bigcup_{j=1}^{J}\{Y_j > d_j\}\right) \middle| null\ hypotheses\right)$$

$$\leq P\left(\bigcup_{j=1}^{J}\{X_j > c_j\} \middle| null\ hypotheses\right) + P\left(\bigcup_{j=1}^{J}\{Y_j > d_j\} \middle| null\ hypotheses\right)$$

$$= \alpha_A + \alpha_B = \alpha$$

换句话说我们可以在$\alpha$水平上检验$H_A \cap H_B$，只要出现至少一个$X_j > c_j$或者$Y_j > d_j$事件（$j = 1,2,\dots,J$），即可拒绝$H_A \cap H_B$，此时基于 CTP，$\alpha$可以进行$\alpha$-reallocation。

由于$X_j > c_j$或者$Y_j > d_j$事件对应的是显著水平$\alpha_A$或$\alpha_B$，所以实际操作中，我们是在显著水平$\alpha_A$或$\alpha_B$上拒绝$H_A \cap H_B$的。更具体地说，假设在所有$J$次分析中，在第$m$次分析时，我们观测到$X_m > c_m$,则我们可以在显著水平$\alpha_A$上拒绝$H_A \cap H_B$;又因为$\alpha_A \leq \alpha$，显然我们还可以在显著水平$\alpha$上拒绝$H_A$。基于 CTP, 仅需要在显著水平$\alpha$上检验$H_B$。整个思路基于 Sequentially rejective Bonferroni-based closed testing procedures（参考 2.3.2.2 节），所以整个过程可以用 graphical approach，基于 alpha-reallocation 的方式展示。

需要注意的是，以上标红的显著水平在序贯设计中是需要通过 error spending function 来转换的，以控制整个动态过程的一类错误率。简而言之，在上面叙述中，我们并不是直接在$\alpha_A$水平上去检验$H_A \cap H_B$，而是在$f(\alpha_A, t)$（spending function，$t$为 timing of analysis）上去检验。同理，在某次 IA（比如$t = t_0$）中拒绝了$H_A \cap H_B$与$H_A$后，我们要在显著水平$f(\alpha, t_0)$上去检验终点 B。

由于 error spending function 在此处为显著水平的单调非减函数，我们有$f(\alpha, t) \geq f(\alpha_A, t)$当$\alpha \geq \alpha_A$，来保证 CTP 中 consonance 的性质，error spending function 的数理性质也应保证$f(\alpha, t)$对应的决策边界$c_t^*$与$f(\alpha_A, t)$对应的决策边界$c_t$满足$c_t^* \leq c_t$。

## 5.6 Truncated multiple test procedures

Note that **MTP** refers to **m**ultiple **t**est **p**rocedures.

Consider a single family of hypotheses, $\mathcal{F} = \{H_1, \ldots, H_n\}$, with p-values, $p_1, \ldots, p_n$. For convenience, we will assume that the hypotheses are equally weighted. The same principle can be applied to construct truncated MTPs for weighted hypotheses. All MTPs are assumed to be of nominal $\alpha$ level.

### 5.6.1 Truncated Holm

In the Holm (1979) MTP the $p$-values are first ordered, $p_{(1)} \leq \cdots \leq p_{(n)}$. Let $H_{(1)}, \ldots, H_{(n)}$ be the corresponding hypotheses. At the first stage, $H_{(1)}$ is tested by comparing $p_{(1)}$ with $\alpha/n$. If $p_{(1)} > \alpha/n$, then all hypotheses are accepted and testing stops. Otherwise $H_{(1)}$ is rejected and one proceeds to test $H_{(2)}$ by comparing $p_{(2)}$ with $\alpha/(n-1)$. In general, $H_{(i)}, \ldots, H_{(n)}$ are accepted and testing stops if

$$p_{(i)} > \frac{\alpha}{n-i+1}\,; \tag{5}$$

otherwise $H_{(i)}$ is rejected and testing continues with $H_{(i+1)}$.

The Holm MTP incorrectly rejects any true hypothesis with probability $\alpha$ and hence is not separable. To see this, consider the problem of testing $H_i : \mu_i = 0$ versus $H_i' : \mu_i > 0$, $(1 \leq i \leq n)$. Suppose that $\mu_j = 0$ for some $j$ and $\mu_i \to \infty$, $i \neq j$. Then $p_i \to 0$ for $i \neq j$ and $p_j$ will be the largest $p$-value. Therefore, $p_j$ will be compared with $\alpha$ and $H_j$ will be rejected with probability $\alpha$.

To make the Holm MTP separable, we truncate its critical constants by taking their convex combination with the Bonferroni MTP constants as follows. In (5), for specified $\gamma$ $(0 \leq \gamma < 1)$, called the *truncation fraction*, we replace the critical constant for comparing $p_{(i)}$ with

$$w_i\alpha = \left[\frac{\gamma}{n-i+1} + \frac{1-\gamma}{n}\right]\alpha\,. \tag{6}$$

We refer to this procedure as the *truncated Holm MTP*. The power of this MTP is strictly increasing in $\gamma$. For $\gamma = 0$ and $\gamma = 1$, this MTP simplifies to the Bonferroni MTP and the Holm MTP, respectively.

Note that the truncated Holm MTP is a step-down shortcut to a closed procedure that tests any intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ using the weighted Bonferroni MTP with weights $w_i(I)$ and finds it significant if $p_i \leq w_i(I)\alpha$ for at least one $i \in I$, where

$$w_i(I) = \frac{\gamma}{|I|} + \frac{1-\gamma}{n}\,.$$

Recalling that a closed procedure rejects $H_I$ iff all $H_J$ for $J \supseteq I$ are significant, an upper bound on the error rate function of the truncated Holm MTP is given by

$$e^*(I) = \begin{cases} \sum_{i \in I} w_i(I)\alpha = [\gamma + (1-\gamma)\,|I|/n]\,\alpha & \text{if } |I| > 0, \\ 0 & \text{if } |I| = 0. \end{cases} \tag{7}$$

Therefore for any $I \subset N$ and $\gamma \in [0, 1)$,

$$e^*(I) < [\gamma + (1-\gamma)]\,\alpha = \alpha\,.$$

Hence the truncated Holm MTP is separable. Note that the $e^*(I)$ function of the truncated Holm MTP satisfies (4).

## 5.6.2 Truncated Hochberg

The Hochberg MTP uses the same Holm critical constants (5) but tests the hypotheses in a step-up manner (it begins with the hypotheses corresponding to the least significant $p$-value). The Hochberg MTP is more powerful than the Holm MTP. However, the Hochberg MTP (as well as the Hommel MTP mentioned in the sequel) requires independence among the $p$-values since it is based on the Simes (1986) test, which assumesindependence; Sarkar and Chang (1997) have shown that the independence assumption can be relaxed to the positive dependence assumption. The error rate function of the Hochberg MTP also equals $\alpha$ under the same configuration for which the error rate function of the Holm MTP equals $\alpha$, namely one hypothesis is true and the others are infinitely false. Hence the Hochberg MTP is not separable.

A *truncated Hochberg MTP* uses the same critical constants (6) as does the truncated Holm MTP, but is more powerful than the latter. At the first stage, this MTP rejects all hypotheses and stops testing if

$$p_{(n)} \le w_n \alpha = \left[\gamma + \frac{(1-\gamma)}{n}\right]\alpha\,;$$

otherwise it accepts $H_{(n)}$ and goes on to test $H_{(n-1)}$. In general, having accepted $H_{(n)},\ldots,H_{(i+1)}$, it rejects $H_{(i)},\ldots,H_{(1)}$ and stops testing if $p_{(i)} \le w_i\alpha$ where $w_i$ is defined in (6); otherwise, it accepts $H_{(i)}$ and goes on to test $H_{(i-1)}$.

As is well-known, the Hochberg (1988) MTP is a conservative shortcut to the closed procedure in which each intersection hypothesis $H_I = \bigcap_{i\in I} H_i$ is tested using the Simes (1986) test. Similarly, the truncated Hochberg MTP is a conservative shortcut to the closed procedure based on the truncated Simes test in which any intersection hypothesis $H_I = \bigcap_{i\in I} H_i$ is rejected if

$$p_{(i)}(I) \le \left[\frac{\gamma}{|I|-i+1} + \frac{1-\gamma}{n}\right]\alpha \quad \text{for at least one} \quad i \in I\,,$$

where $p_{(i)}(I)$ is the $i$th ordered $p$-value in the index set $I$ and $\gamma \in [0,1)$. Therefore an upper bound on the error rate function of the truncated Hochberg MTP is given by

$$e^*(I) = 1 - P\left\{p_{(i)}(I) > \left[\frac{\gamma}{|I|-i+1} + \frac{1-\gamma}{n}\right]\alpha \text{ for all } i \in I\right\}$$

if $|I| > 0$ and $e^*(I) = 0$ if $|I| = 0$. Using the Simes (1986) identity, it is readily seen that $e^*(I) < \alpha$ for $I \subset N$ and $\gamma \in [0,1)$. Therefore, the truncated Hochberg MTP is separable. In general, $e^*(I)$ above does not satisfy the monotonicity condition (4); therefore the latter may need to be enforced as explained following its statement. For independent $p$-values, $e^*(I)$ can be computed using the recursive formula given in the following result due to Sen (1999).

**Präposition 3.1** *Let $U_{(1)} < \ldots < U_{(k)}$ denote the order statistics of $k \ge 1$ i.i.d. observations from a uniform $(0,1)$ distribution. For any $0 < a_1 < \ldots < a_k < 1$,*

$$P(a_1,\ldots,a_k) = P(U_{(i)} > a_i \text{ for all } i = 1,\ldots,k) = k!H_k(1)\,,$$

*where*

$$H_i(u) = \int_{a_i}^{u} H_{i-1}(v)\,\mathrm{d}v, \quad i = 1,\ldots,k \quad \text{and} \quad H_0(u) = I(u \ge a_1)\,,$$

*and $I(\cdot)$ is an indicator function.*

## 5.6.3 Truncated Fallback

Wiens (2003) proposed a step-down MTP in which the hypotheses are *a priori* ordered (in contrast to the Holm MTP which orders the hypotheses according to their observed $p$-values). The total $\alpha$ is allocated to the $n$ ordered hypotheses as $\alpha_1, \ldots, \alpha_n$ such that $\sum_{i=1}^{n} \alpha_i = \alpha$. For simplicity, we shall restrict to the equal allocation case: $\alpha_i = \alpha/n$ $(1 \le i \le n)$. The MTP begins by testing $H_1$ at level $\alpha/n$; more generally, it tests a hypothesis $H_i$ at level $(i-t)\,\alpha/n$, where $t$ is the index of the last accepted hypothesis ($t = 0$ if none of the previous hypotheses is accepted). This MTP also follows the "use it or lose it" principle so that the $\alpha_i$'s for the rejected hypotheses in the sequence are carried forward to test the later hypotheses.

This fallback MTP is not separable. Suppose, for example, that $H_1, \ldots, H_{n-1}$ are infinitely false and $H_n$ is true, so that $p_1, \ldots, p_{n-1} \to 0$ and $p_n$ is compared with $\alpha$. Then the probability of rejecting $H_n$ is $\alpha$.

The *truncated fallback MTP* tests $H_i$ at level

$$w_i(t)\alpha = \left(\frac{\gamma(i-t)}{n} + \frac{1-\gamma}{n}\right)\alpha\,,$$

where $0 \le \gamma < 1$ and $t$ is the index of the last accepted hypothesis ($t = 0$ if none of the previous hypotheses is accepted).

Extending the arguments in the proof of Theorem 1 of Wiens and Dmitrienko (2005), it can be shown that this MTP is a shortcut to a closed procedure which rejects any intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ if $p_i \le w_i(t_I)\alpha$ for at least one $i \in I$, where $t_I$ is the largest index in $I$ that is smaller than $i$ if $i$ is not the smallest index in $I$ and $t_I = 0$ if $i$ is the smallest index in $I$. Therefore, using the closure principle and the Bonferroni inequality, an upper bound on the error rate function of the truncated fallback MTP is given by

$$e^*(I) = \begin{cases} \sum_{i \in I} w_i(t_I)\alpha & \text{if} \quad |I| > 0 \\ 0 & \text{if} \quad |I| = 0 \end{cases}.$$

Note that $e^*(I) < \alpha$ for any $I \subset N$, and hence the truncated fallback MTP is separable if $\gamma \in [0, 1)$. Also, $e^*(I)$ satisfies the conditions (4).

## 5.6.4 Truncated Dunnett

The Dunnett (1955) MTP can be thought of as a parametric version of the Bonferroni MTP and the step-down Dunnett MTP (Marcus, Peritz and Gabriel, 1976) is analogous to the Holm MTP. The step-down Dunnett MTP does not satisfy the separability condition because it incorrectly rejects any true hypothesis with probability $\alpha$. The *truncated Dunnett MTP* is defined as a convex combination of the regular and step-down Dunnett MTPs with $0 \leq \gamma < 1$. Let $t_1, \ldots, t_n$ be the test statistics associated with $H_1, \ldots, H_n$. Let $t_{(1)} > \ldots > t_{(n)}$ be the ordered test statistics and $H_{(1)}, \ldots, H_{(n)}$ denote the corresponding null hypotheses. Further, let $T_1, \ldots, T_n$ denote the random variables corresponding to the observed statistics $t_1, \ldots, t_n$ and assume that they follow a multivariate $t$-distribution under the global null hypothesis.

For any $I \subseteq N$, let $c(I)$ be the critical value for the maximum test statistic associated with $H_i$, $i \in I$, such that

$$P\left(\max_{i \in I} T_i > c(I) \mid H_I = \bigcap_{i \in I} H_i\right) = \alpha\,.$$

The computation of these critical values can be performed using the algorithm for calculating multivariate $t$ probabilities due to Genz and Bretz (2002).

For any $i = 1, \ldots, n$, let $I_{(i)} = \{(i), \ldots, (n)\}$. The truncated Dunnett MTP begins with the hypothesis, $H_{(1)}$, corresponding to the most significant $t$-statistic, $t_{(1)}$. This hypothesis is rejected if $t_{(1)} > c(I_{(1)})$ and is accepted otherwise. If $H_{(1)}$ is rejected, the next hypothesis in the sequence, $H_{(2)}$, is tested. In general, the MTP rejects $H_{(j)}$ if

$$t_{(i)} > (1-\gamma)c(I_{(1)}) + \gamma c(I_{(i)}) \text{ for all } i = 1, \ldots, j\,.$$

Otherwise, $H_{(j)}, \ldots, H_{(n)}$ are accepted and testing stops. The truncated Dunnett MTP simplifies to the regular Dunnett MTP if $\gamma = 0$ and to the step-down Dunnett MTP if $\gamma = 1$.

The computation of the error rate function for the truncated Dunnett MTP can be performed by using its closed representation. This MTP is equivalent to a closed testing procedure that rejects the intersection hypothesis $\bigcap_{i \in I} H_i$, $I \subseteq N$, if

$$\max_{i \in I}\ T_i > (1-\gamma)\, c(N) + \gamma c(I)\,.$$

Therefore, using the same argument as used for the Holm MTP, an upper bound on the error rate function of the truncated Dunnett MTP is given by

$$e^*(I) = \begin{cases} P(\max_{i \in I} T_i > (1-\gamma)\, c(N) + \gamma c(I)) & \text{if } |I| > 0 \\ 0 & \text{if } |I| = 0 \end{cases}.$$

It is easy to see that the truncated Dunnett MTP satisfies the separability condition for any $I \subset N$ if $0 \leq \gamma < 1$. However, the upper bound $e^*(I)$ on its error rate function may not satisfy the monotonicity condition (4), in which case the latter may need to be enforced as explained following its statement.

## 5.7 Weight assignment algorithm for Bonferroni tree gatekeeping procedures

The algorithm is given by Dmitrienko et al. (Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;, 2010).

Assuming the multiple testing problem formulated in Section 5.5.1, consider the closed family associated with the $n$ null hypotheses in Families $F_1, \ldots, F_m$. For each intersection hypothesis $H$, define the indicator functions $\delta_{ij}(H)$ and $\xi_{ij}(H)$ as follows. Let $\delta_{ij}(H) = 1$ if $H$ contains $H_{ij}$ and 0 otherwise, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Further, for $i = 2, \ldots, m$ and $j = 1, \ldots, n_i$, let $\xi_{ij}(H) = 0$ if $H$ contains at least one hypothesis from $R_{ij}^S$ or all hypotheses from $R_{ij}^P$. Otherwise, let $\xi_{ij}(H) = 1$. A Bonferroni tree gatekeeping procedure is defined by specifying a weighted Bonferroni test for each intersection hypothesis $H$. To accomplish this, it is sufficient to set up an $n$-dimensional weight vector for $H$ denoted by $v_{ij}(H)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The $p$-value for $H$ is given by

$$p_H = \min_{i,j} \frac{p_{ij}}{v_{ij}(H)},$$

where $p_{ij}$ is the $p$-value for $H_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Note that $p_{ij}/v_{ij}(H)$ can be set to 1 if $v_{ij}(H) = 0$. Based on the closure principle, the adjusted $p$-value for $H_{ij}$ is found by computing the maximum $p_H$ over all intersection hypotheses containing $H_{ij}$.

The weight vector for $H$ is constructed sequentially by defining $m$ subvectors

$$(v_{i1}, \ldots, v_{in_i}), \quad i = 1, \ldots, m,$$

using the algorithm described below (it is assumed in the algorithm that $0/0 = 0$).

**Family $F_1$.** Let

$$v_{1j}(H) = v_1^*(H) w_{1j} \delta_{1j}(H), \quad j = 1, \ldots, n_1,$$

where $v_1^*(H) = 1$, and let $v_2^*(H)$ denote the remaining weight, i.e.,

$$v_2^*(H) = v_1^*(H) \left( 1 - \sum_{j=1}^{n_1} w_{1j} \delta_{1j}(H) \right).$$

**Family $F_k$, $k = 2, \ldots, m-1$.** Let

$$v_{kj}(H) = v_k^*(H) w_{kj} \delta_{kj}(H) \xi_{kj}(H), \quad j = 1, \ldots, n_k.$$

The remaining weight is given by

$$v_{k+1}^*(H) = v_k^*(H) \left( 1 - \sum_{j=1}^{n_k} w_{kj} \delta_{kj}(H) \right).$$

**Family $F_m$.** Let

$$v_{mj}(H) = v_m^*(H) w_{mj} \delta_{mj}(H) \xi_{mj}(H) / \sum_{l=1}^{n_m} w_{ml} \delta_{ml}(H) \xi_{ml}(H),$$

where $j = 1, \ldots, n_m$.