|  |
| --- |
| Biostatistics |
| MULTIPLICITY |
| IN CLINICAL TRIALS |

|  |
| --- |
| Yupeng Li  12-15-2021 |

[1 Introduction 5](#_Toc109914156)

[1.1 Testing A Single Hypothesis 5](#_Toc109914157)

[1.2 Testing Multiple Hypotheses 5](#_Toc109914158)

[1.3 Familywise error rate (FWER) 6](#_Toc109914159)

[1.4 False discovery rate (FDR) and false discovery proportion (FDP) 7](#_Toc109914160)

[1.5 The reasons to consider multiple testing adjustments 8](#_Toc109914161)

[1.5.1 FDA presenters’ view: example of substantial underreporting of true error rate 8](#_Toc109914162)

[1.5.1.1 Example 1 9](#_Toc109914163)

[1.5.1.2 Example 2 10](#_Toc109914164)

[2 Multiple testing principles 11](#_Toc109914165)

[2.1 Union-intersection testing 11](#_Toc109914166)

[2.1.1 When will Type I error occur 12](#_Toc109914167)

[2.2 Intersection-union testing 12](#_Toc109914168)

[2.3 Closure principle 12](#_Toc109914169)

[2.3.1 The mechanism of controlling 𝜶 level 13](#_Toc109914170)

[2.3.2 Bonferroni-based closed testing procedures 14](#_Toc109914171)

[2.3.2.1 Class of Bonferroni-based closed testing procedures 14](#_Toc109914172)

[2.3.2.2 Sequentially rejective Bonferroni-based closed testing procedures 16](#_Toc109914173)

[2.3.2.3 Graphical visualization 18](#_Toc109914174)

[2.3.3 Properties of closed testing procedures 18](#_Toc109914175)

[2.3.3.1 Monotone procedures 18](#_Toc109914176)

[2.3.3.2 Consonant procedures 19](#_Toc109914177)

[2.3.3.3 -exhaustive procedures 19](#_Toc109914178)

[2.4 Partitioning principle 19](#_Toc109914179)

[2.5 Summary 20](#_Toc109914180)

[3 Multiple Testing Procedures 22](#_Toc109914181)

[3.1 Classiﬁcation of multiple testing procedures 22](#_Toc109914182)

[3.1.1 Single-step and stepwise procedures 22](#_Toc109914183)

[3.1.2 Distributional assumptions 23](#_Toc109914184)

[3.1.3 Hierarchical or non-hierarchical structures of testing multiple hypotheses 23](#_Toc109914185)

[3.2 Multiple Testing Procedures 26](#_Toc109914186)

[3.2.1 Bonferroni method 26](#_Toc109914187)

[3.2.2 Holm (Bonferroni based) procedure 27](#_Toc109914188)

[3.2.2.1 Holm's weighted procedure 28](#_Toc109914189)

[3.2.2.2 How closed Bonferroni procedure gives the Holm's procedure 28](#_Toc109914190)

[3.2.3 Fixed-sequence procedure 29](#_Toc109914191)

[3.2.4 Fallback procedure 30](#_Toc109914192)

[3.2.5 Simes method 30](#_Toc109914193)

[3.2.6 Hochberg Procedure 33](#_Toc109914194)

[3.2.7 Hommel procedure 34](#_Toc109914195)

[3.2.8 Dunnett Test 35](#_Toc109914196)

[3.3 Novel multiple testing techniques 38](#_Toc109914197)

[3.3.1 Gatekeeping Approaches 38](#_Toc109914198)

[3.3.1.1 Motivation 38](#_Toc109914199)

[3.3.1.2 Clinical trials with serial gatekeepers 39](#_Toc109914200)

[3.3.1.3 Clinical trials with parallel gatekeepers 44](#_Toc109914201)

[3.3.1.4 Clinical trials with tree-structured gatekeepers 53](#_Toc109914202)

[3.3.2 Graphical approaches 57](#_Toc109914203)

[3.3.2.1 The definition of graphical approach 57](#_Toc109914204)

[3.3.2.2 Algorithm to update the graph 59](#_Toc109914205)

[3.3.2.3 Example 1 - Fixed sequence test 59](#_Toc109914206)

[3.3.2.4 Example 2 - Fallback procedure 60](#_Toc109914207)

[3.3.2.5 Example 3 - Bonferroni–Holm procedure 60](#_Toc109914208)

[3.3.2.6 Shifting signiﬁcance levels between families of hypotheses 62](#_Toc109914209)

[3.3.2.7 Late phase development of a new drug for the indication of multiple sclerosis 65](#_Toc109914210)

[3.3.3 Gatekeeping approach versus Graphical approach 69](#_Toc109914211)

[3.3.4 Adaptive Designs and Confirmatory Hypothesis Testing 71](#_Toc109914212)

[3.3.4.1 Causes of multiplicity and bias in adaptive designs 71](#_Toc109914213)

[3.3.4.2 Repeated hypothesis testing at interim analyses in group sequential designs 72](#_Toc109914214)

[3.3.4.3 Group sequential Holm procedure with multiple primary endpoints 75](#_Toc109914215)

[3.3.4.4 Weighted parametric group sequential design (WPGSD) 94](#_Toc109914216)

[4 Bibliography 98](#_Toc109914217)

[5 Appendix 100](#_Toc109914218)

[5.1 Error Spending Methods 100](#_Toc109914219)

[5.2 Adjusted significance levels and p-values 102](#_Toc109914220)

[5.3 Proof of 2-stage gatekeeping procedure controls the FWER at the α level 103](#_Toc109914221)

[5.4 A detailed example of sequentially rejective Bonferroni-based closed testing procedures 104](#_Toc109914222)

[5.5 Closure testing principle in group sequential Holm procedure 106](#_Toc109914223)

[5.6 Truncated multiple test procedures 108](#_Toc109914224)

[5.6.1 Truncated Holm 108](#_Toc109914225)

[5.6.2 Truncated Hochberg 109](#_Toc109914226)

[5.6.3 Truncated Fallback 110](#_Toc109914227)

[5.6.4 Truncated Dunnett 111](#_Toc109914228)

[5.7 Weight assignment algorithm for Bonferroni tree gatekeeping procedures 112](#_Toc109914229)

TABLE OF FIGURES

Figure 1 - The relationship among FDR, FDP and alpha 8

Figure 2 - Two key statistical approaches for the analyses of the Primary Endpoint and Secondary Endpoint hypotheses of clinical trials. 9

Figure 3 - PE is primary endpoint; SE is secondary endpoint. 10

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. 10

Figure 5 - The FWER is the probability of rejecting at least one hypothesis when all hypotheses are true. 11

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. 11

Figure 7 - Graphical presentation 12

Figure 8 - closure principle with 2 hypotheses and it connection to -recycling and the graphical method. 15

Figure 9 - and are primary hypotheses and is the secondary hypothesis. Note that is tested only when at least one primary hypothesis is rejected. 17

Figure 10 - Closure principle tables for Figure 9 with Bonferroni weights satisfying consonance. 17

Figure 11 - Graphical representation of the Figure 9. 17

Figure 12 - Graphical presentation 20

Figure 13 - Graphical presentation for the three mutually exclusive hypotheses 20

Figure 14 - Union-intersection testing 21

Figure 15 - Intersection-union testing 21

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. 31

**Figure 17** - Bonferroni's rejection region. The visible area is the rejection region. 32

**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. 32

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, comparisons, correlation ; dashed curve, comparisons, correlation ). The Simes test is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025. 33

Figure 20 - Scenario 1, left panel; Scenario 2, right panel. Family 1 serves as a gatekeeper in both scenarios. Note that in Scenario 1, both tests can be carried out at the pre-specific level since this testing procedure is a special case of the fixed-sequence approach. But multiplicity adjustment should be implemented in Scenario 2. 40

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-P comparison), (M-P comparison) and (L-P comparison) for the th endpoint are included in , . The hypotheses are equally weighted within each family and the FWER is set at a two-sided . 43

Figure 22 - A problem with a parallel gatekeeper . 44

Figure 23 - 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. 52

Figure 24 - Decision tree in the combination-therapy clinical trial example (Noninf, Noninferiority; Sup, Superiority). 53

Figure 25 - Graphical illustration of the weighted Bonferroni–Holm procedure with two hypotheses. 55

Figure 26 - Example multiple test procedures to illustrate 57

Figure 27 - Graphical illustration of the ﬁxed sequence test with three hypotheses 57

Figure 28 - A break down of graphical illustration of the fixed sequence test; Note that Green=rejection and Red=no rejection (and stop) in this figure. 58

Figure 29 - Improvement of the fallback procedure by Wiens BL, Dmitrienko A. with . 58

Figure 30 - Graphical illustration of the Bonferroni–Holm procedure with 𝑚=3 hypotheses and initial allocation 59

Figure 31 - A vivid demonstration 59

Figure 32 - The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values , , and 60

Figure 33 - A vivid presentation of updating a graph with -edge; The testing starts from hypothesis . 61

Figure 34 - A vivid presentation of updating a graph with ϵ-edge. 62

Figure 35 - Graphical illustration of the gatekeeping procedure with four hypotheses. 63

Figure 36 - Graphical illustration of the improved gatekeeping procedure with four hypotheses by adding 𝜀-edge. The tests start from hypothesis and the 𝜀-edge can be treated as a “blocker”. 63

Figure 37 - A fixed sequence test to the six hypotheses. 64

Figure 38 - Visualization of different implementations for Strategy 2. 65

Figure 39 - Visualization of implementations for Strategy 3. 66

**Figure 40 - Closed test.** 74

Figure 41 - Rejection region at the interim analysis for MONET1 (GSHv). 78

Figure 42 - A flowchart of multiple testing strategy. 85

Figure 43 - Timing of Primary and Key Secondary Endpoints Analyses. 86

Figure 44 - Type I Error Reallocation Strategy Following Closed Testing Principle. 86

# 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

## 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).

## 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 𝛼 = 0.05. If we have two null hypotheses and do two independent tests, each at level 𝛼 = 0.05:

The probability of rejecting at least one true null hypothesis is

Which is . 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

## 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 is the index set of true null hypotheses, we require that

,

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

* Let be the mean improvement in the placebo arm, be the mean improvement in the ith dose group.
* is a non-negative constant deﬁning the clinically important diﬀerence.

The supremum is taken over all satisfying null hypothesis for and for , and the maximum is taken over all index sets .

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 𝛼 in advance. The choice of 𝛼 allocation depends on specific problem and it may lead to different results. We need to determine the methods of adjustment during trial design.*

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

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

The FDR is said to be controlled at the 𝛾 level if

To ensure the control of the FDR at 𝛾 level, one can choose an acceptable probability of exceedance 𝛼 and require that

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

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

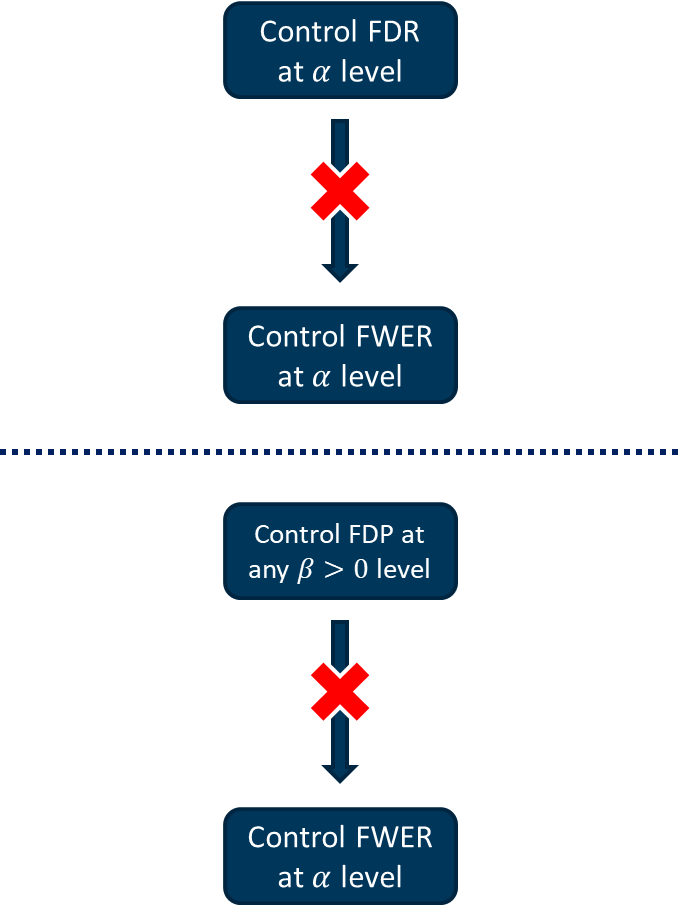


Figure 1 - The relationship among FDR, FDP and alpha

## 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.

### 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.

Graphical user interface, text, application

Description automatically generated

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

#### 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 (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?**

Text

Description automatically generated

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

#### 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: and .

Diagram

Description automatically generated

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.

# 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:

* denote the hypotheses corresponding to the multiple objectives and they are tested against the alternative hypotheses
* denotes the global hypothesis

## Union-intersection testing

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

Text

Description automatically generated

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 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.

Text, letter

Description automatically generated

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.

### 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 iff which is the rejection region. is the statistics for hypothesis and is the () quantile from the marginal distribution of

In this example, we will test

We suppose is true so that Type I error will occur when or is rejected.

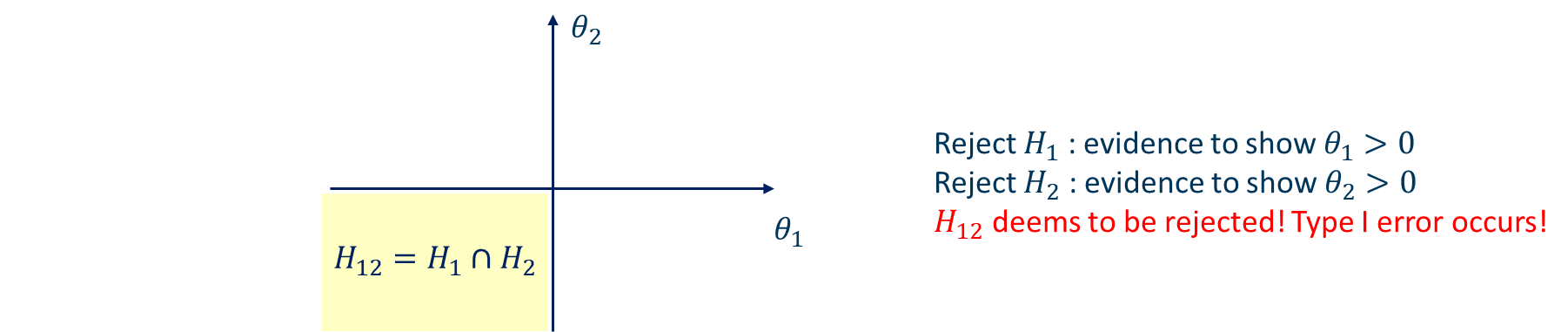


Figure 7 - Graphical presentation

## 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 , defined as the union of the hypotheses, is tested v.s. the union of the alternative hypotheses ():

When the global hypothesis is rejected, one concludes that all 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.

## 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 , consider an intersection hypothesis defined as
* Establish implication relationships. An intersection hypothesis that contains another intersection hypothesis is said to imply it, i.e., implies if
* Define local level tests for individual intersection hypotheses. Let denotes the p-value produced by the associated local test and reject iff (if and only if) for all .

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

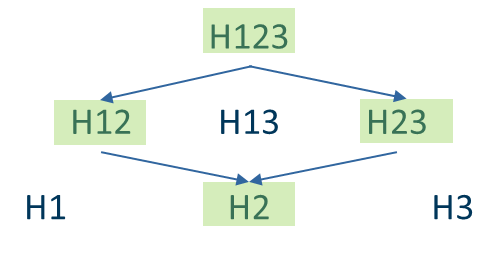
reject iff for all index sets that include

### The mechanism of controlling 𝜶 level

Suppose three null hypotheses to be tested using closure principle

Here where ; And a fully closure set is . Obviously any

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



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

{reject using closure} = {reject using closure} {reject using closure} {reject using closure} {reject using closure} {reject using closure} ；

Similarly, we have

{reject using closure} {reject using closure};

{reject using closure} {reject using closure}.

Therefore,

{reject using closure} {reject using closure} {reject using closure} {reject using closure}.

That is,

### Bonferroni-based closed testing procedures

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.

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

**Problem**

* Testing hypotheses and let .
* Applying the closure principle leads to consideration of the intersection hypotheses , where .
* For each intersection hypothesis we assume a collection of non-negative weights (These weights quantify the relative importance of the hypotheses included in the intersection ), where and .

In this section we assume that each intersection hypothesis is tested with a weighted Bonferroni test[[1]](#footnote-2).

Consequently, we obtain the multiplicity adjusted p-values

For the weighted Bonferroni test for , where

This defines Class 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 .

**Example**

Consider the simple two-hypothesis problem where the intersection hypothesis is with and .

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

If is rejected, so is either or , since they are tested subsequently at level . In other words, if is rejected (the smaller of the two p-values is less than ), the remaining elementary hypothesis is tested at level .

Suppose the significance level is .

If is rejected when

It means that

Based on this derivation, we can conclude that or . Therefore we can say that If is rejected at level , so is either or .

Text

Description automatically generated

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

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

It can further be shown that under a mild monotonicity condition on the weights the closure principle leads to powerful **consonant** [[2]](#footnote-3)multiple testing procedures.

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 of shortcut procedures characterized by the property

This condition ensures that if an intersection hypothesis is rejected, there is an index such that () and the corresponding elementary is deemed to be rejected at level by the closed testing procedure.

Therefore, short-cut procedures of order can be constructed:

**Instead of testing hypotheses (as usually required by the closure principle), it is sufficient to test the elementary hypotheses in steps.**

**Algorithm**

Therefore, shortcut procedures from can be carried out with the following -step procedure: **Start:** testing the global intersection , ;

**Subsequent Step:** If it is rejected, there is an index such that is rejected by the closed testing procedure;

**Subsequent Step:** continues testing the global intersection until the first non-rejection.

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 .

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.

Diagram

Description automatically generated

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

Graphical user interface, application

Description automatically generated

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

Diagram

Description automatically generated

Figure 11 - Graphical representation of the Figure 9.

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

#### Graphical visualization

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

It was shown above that Class includes a variety of Bonferroni-based testing procedures, such as

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

Using procedures in this class , 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 for each of the intersection hypotheses , .

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.

### Properties of closed testing procedures

This section briefly describes important properties of closed testing procedures.

#### Monotone procedures

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

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.

#### Consonant procedures

A closed testing procedure is termed **consonant** if the rejection of an intersection hypothesis withand always leads to the rejection of at least one implied by , i.e., with .

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 , , without rejecting any other intersection hypotheses.

#### -exhaustive procedures

An -exhaustive procedure is a closed testing procedure based on intersection hypothesis tests the size of which is exactly (Eugene Grechanovsky & Yosef Hochberg, 1999). In other words, for any intersection hypothesis ,.

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

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

## 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

Into 3 mutually exclusive hypotheses:

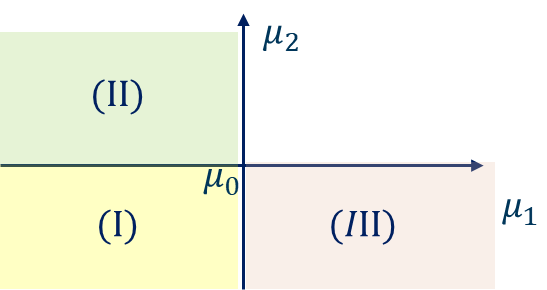


Figure 12 - Graphical presentation

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,

1. If is rejected, we conclude that or
2. If and are rejected, we conclude that (similarly, rejecting and implies that );
3. If , and are all rejected, we conclude that and .

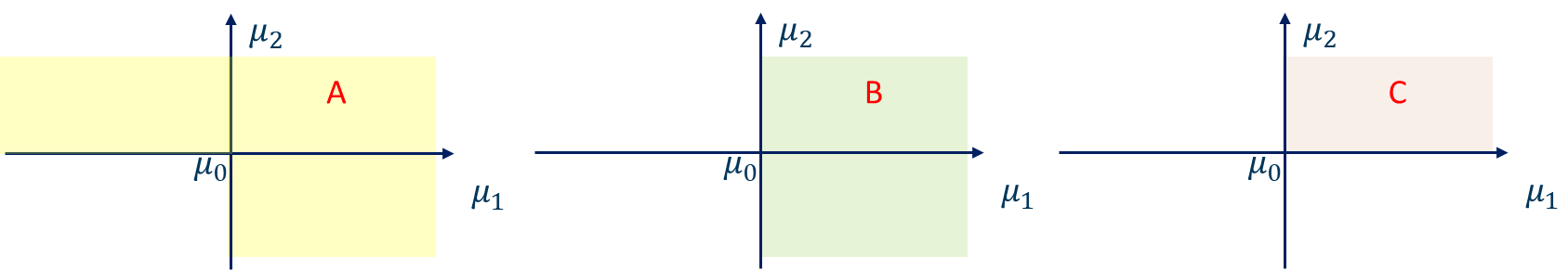


Figure 13 - Graphical presentation for the three mutually exclusive hypotheses

**General Case**

For hypotheses , the steps of partitioning principle are:

* For a given parameter space , choose an appropriate partition ;
* Test at unadjusted level;
* If all intersecting which are rejected, we can conclude that 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 level.

## Summary

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

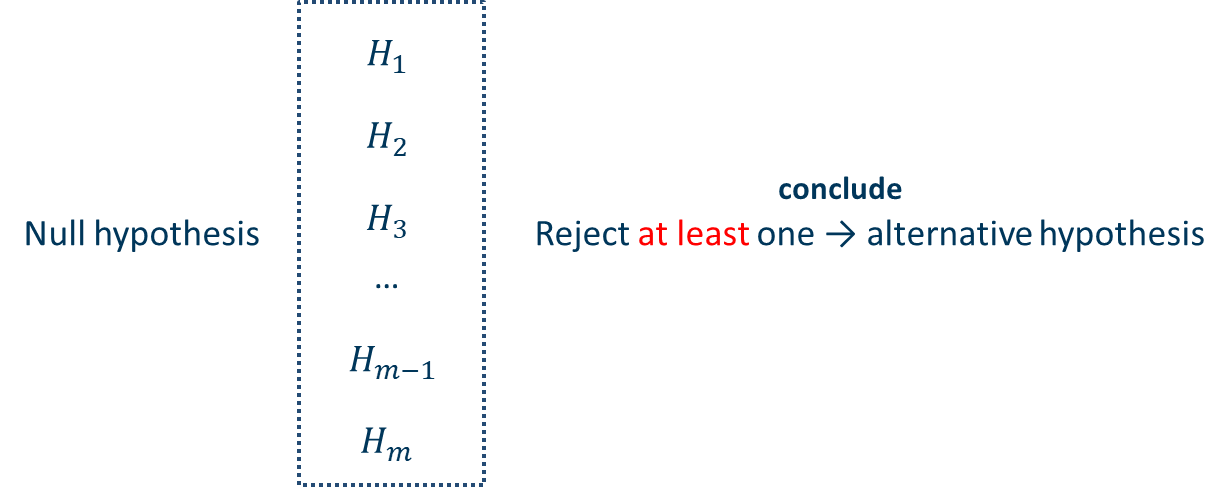


Figure 14 - Union-intersection testing



Figure 15 - Intersection-union testing

# Multiple Testing Procedures

## 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.

### 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.

### 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.

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

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

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**

* Simple procedures for hierarchical hypotheses
  + Fixed-sequence procedure
  + 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

## Multiple Testing Procedures

### Bonferroni method

It uses for all inference; for :

Reject if .

With adjusted p-values ,

Reject if .

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**:

where denotes the event of rejecting .

For ,

According to the definition of Type 1 error， FWER=Prob{reject H1 or reject H2}when H1 and H2 are true，which is to say,

By Boole’s inequality,

**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.

### Holm (Bonferroni based) procedure

Let denote the ordered unadjusted p-values with associated null hypotheses .

Then we have the following stepwise procedure:

Table 1 - Raw p-values.

|  |  |
| --- | --- |
| If | Reject and continue; else stop |
| If | Reject and continue; else stop |
| … |  |
| If | Reject and continue; else stop |
| … |  |
| If | Reject and continue; else stop |
|  |  |

With , define adjusted p-values using

* …

Where

;

.

**Properties**

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

* Bonferroni uses the same threshold for all hypotheses;
* Holm uses the larger thresholds .

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

#### 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 and order . The hypothesis and weight correspond to . Reject when

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

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

Theorem 1 – denote the number of incorrectly reject true hypotheses, the number of true hypotheses and the number of hypotheses respectively.

Text

Description automatically generated

For each non-empty index set , consider an intersection hypothesis (global null) defined as

For each index set , consider a valid level-α test for testing (reject if )

We use Bonferroni’s procedure to construct by Theorem 1:

We now show that we can avoid testing number of tests and simply run tests, where is the number of hypotheses.

1. Fix any and consider any such that , i.e. the th smallest p-value (among the p-values) is in fact the minimum p-value among those indexed in .
2. Let denote the index set corresponding to the largest p-values. We claim that

This is because if then

Note that since is the smallest p-value in , we must have .

1. Since by the definition of Bonferroni’s procedure[[3]](#footnote-4) which can control type I error in a strong sense
2. Using the previous claim, we note the following:

is rejected by closed test

;

**While do**

;

Reject .

and … and

and … and

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.

### 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.

is pre-specified (this order normally reflects the clinical importance of the multiple analyses).

Testing begins with the first hypothesis, , 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 ,, is rejected at the th step if

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)

### 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）

A picture containing text, clock

Description automatically generated

问题是什么？

Improved Fallback procedure

Diagram

Description automatically generated

### Simes method

Let denote the ordered unadjusted p-values with associated null hypotheses .

Simes method uses all ordered p-values and

global hypothesis global hypothesis of no treatment effect (union-intersection, see Section 2.1) if for at least one =1,2,…m;

Where

and are independent.

**Properties**

* Simes’ adjusted p-value uses , which is less than or equal to Bonferroni’s .
* Simes cannot be used to test the individual hypotheses .
* Type I error rate is at most 𝛼 under independence or (certain types of) positive dependence of p-values.

**Simes V.S. Bonferroni**

Suppose and independence of and . The rejecting probability is visualized in Figure 16.

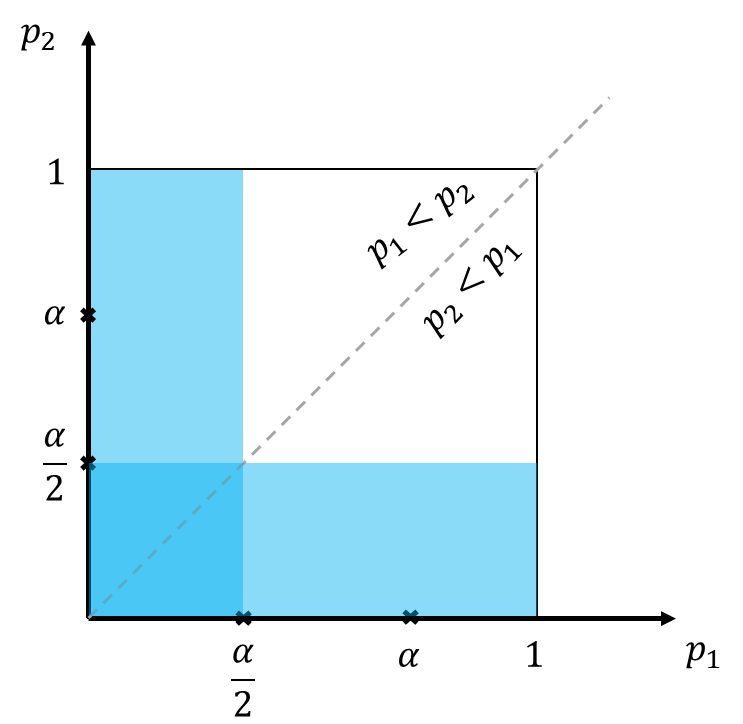
**Bonferroni test**: Reject if ; see **Figure 17**:

**Simes method**: Reject if or ; see **Figure 18**:

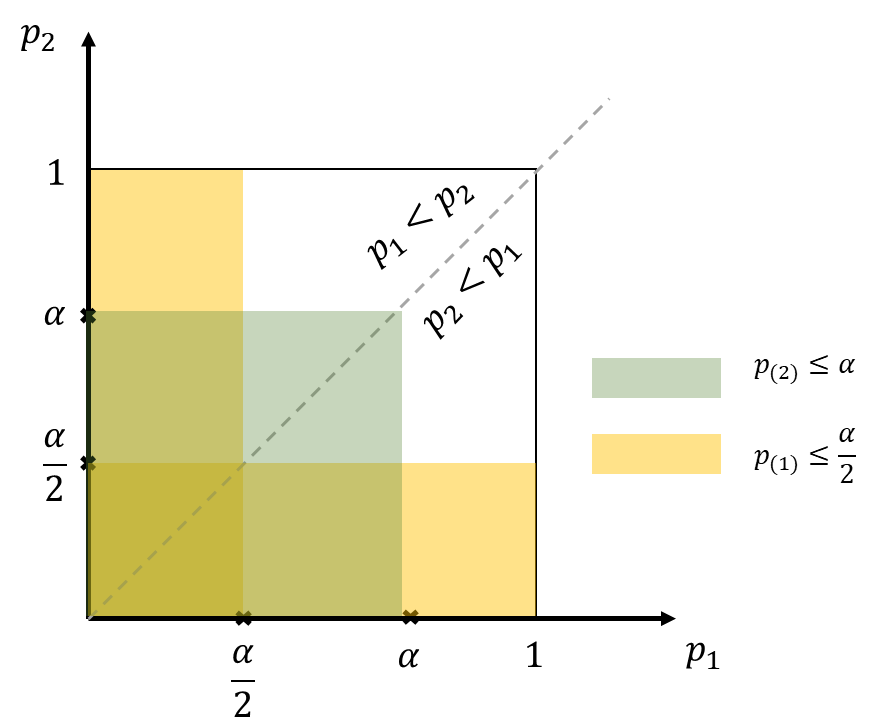
Chart, histogram

Description automatically generated

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 and common correlation coefficient in the case of normally distributed test statistics.

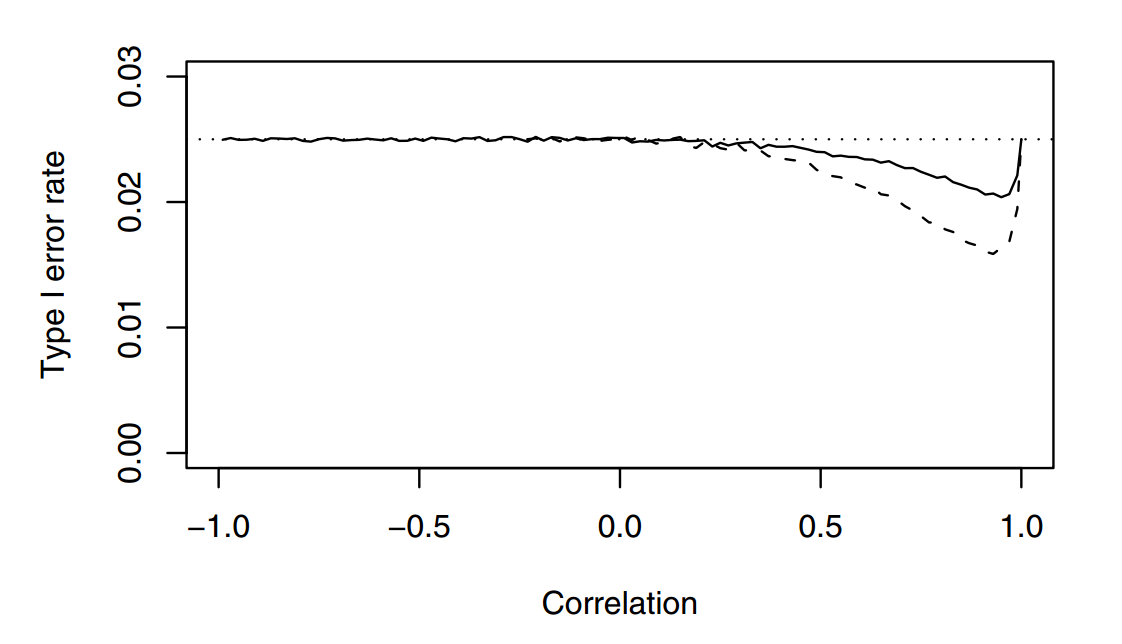


Figure 19 - Type I error rate of the Simes test under the global null hypothesis as a function of the number of comparisons and correlation (solid curve, comparisons, correlation ; dashed curve, comparisons, correlation ). The Simes test is carried out at the one-sided 0.025 level. The dotted line is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

### 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 denote the ordered unadjusted p-values with associated null hypotheses .

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 | Reject and stop; else retain and go to next step |
| If | Reject and stop; else retain and go to next step |
| … |  |
| If | Reject and stop; else retain and go to next step |
| … |  |
| If | Reject ; else retain |
|  |  |

Adjusted p-values are

;

.

**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.

### 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 | Retain and go to next step; else reject all hypotheses and stop. |
| … |  |
| If | Retain and go to next step; else reject all hypotheses and stop. |
| … |  |
| If | Retain ; else reject . |

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 if and only if whereas the Hommel procedure can reject this hypothesis when , e.g., is rejected if .

### 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.,

where

denotes observation in group ;

the effect of treatment group 𝑖;

are independent and identically normally distributed with mean 0 and variance , i.e. .

The ANOVA 𝐹-test tests the global null .

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

Consider the pairwise t-tests which have the following properties:

* , where and are the ordinary least squares (OLS) of and ;
* under , where denotes the univariate 𝑡-distribution with degrees of freedom;
* follows the 𝑚-variate 𝑡-distribution with 𝜈 degrees of freedom and correlations

**Rejection Rule**

For the 𝑚 individual null hypotheses, reject if where the quantile is computed such that

.

It should be noted that follows 𝑚-variate 𝑡-distribution (see Section 5.1 for details) with 𝜈 degrees of freedom and correlations , for .

In other words, is the quantile of the distribution of the maximum of -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.

**Stepwise Dunnett Test**

Let denote the ordered test statistics with associated null hypotheses .

Table 4 - Algorithm for stepwise Dunnett test procedure.

|  |  |
| --- | --- |
| If | Reject and continue; else stop. |
| If | Reject and continue; else stop. |
| … |  |
| If | Reject and continue; else stop. |
| … |  |
| If | Reject . |

\*Note that denotes the quantile of the distribution of the maximum of -distributed random variables and is computed from the corresponding multivariate -distribution. For example, if rejected, then the quantile is computed from a -variate -distribution.

Other properties for stepwise Dunnett test:

* It can be shown that , where is the quantile from the univariate -distribution with 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*.

## Novel multiple testing techniques

### 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.

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 to be used for the primary family. Thus, maximizing the study power for those critical endpoints.

#### 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).

#### 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.

##### 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:

: the true treatment differences for the comparisons of A versus B

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

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

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

**Sets of hypotheses**

A versus B (noninferiority): versus

A versus B (superiority): versus

A versus B (noninferiority): versus

Testing begins with the first family that includes the test for .

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 ; in Scenario 2, this family includes the tests for and . The decision rules are depicted in the Figure 20.

Diagram

Description automatically generated

Figure 20 - Scenario 1, left panel; Scenario 2, right panel. Family 1 serves as a gatekeeper in both scenarios. Note that in Scenario 1, both tests can be carried out at the pre-specific level since this testing procedure is a special case of the fixed-sequence approach. But multiplicity adjustment should be implemented in Scenario 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 . Each family ( 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.

Table

Description automatically generated

Note that is the weight representing the importance of hypotheses within and is the associated p-value. Multiplicity adjusted p-values for the hypotheses in is denoted by . 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 is a **serial gatekeeper**, hypothesis in are tested iff

**Single decision-making branch**

Within each family , hypotheses are tested at the nominal level.

For example, the hypotheses in can be tested using an Intersection-Union test; i.e. all hypotheses are rejected in if and all hypotheses are retained otherwise. Any FWER-controlling test can be used in , including all popular multiple tests.

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 :

* : the largest p-value in family .

: the adjusted p-value for hypothesis produced by the test used in the last family, ;

The adjusted p-value for is given by

**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 ).

Consider a multiple testing problem with families and assume that each one contains hypotheses, i.e., .

In this case there are branches and the th branch includes the hypotheses . Hypotheses within each branch are tested sequentially as follows:

* Consider the th branch,. The hypothesis is tested first at an level. If is rejected, the next hypothesis in the sequence, i.e., , is tested, otherwise testing within this branch stops.
* In general, the hypothesis is rejected if for all .

**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.

Diagram, schematic

Description automatically generated

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-P comparison), (M-P comparison) and (L-P comparison) for the th endpoint are included in , . The hypotheses are equally weighted within each family and the FWER is set at a two-sided .

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 .

Table

Description automatically generated

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.

#### Clinical trials with parallel gatekeepers

If family 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 , to proceed to . Hypothesis in are tested at the level iff

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.

Diagram, schematic

Description automatically generated

Figure 22 - A problem with a parallel gatekeeper .

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 , the resulting parallel gatekeeping procedures may lack transparency and their implementation can be computationally intensive since it takes order- steps to test 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)).

##### -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 (e.g., ) and its associated p-value is significant at this level, then this 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.

###### α-Propagation rules

|  |  |
| --- | --- |
| Family 1 | Family 2 |
| * Procedure 1 at level; * index set of null hypotheses accepted in family 1 (null hypotheses, is the index set); * Component procedure (Procedure 1): we should use separable procedure (will be defined later) with local FWER control. | * **Procedure 2** at level; * , where is error rate function (will be define later) of procedure 1 and ; * Family 2 (null hypotheses, is the index set); * Component procedure (Procedure 2): we can use any procedure with local FWER control. |

**Error rate function**

Assume that all null hypotheses , are true Error rate function is probability of rejecting at least one true null hypothesis

Where “Reject ” represents the event (the subset of the sample space) that corresponds to rejection of and the supremum of the probability is taken over the entire null space defined by , including any false hypotheses . Thus is the maximum probability of making at least one type I error in the sub-family .

Based on the definition above, we have:

**(1)**

:

, all null hypotheses are rejected in ; Here we have .

Null hypotheses in are tested at full level.

**(2)**

:

, no null hypotheses are rejected in ; Here we have .

Null hypotheses in are not tested.

For an example, error rate function of Bonferroni procedure is , where is the cardinality of set , i.e., the number of elements in index set .

**Separable procedures**

Procedure 1 is separable if provided is a proper subset[[4]](#footnote-5) of .

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

*\*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 .

|  |  |
| --- | --- |
| Bonferroni procedure | Holm procedure |
| Problem with three null hypotheses  Text  Description automatically generated | Problem with three null hypotheses  Text  Description automatically generated |
| * Error rate function of Bonferroni procedure is ; * Bonferroni procedure is separable because if is a proper subset of . | * Error rate function of Holm procedure is unless is empty; * Holm procedure is not separable. |

**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.

###### 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 level. (proof is provided on Section )

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

**Characteristics to define the multistage gatekeeping procedure**

families;

for ;

and be the index set corresponding to the accepted hypotheses in ;

The algorithm for applying the procedure is:

**Start**

Initialize

**Stage ~ Stage**

Test at level using any separable procedure with a suitable upper bound on the error rate function . Set

If , i.e. no hypotheses are rejected in , then apply the following procedures, stop testing and accept all hypotheses in

Otherwise, go to next step.

**Start**

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

**Remarks**

* If all hypotheses are rejected at the -th stage (), then and . Thus full 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.

##### 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.

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

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.

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 for some sufficiently large value of .

The adjusted p-value, , for hypotheses is the smallest (corresponding to the smallest ) for which 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.

##### 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[[5]](#footnote-7) )
  + 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 . 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.

Table

Description automatically generated

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

* The hypotheses in and 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 (, , and ) to evaluate the impact of this parameter on the outcomes of the four analyses.

**Scenario 1**

Let . The hypotheses and are tested using the truncated Holm test at . The smaller p-value, , is less than

And thus is rejected.

Further, , is compared to

The corresponding hypothesis cannot be rejected.

To find the fraction of that can be carried over to the hypotheses in , note that the set of retained hypotheses in includes only one hypothesis. Thus,

Where and .

Applying the regular Holm test in at , it is easy to verify that and are rejected at level .

##### 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 to and , respectively, at an -level of (see Figure 23).

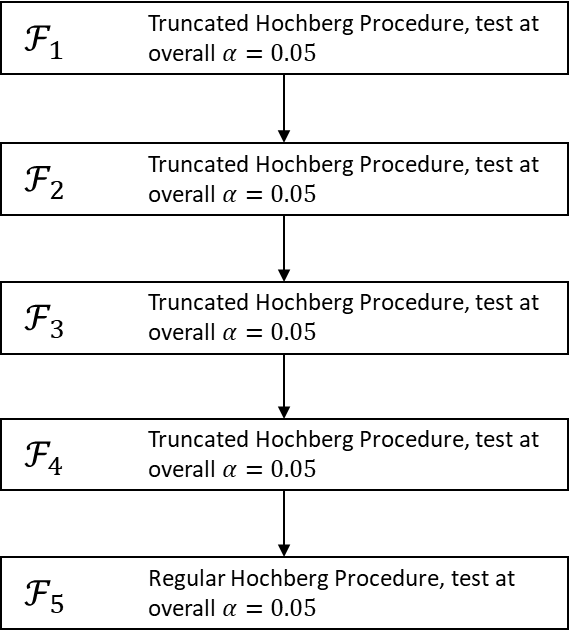


Figure 23 - A decision tree associated with parallel gatekeeping testing strategy.

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 and , 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 and check if ;
2. If , reject all hypotheses and stop. Else continue;
3. If , reject all hypotheses and stop. Else continue;
4. …
5. If , reject all hypotheses and stop. Else continue;
6. …

#### 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.

##### 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

* is a hypothesis in a family , where ;
  + denotes serial rejection set includes hypotheses from to ;
  + denotes parallel rejection set includes hypotheses from to (do not overlap with ).

The is testable if:

* All hypotheses are rejected in and at least one hypothesis is rejected in .

The following two conditions hold:

and

Tree gatekeeping procedures simplifies to serial gatekeeping procedures if and is empty for all hypotheses , , and to parallel gatekeeping procedures if is empty and for all hypotheses , .

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.,

* 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**

;

;

;

.

The hypothesis can be tested only if there is a significant result in and at least one significant result in as shown in Figure 23.

Diagram, schematic

Description automatically generated

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.

##### 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 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 includes (noninferiority hypothesis for A versus B).
* Family includes (superiority hypothesis for A versus B) and (noninferiority hypothesis for A+B versus B).
* Family includes (superiority hypothesis for A+B versus B) and (noninferiority hypothesis for A+B versus A).
* Family includes (superiority hypothesis for A+B versus A).

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

Diagram

Description automatically generated

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

Text, letter

Description automatically generated

and the parallel rejections sets are empty. For an example, becomes testable if all hypotheses (which is actually ) in are rejected and at least one hypothesis in 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.

Table

Description automatically generated

The table shows that:

* , is rejected at the two-sided 0.05 level and thus the hypotheses and become testable.
* Both of and are also rejected and, since is included in the serial rejection sets of the hypotheses in , the tree gatekeeping procedure tests and at the next step.
* The adjusted p-value for is significant but the adjusted p-value for is not. Since the hypothesis depends on , the former is retained without testing.

### 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.

#### 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 to the individual hypotheses;
* a sequentially rejective multiple test procedure (since after rejecting, for example, H1, only H2 remains to be tested).

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

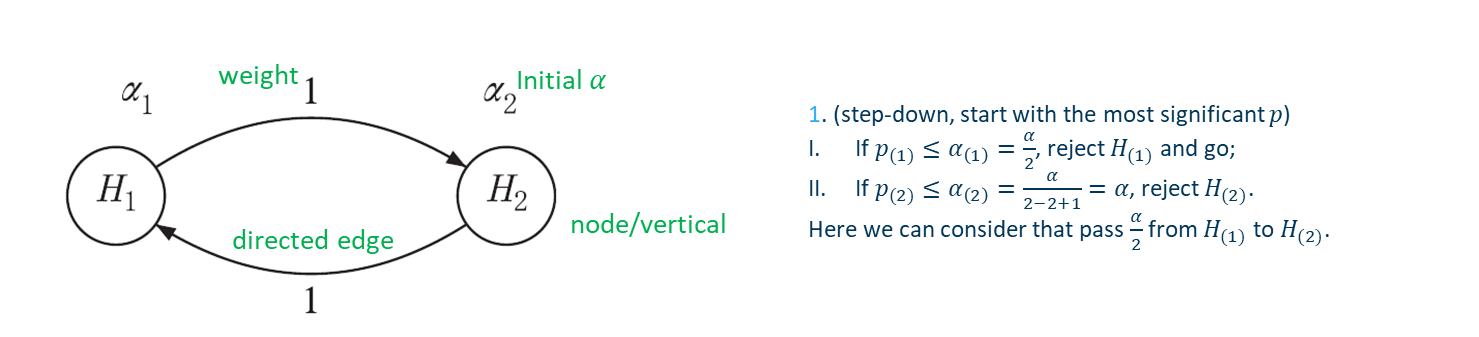
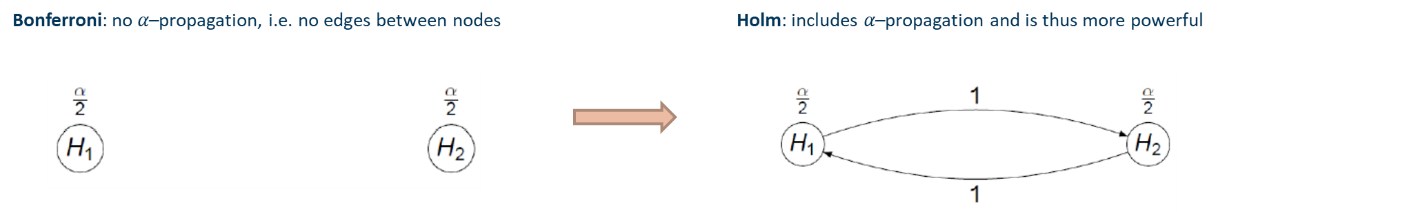
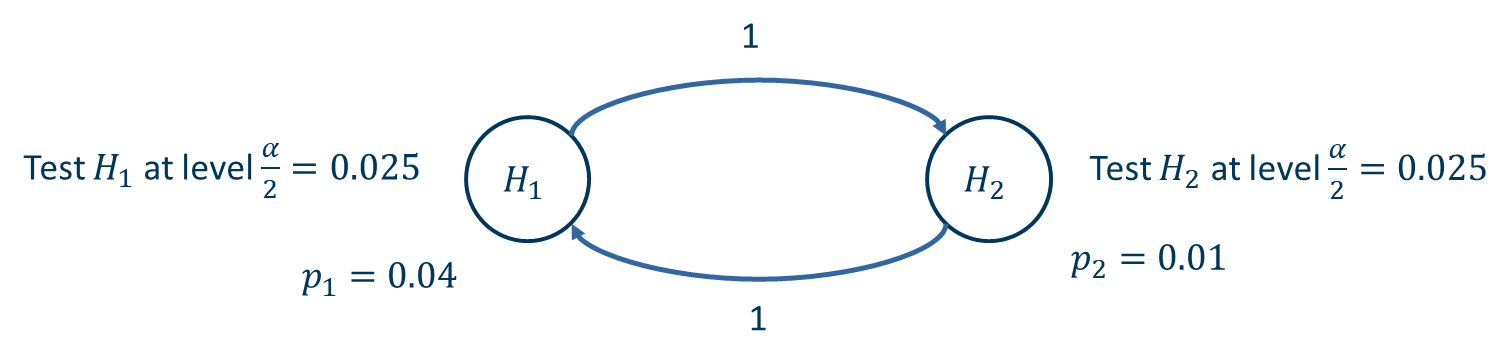


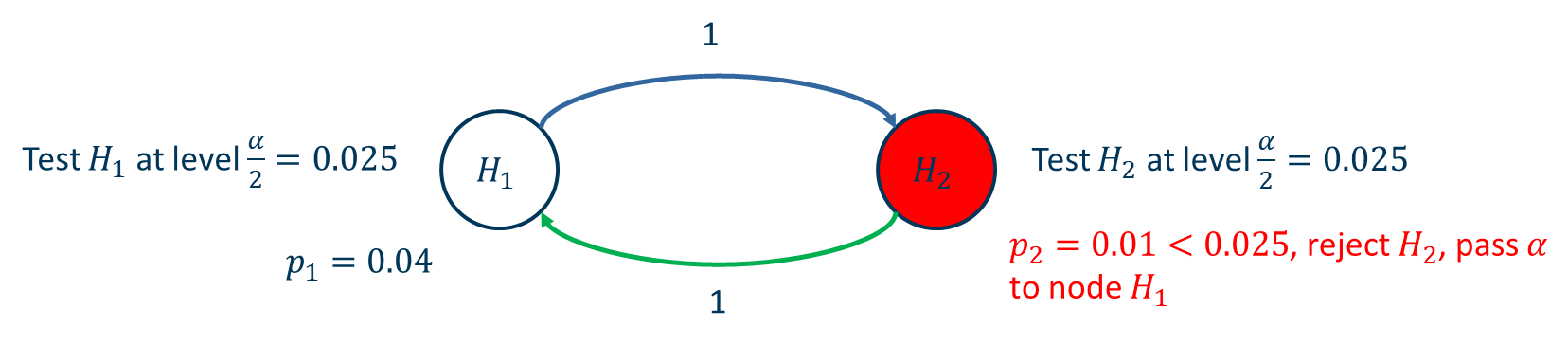
Figure 26 - Graphical illustration of the weighted Bonferroni–Holm procedure with two hypotheses.

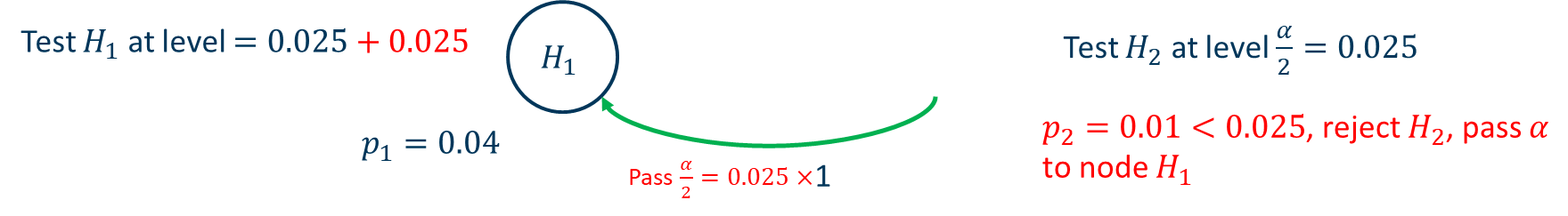
A vivid demonstration is provided below **(m=2, =0.05)**:

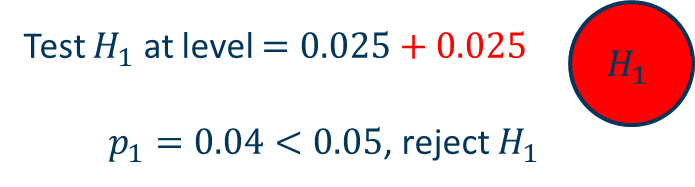




Since Bonferroni-Holm is a step-down procedure, it will start with the most significant p-value which is :







**Formal Definition**

* Let initial levels with
* transition matrix where is the fraction of the level of that is propagated to with , , and , ∀𝑖 = 1, … , 𝑚
* defines a directed graph with an associated multiple test

To illustrate the proposed graphs, consider the following Figure 26 for an example involving hypotheses. For the graph we have and transition matrix

Diagram

Description automatically generated

Figure 27 - Example multiple test procedures to illustrate

#### Algorithm to update the graph

Let , a graph

1. Set
2. Let .
3. If , reject ; otherwise stop.
4. Update the graph:

1. If , go to step 1; otherwise stop.

#### 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 , 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 precedes , which in turn precedes . Note the initial allocation of the overall signiﬁcance level to the individual hypotheses. If, for example, is rejected, the initially allocated signiﬁcance level (at the vertex ) is passed on fully to (as indicated by the directed edge with associated weight 1). Accordingly, and

Diagram

Description automatically generated

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

Chart

Description automatically generated

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.

#### Example 2 - Fallback procedure

*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 ). In the notation from Algorithm, and

**Simple demonstration**

1. Test at . Suppose , not reject , retain it;
2. Test at . Suppose , not reject , retain it;
3. Test at . Suppose , reject , pass to and to ;

*Fallback*

1. Test at ...
2. Test at ...

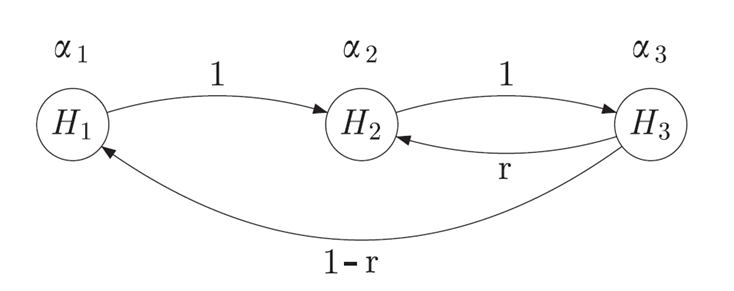


Figure 30 - Improvement of the fallback procedure by Wiens BL, Dmitrienko A. with .

#### Example 3 - Bonferroni–Holm procedure

As seen from the Figure 30, where and

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 and overall signiﬁcance level . Figure below displays the resulting sequentially rejective test procedure.

The hypothesis is rejected at the ﬁrst step, since . The associated local signiﬁcance level is split equally and passed on to the remaining (not yet rejected) hypotheses and , as indicated by the directed edges in the left graph in Figure 31.

Based on the algorithm proposed before, we have:

where since is rejected.

Similarly, will be updated to 1 and will be updated to 0.025, which is the middle graph.

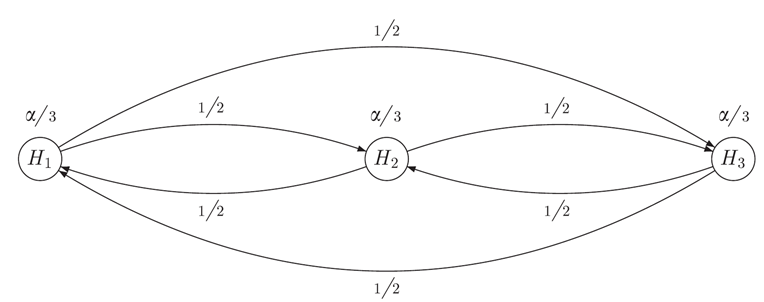


Figure 31 - Graphical illustration of the Bonferroni–Holm procedure with 𝑚=3 hypotheses and initial allocation

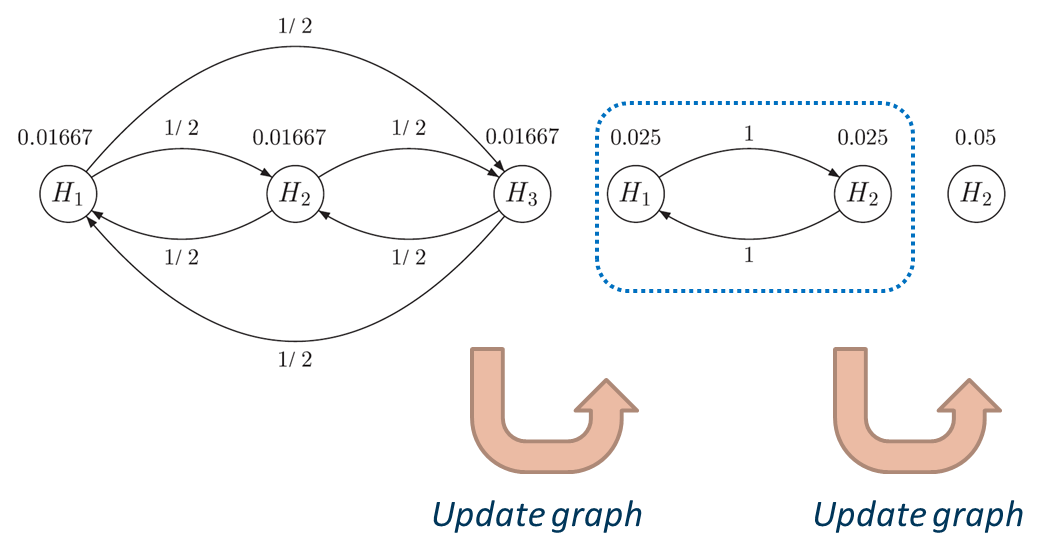


Figure 32 - A vivid demonstration

#### 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 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.

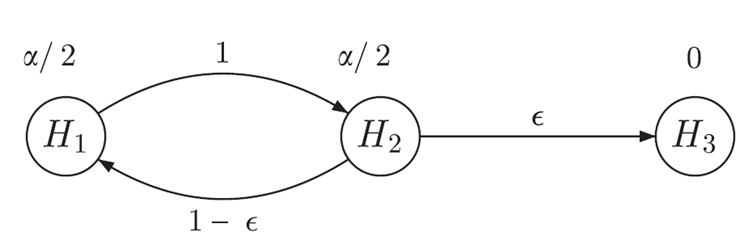


Figure 33 - The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values , , and

##### Example 1

As an example consider the test of three hypotheses , ,and , where , are of primary interest and is of interest only if and can be both rejected.

The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values , , and

If both and can be rejected, then is tested at level .

Additionally, to achieve weights that sum to one, the weight of the edge is set to (instead of 1 in the Bonferroni–Holm procedure).

**Simple demonstration**

As , is rejected in the ﬁrst step and its level is shufﬂed to hypothesis , since by the above calculation rules and .

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

In particular, the edge from to gets the weight .

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



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

##### Example 2

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

* , ,
* the observed p-values , , , .

If both hypotheses and are rejected, the significance level is shuffled to and according to the weights and :

* receives and receives .

If both and can be rejected, then is tested at level .

**Simple demonstration**

As , is rejected in the ﬁrst step and its level is shufﬂed to hypothesis , since by the above calculation rules and . The level for is updated to .

Node is dropped and the graph will be updated to

* Weight for edge to :
* Weight for edge to :

The edges of are updated as shown in the middle graph in Figure 34. Since is rejected at , Now will receive the level , will receive the level

Then the tests within ‘family’ consists of and will be perform as the right graph illustrating.

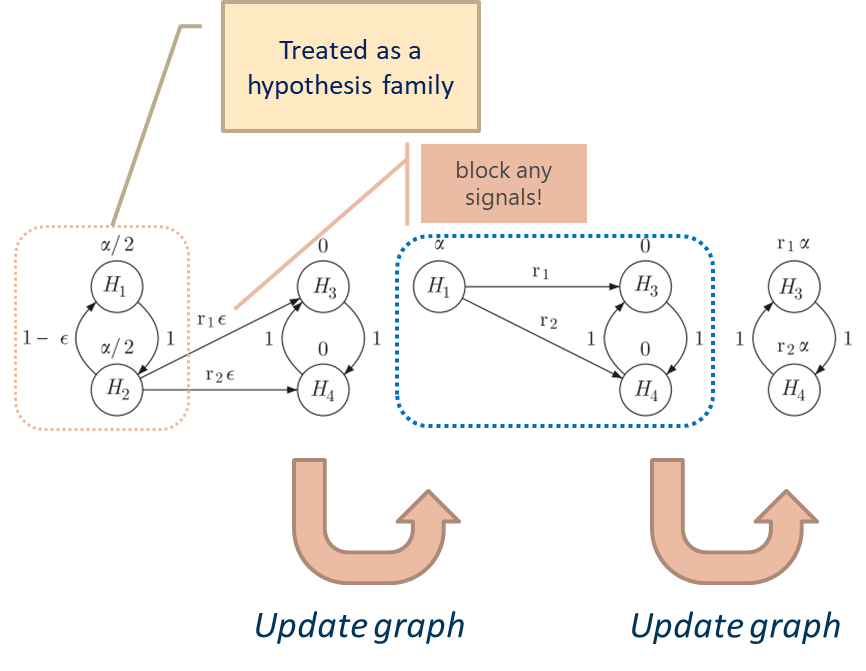


Figure 35 - A vivid presentation of updating a graph with ϵ-edge.

##### 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).

* the observed p-values , , , .

***Simple demonstration***

As , is rejected in the ﬁrst step and its level is shufﬂed to hypotheses and , which are both now assigned level .

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

*Thus, the gatekeeping procedure can be improved by adding -edges from to and from to (see the below figure).*

The only outgoing edge from after the rejection of and is the -edge to and is thus assigned the weight 1.

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

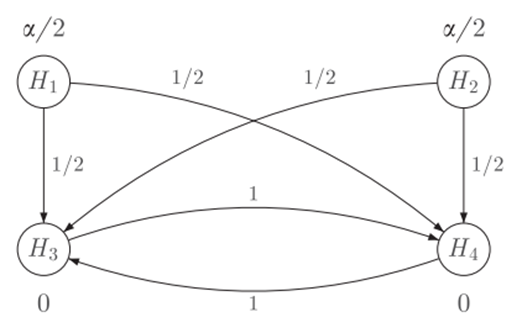


Figure 36 - Graphical illustration of the gatekeeping procedure with four hypotheses.

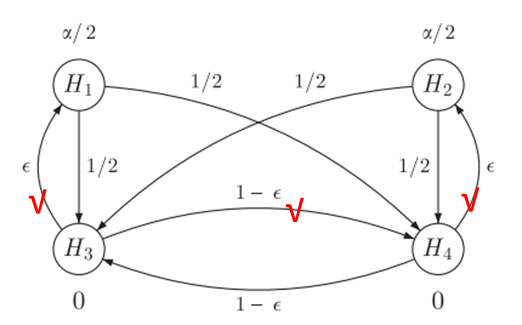


Figure 37 - Graphical illustration of the improved gatekeeping procedure with four hypotheses by adding 𝜀-edge. The tests start from hypothesis and the 𝜀-edge can be treated as a “blocker”.

#### 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:

where

denotes the mean difference of treatment with different dose level and control

(1-high dose; 2-low dose)

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).

##### Strategy 1

Consider a fixed sequence test to the six hypotheses being and to test each hypothesis sequentially at level .

The sequence H11→ H21→ H12→ H22→H13→ H23 is a reasonable possibility, see Figure 37.

A picture containing text, clock

Description automatically generated

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 , none of the subsequent hypotheses can formally be rejected, even if their p-values are very small.

##### 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

* = {H11, H12, H13} – high dose level
* = {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 . 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 .

Diagram, arrow

Description automatically generated

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.

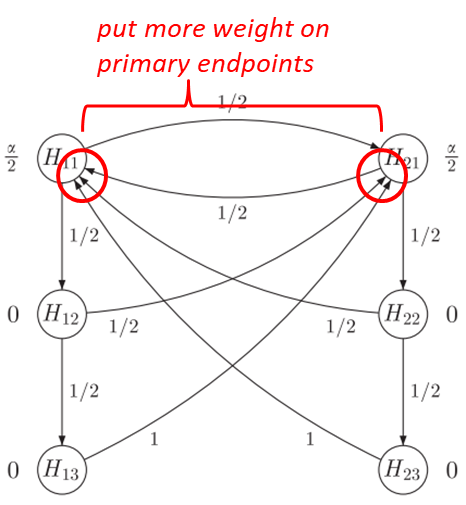


Figure 39 - Visualization of different implementations for Strategy 2.

##### 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 for all , 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 .

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.

### 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[[6]](#footnote-8) 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[[7]](#footnote-9) and non-hierarchical procedures into one framework. A graphical multiple testing procedure with “no-dead-end” provides an opportunity to fully recycle 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
  + Other extensions of gatekeeping procedures

**Integrate non-hierarchical and hierarchical hypotheses**

* Graphical approaches

### 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).

#### 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.

Table

Description automatically generated

#### 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.

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.

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

**Assumptions**

* a parallel group 2-arm trial with planned interim analyses, including the final analysis;
* the treatment effect in comparison to a control denoted by the single parameter of interest, , that can take on any real value;
* there is only one null hypothesis versus alternative hypothesis ;
* the amount of statistical information available at interim analysis is ;
* 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 ;
* in the most common case, ’s follow (asymptotically) a multivariate normal distribution (MVN) with and .

**Concepts**

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

A pair of hypotheses ( and ) 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.

##### Stopping[[8]](#footnote-10) boundaries

**At the first interim analysis ()**

the test statistic is compared to lower and upper “stopping boundaries” and , respectively. If , the trial stops and the null hypothesis is retained (or equivalently “futility” is declared). If , is rejected in favor of and the trial is also stopped. If the trial continues to the next planned interim or final analysis.

**At the interim analysis ()**

If , the trial stops and the null hypothesis is retained (or equivalently “futility” is declared). If , is rejected in favor of and the trial is also stopped. If the trial continues to the next planned interim or final analysis.

**At the final analysis ()**

If the trial is not stopped at any of the interim analyses, a final test is done with being compared to the decision thresholds . In this case 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 at any of the interim or at the final analysis, is guaranteed not to exceed a predefined level , e.g., one-sided .

The probability to reject the null hypothesis at interim analysis , given a true treatment is

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

When , is called

* the level spent at interim analysis .

Under the alternative hypothesis with , is called

* Type II error rate spent at interim decision .

**Overall Type I error rate**  is given by

**Overall Type II error rate**  is given by

It should be noted that, if at the final analysis a decision is taken with respect to rejection or retention of , i.e., if , then . 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**).

#### 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 𝛼 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 𝛼 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].

##### 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;
* times of interim analyses and is for the specific analysis;
* **Wald-statistics** for A and for B based on cumulative data;
* Null hypothesis () of no treatment effect on the endpoint A (B);
* allocation: and , where ;
* Group sequential boundaries:
  1. for A at significance level , and for A at significance level . Based on consonant property, there is .
  2. for B at significance level , and for B at significance level ; Based on consonant property, there is .

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

Text, letter

Description automatically generated

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

**Figure 41 - Closed test.**

Diagram

Description automatically generated

**Boundary values for endpoints**

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.

1. After one hypothesis is rejected, one may continue using the predefined interim boundaries with or for . In other words, the boundaries can be left unchanged at the interim analyses except at the final analysis .
2. Alternatively, or may be updated with completely different values after the 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 level because it is a closed test. To see this, consider the closed family

An level for global test for is as follows:

* Reject if endpoint A cross its level group sequential boundary () at any interim look , or if endpoint B cross its level group sequential boundary () at any interim look :

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 is split between the two endpoints each with independent group sequential procedures and no reallocation. For ease of reference, we label the naïve approach as **group sequential Bonferroni (GSB)**.

**Low-dimensional**

Scenarios are:

* one interim analysis and one final analysis are planned ;
* the nondecreasing functions defined over are and ; Such that ;
* the information fraction for endpoint A at the interim analysis is ;
* the information fraction for endpoint B at the interim analysis is ;
* the information fraction for endpoint A at the final analysis is , where ;
* the information fraction for endpoint B at the final analysis is , where ;
* boundary values and they are calculated by using Lan-DeMets error spending function.

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

* An error spending function that approximates the O’Brien–Fleming boundary is given by

Where is the standard normal cdf and is the quantile of the standard normal distribution. Note that is two-sided level.

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

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

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

**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 in the patients with non-squamous histology and 80% power (13 months v 10 months) for OS with two-sided 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 and are the log-rank statistics at interim and ﬁnal analysis for the overall population):

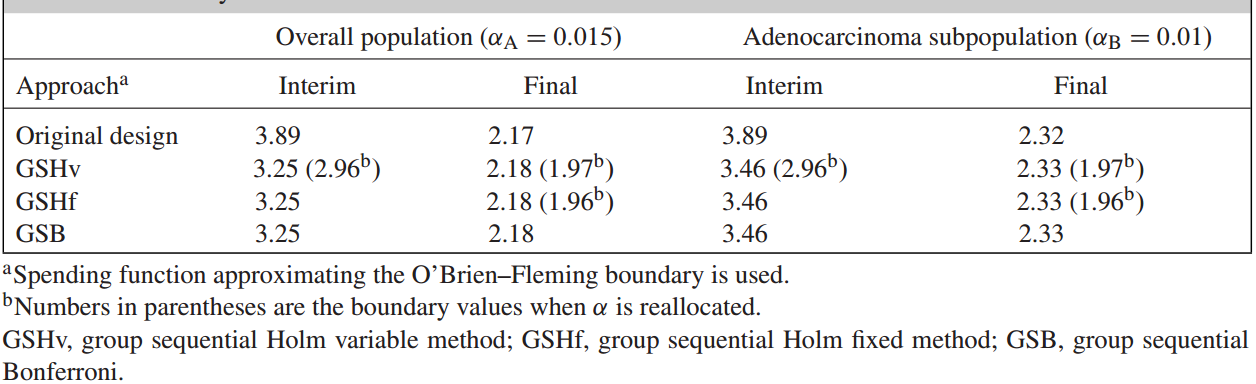
It can be shown that for the overall population. Similarly, in the adenocarcinoma subpopulation. Rejection region at the interim analysis for this method GSHv is shown in Figure 41 where is split between the overall population and the adenocarcinoma subpopulation (0.015 and 0.01, one-sided respectively).

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

It can be shown that for the overall population.

Similarly, in the adenocarcinoma subpopulation. Table 12 summarizes the boundary values of the various methods.

Table 12 - Boundary values for the MONET1 trial.



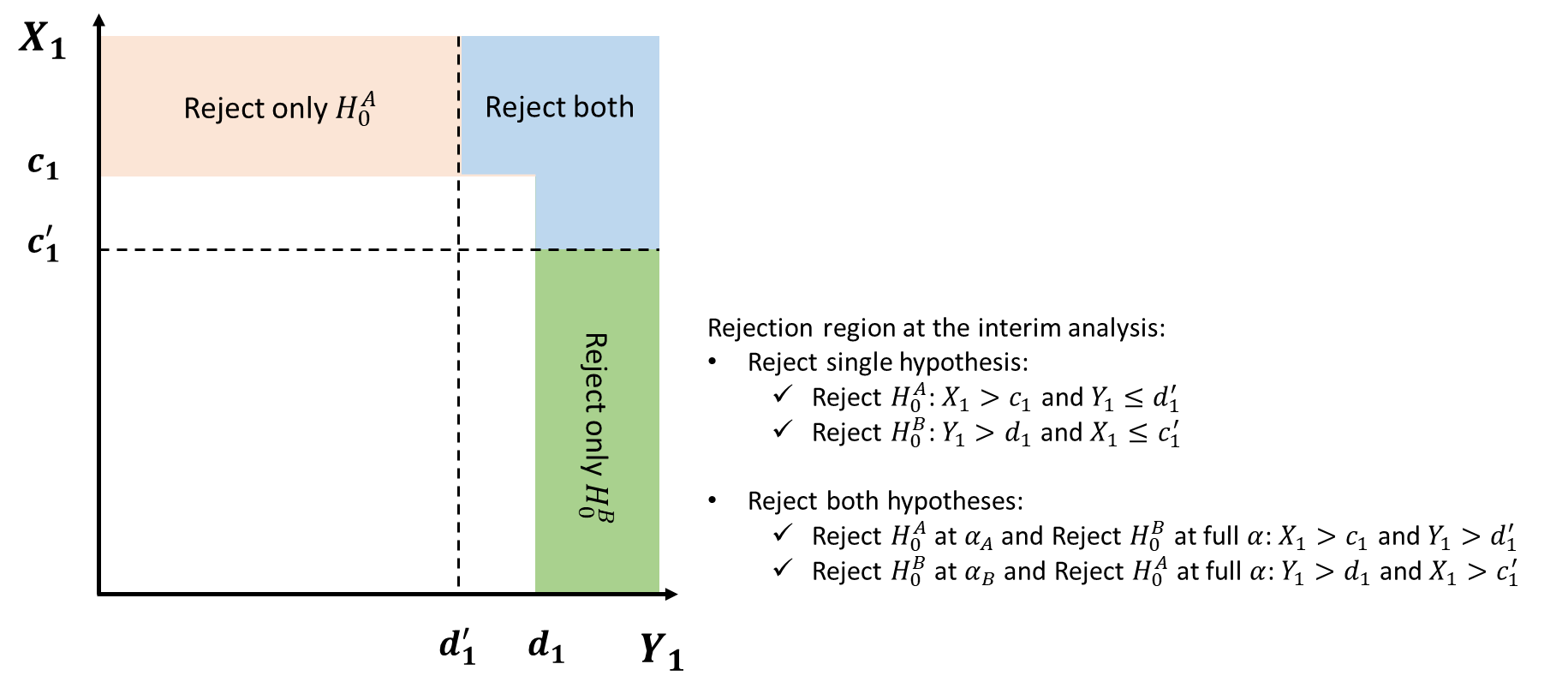


Figure 42 - Rejection region at the interim analysis for MONET1 (GSHv).

##### Extension of methodology

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

**Mathematical notations**

: the number of endpoints of interest, which are also the number of associated hypotheses to be tested ;

: the total number of interim analyses;

: the index set of all endpoints;

: weight for endpoint where ;

: type I error rate allocated for , at each interim analysis, where ;

: Wald statistics to test at each interim look;

: group sequential boundary for at significance level at each interim look such that

**Algorithm**

1. Test single hypothesis
   1. None of the endpoints crossed its boundary at any of the looks, retain all and stop;
   2. Any one of the endpoints crossed its boundary at any of the looks, then efficacy with respect to this endpoint can be claimed.

**Details**

Test with at significance level :

: earliest interim look where at least one endpoint can be rejected;

set of the endpoints that crossed their boundaries at interim look ;

: which the set of remaining endpoints.

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

***updates***

For , the significance level assigned to all the individual hypotheses will be updated to

A screenshot of a computer

Description automatically generated with low confidence

Where and .

***Boundaries updates***

The boundaries for at significance level will be updated using the error spending approach that satisfies the following equation:

A picture containing text

Description automatically generated

1. Repeatedly to test single hypothesis for any interim look () using updated boundary values .
   1. None of the endpoints crossed its boundary at any of the look, retain all and stop;
   2. Any one of the endpoints crossed its boundary at some interim look (), then efficacy with respect to this endpoint can be claimed:

**Details**

: set for cross boundaries at interim look ;

: which the set of remaining endpoints.

***updates***

A picture containing graphical user interface

Description automatically generated

***Boundaries updates***

Text

Description automatically generated with low confidence

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 , the boundaries should satisfy

where and .

For GSHf is fixed at for , otherwise for GSHv.

**A simple illustration**

Let , (3 interim analyses (IA) + 1 final analysis).

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | Variables |
| IA1 | Test with (calculated based on significance level ) 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 | are updated to ;  Calculate based on |
| IA2 |  | Test with () 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 | is updated to ;  Calculate based on |
| IA3 |  |  | the endpoint 3 is not rejected | None of the endpoint crossed its boundary at 3rd look, retain all and stop testing; |
| Final |  |  | Test with () 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. |

##### 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 . 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 representing the local signiﬁcance levels . The transition weight associated with a directed edge between any two vertices and 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 at the head is rejected.

The sum of the transition weights with tail on node is restricted by for and there are no elementary loops (edges where head and tail coincide); that is, for .

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.

Text, letter

Description automatically generated

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 hypotheses and weights , , , it is equivalent to the group sequential graphical procedure by setting , , and , .

**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 hypotheses, grouped into

* Two primary hypotheses , comparing low and high doses, respectively, against placebo for HbA1c;
* Two secondary hypotheses , (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:

* and

A picture containing text, necklet, accessory

Description automatically generated

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 are represented by nodes with associated weights representing the local signiﬁcance levels , .

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 .

Assume that both doses are considered to be equally important and thus let . Further-more, assume that if one of two primary hypotheses can be rejected at local signiﬁcance level , 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 , 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 .

##### 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[[9]](#footnote-11) BID[[10]](#footnote-12)).
* Capecitabine will be given at 1000 mg/m2 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 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.

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.

Diagram, table

Description automatically generated

**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 (PFSBM) 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:

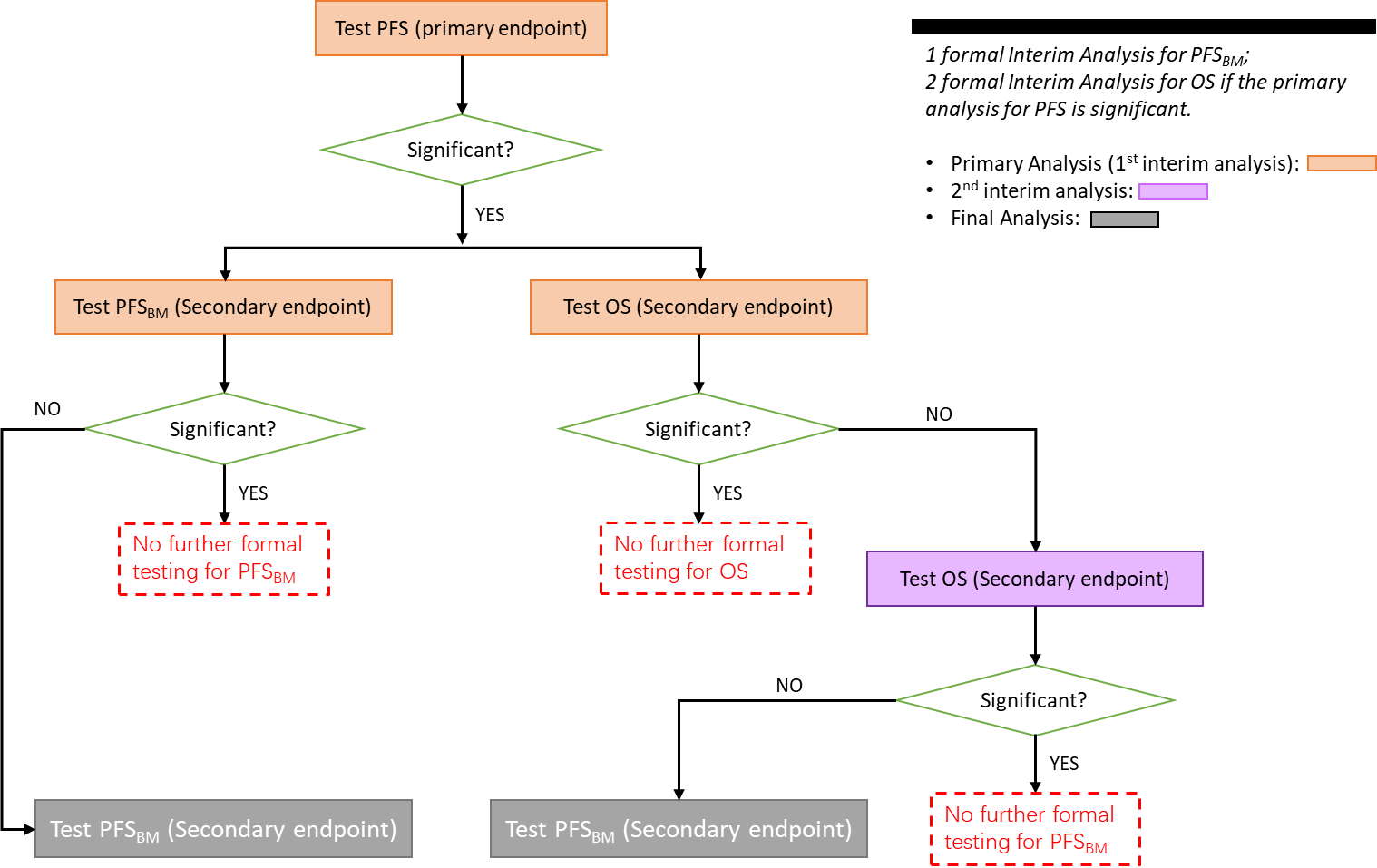


Figure 45 - A flowchart of multiple testing strategy.

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.  An interim analysis for the key secondary endpoints PFSBM and OS in the ITT population will also be performed at this time if PFS is statistically significant. |
| 2ND Analysis | * If PFSBM is NOT statistically significant at the time of the primary analysis of PFS, the final analysis of PFSBM will be performed when approximately 220 PFS 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 PFSBM.   ***or***   * If PFSBM is statistically significant at the time of the primary analysis of PFS, no further formal testing of PFSBM 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.

Text, letter

Description automatically generated

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).

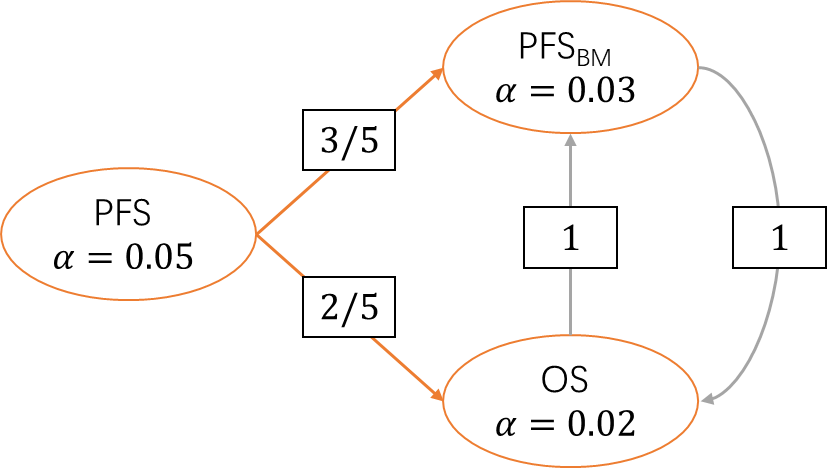


Figure 47 - Type I Error Reallocation Strategy Following Closed Testing Principle.

If PFS is statistically significant, then the key secondary endpoints PFSBM and OS will be tested using group sequential boundaries:

* The split between PFSBM and OS is and , 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 is the ratio between number of events at interim analysis and number of events at final analysis. For illustration purpose, we assume for PFSBM and , 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 | |
| **PFSBM** | **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 PFSBM and OS. The boundary at interim analysis is determined according to the Lan-DeMets O’Brien-Fleming (LD(OF)) approximation spending function

for two-sided tests, where is the upper critical point (quantile) of the standard normal distribution.

The GSHv procedure operates as follows:

1. Begin with a 0.03-level group sequential boundary for PFSBM 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 formal statistical testing for PFSBM and OS will be conducted.

Table 16 - **Initial** LD(OF) nominal P-value boundaries for PFSBM (2 analyses) and OS (3 analyses).

|  |  |  |
| --- | --- | --- |
| Analysis | PFSBM () | OS () |
| 1 | 0.0139 | 0.0023 |
| 2 | 0.0259 | 0.0069 |
| 3 |  | 0.0176 |

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

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

**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 PFSBM (2 lo analyses) and OS (3 analyses) at level.

|  |  |  |
| --- | --- | --- |
| Analysis | PFSBM () | OS () |
| 1 | 0.0258 | 0.0092 |
| 2 | 0.0425 | 0.0194 |
| 3 |  | 0.0429 |

1. 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 PFSBM was found significant at final analysis at level, but OS was not significant at level, then OS can be **tested again** at the level.
2. If PFSBM 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 PFSBM 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 in Table 18 and Table 19 are for illustration purpose only.

If both PFSBM 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 level.

Table 18 - Initial LD (OF) boundaries for PFSBM (2 analyses) and OS (2 analyses).

|  |  |  |
| --- | --- | --- |
| Analysis | PFSBM () | OS () |
| 1 | 0.0139 | 0.0023 |
| 2 | 0.0259 | 0.0193 |

Table 19 - LD (OF) boundaries for PFSBM (2 analyses) and OS (2 analyses) at level.

|  |  |  |
| --- | --- | --- |
| Analysis | PFSBM () | OS () |
| 1 | 0.0258 | 0.0092 |
| 2 | 0.0425 | 0.0471 |

#### 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.

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.

##### 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. to test that the OS in experimental treatment is superior to the control in the squamous cell carcinoma subgroup (Population 1);
2. to test the superiority in the subgroup with PD-L1 CPS ≥10 (Population 2);
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.

Diagram

Description automatically generated

Figure 48 - The 3 populations of Scenario 1.

In current practice, the closed weighted Bonferroni approach is often used to split 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.

##### Methodology

In this section, the notations for graphical approaches are exactly the same as those in Section 3.3.2.

###### Correlated hypothesis

Among the tests of the 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 planned analyses if the trial continues to the end for all hypotheses. We assume further that the distribution of the tests of individual hypotheses at all 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 be the standardized normal test statistic for hypothesis , analysis .

Let be the the number of observations (or number of events for time-to-event endpoints) collected cumulatively through stage for hypothesis .

We use the wedge operator in to denote he number of observations (or events) included in both and .

The key of the parametric tests is to utilize the correlation among the test statistics. The correlation between and is

The full correlation matrix of and () can be derived this way and is referred to as the complete correlation structure (CCS).

###### Well-ordered family of group sequential bounds

A restriction of spending functions to those that produce well-ordered bounds as -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 . For simplicity of notation, we just assume that bounds will be defined for at all significance levels of satisfying ,

For a well-ordered family Maurer and Bretz (Willi Maurer & Frank Bretz, 2013), we require for any

that

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 is strictly increasing in and as , we have .

# 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.

# Appendix

## Error Spending Methods

From https://documentation.sas.com/doc/en/pgmsascdc/9.4\_3.3/statug/statug\_seqdesign\_details37.htm

Text, letter

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

## 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 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 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.

For example, , if all intersection hypotheses containing are rejected. If , , denotes the p-value for testing the intersection hypothesis , the adjusted p-value for is the largest p-value associated with the index sets including :

The hypothesis is rejected if the adjusted p-value does not exceed the pre-specified level, i.e., . 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).

## Proof of 2-stage gatekeeping procedure controls the FWER at the α level



Text

Description automatically generated

## A detailed example of sequentially rejective Bonferroni-based closed testing procedures

For Figure 9, Figure 10 and Figure 11, the details regarding the use of closure principle are

Graphical user interface, application

Description automatically generated

The parts in red are assumptions.

1. We assume that the unadjusted p-value for hypothesis satisfies and .
2. We assume that the unadjusted p-value for hypothesis satisfies and .
3. Let and the vector of weights for is .

Since is tested only when at least one primary hypothesis is rejected, the weights for the three hypotheses are set to

Where .

We test by comparing . Based on the **assumption A**, can be rejected and the corresponding elementary with the minimum adjusted p-value can also be rejected at level . (See Section 2.3.2.1)

1. Since is rejected in **Step 1**, the index set is updated to . Based on the Section 2.3.2.2, we have

And the procedure should meet the condition

The vector of weights for is should satisfy

Therefore we can set the updated

It is obvious that and

We test by comparing .

Based on the **assumption B**, can be rejected and the corresponding elementary with the minimum adjusted p-value can also be rejected at level . (See Section 2.3.2.1)

1. Since both and are rejected, then one has to simply test at level .

The above complicated steps can be simplified to the following graph:

Diagram

Description automatically generatedText

Description automatically generated

## 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; 有次期中分析（IA）及一次最终分析（FA），总共次分析。

根据闭合原理（CTP），想要在水平上拒绝假设，需要在显著水平上拒绝以及。

对于，在成组序贯设计中，如果在所有次分析中，终点A对应的检验统计量在任意一次分析 （）中超过了其对应的拒绝域边界值（这个边界值是基于显著水平)计算的），则可以拒绝。同样的，以上叙述对终点B适用。

基于水平对做假设检验的数学表达式为

换句话说我们可以在水平上检验，只要出现至少一个或者事件（）,即可拒绝，此时基于CTP，可以进行-reallocation。

由于或者事件对应的是显著水平或，所以实际操作中，我们是在显著水平或上拒绝的。更具体地说，假设在所有次分析中，在第次分析时，我们观测到,则我们可以在显著水平上拒绝;又因为，显然我们还可以在显著水平上拒绝。基于CTP，仅需要在显著水平上检验。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，我们并不是直接在水平上去检验，而是在（spending function, 为timing of analysis）上去检验。同理，在某次IA（比如）中拒绝了与后，我们要在显著水平上去检验终点B。

由于error spending function在此处为显著水平的单调非减函数，我们有当，来保证CTP中consonance的性质，error spending function的数理性质也应保证对应的决策边界与对应的决策边界满足。

## Truncated multiple test procedures

Note that **MTP** refers to **m**ultiple **t**est **p**rocedures.

Consider a single family of hypotheses, , with p-values, . 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 level.

### Truncated Holm

Text

Description automatically generated

Text

Description automatically generated

### Truncated Hochberg

Text

Description automatically generated

Text, letter

Description automatically generated

### Truncated Fallback

Text, letter

Description automatically generated

Text

Description automatically generated

### Truncated Dunnett

Text

Description automatically generated

## Weight assignment algorithm for Bonferroni tree gatekeeping procedures

The algorithm is given by Dmitrienko et al. (Alex Dmitrienko; Ajit C. Tamhane; Frank Bretz;, 2010).

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Text, letter

Description automatically generated

1. With the weighted Bonferroni test, we reject if for at least one , where denotes the unadjusted p-value for . [↑](#footnote-ref-2)
2. A closed testing procedure is termed **consonant** if the rejection of an intersection hypothesis with and always leads to the rejection of at least one implied by , i.e., with . [↑](#footnote-ref-3)
3. Bonferroni test: Reject if [↑](#footnote-ref-4)
4. 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. [↑](#footnote-ref-5)
5. Assume no treatment effect between treatment groups. [↑](#footnote-ref-7)
6. 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. [↑](#footnote-ref-8)
7. 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. [↑](#footnote-ref-9)
8. 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). [↑](#footnote-ref-10)
9. Oral administration [↑](#footnote-ref-11)
10. Twice daily [↑](#footnote-ref-12)