Some Statistical Methods for Multiplicity Issues in Clinical Trials

Project Work

Submitted to

PONDICHERRY UNIVERSITY

in partial fulfilment of the requirements

for the award of the degree of

Master of Science

in

Statistics

BY

KARRA SAIKIRAN

Reg No: 22375032

DEPARTMENT OF STATISTICS

PONDICHERRY UNIVERSITY

PONDICHERRY

APRIL, 2024

PONDICHERRY UNIVERSITY
R.V. NAGAR, KALAPET,
PUDUCHERRY-605014

Certified that the project work entitled **Some Statistical Methods for Multiplicity Issues in Clinical Trials** is a bonafide record of work carried out by the following student.

KARRA SAIKIRAN (22375032)

of M.Sc. (Statistics) submitted in partial fulfilment of the requirement for the award of degree of Master of Science in Statistics, during the academic year **2022-24**.

(Supervisor) Head of the Department

Submitted for M.Sc., Degree Examination held on ______________

Examiners
1.
2.

GSK India Global Services Private Limited
Level 1,2 & 3, Luxor North Tower
Bagmane Capital Business Park
Outer Ring Road, Mahadevupura
KR Puram Hobli
Bangalore - 560037

TO WHOMSOVER IT MAY CONCERN

Date: 02-May-2024

This is to Certify that Saikiran Karra is pursuing Internship with Development Biostatistics, GCC GSK India from 15th January 2024, under the guidance of Ramiya Ravindranath & will be completing Internship on 15th July 2024.

We wish him every success in life and career.

Yours sincerely,

For GSK India Global Services Private Limited

Nilesh Kumar
Director - Human Resources - Global Capability Centre

CERTIFICATE

This is to certify that **Karra Saikiran** (**22375032**), a student of M.Sc. Statistics from the Department of Statistics, Pondicherry University has completed his fourth semester project under our guidance at **GlaxoSmithKline (GSK), Bengaluru from January to May 2024**. The project work entitled **“Some Statistical Methods for Multiplicity Issues in Clinical Trials”** embodies the novel work done by him.

Signature__________

Dr. J. Prabhakara Naik
Associate Professor
Department of Statistics
Pondicherry University
Puducherry – 605014
(Faculty Guide)

Dr. Anjali Pandey
Manager, Statistics
Biostats, India
GSK, Bengaluru
Bengaluru - 560001
(Guide)

ACKNOWLEDGEMENT

I take this opportunity to acknowledge my deep sense of gratitude to **Dr. J. Prabhakara Naik**, Associate Professor and Coordinator for internships and placements, Department of Statistics, Pondicherry University for providing this opportunity to work with the reputed company **GlaxoSmithKline (GSK), Bengaluru.**

I am thankful to my guide, **Dr. J. Prabhakara Naik,** Associate Professor, Department of Statistics, Pondicherry University, Puducherry, for his guidance and overall support throughout the project without which it would have been impossible to complete this work.

I express my sincere thanks to **Dr. Anjali Pandey**, Manager Statistics, and **Dr. Ramiya Ravindranath,** Associate Director Statistics, GSK, Bengaluru for providing invaluable guidance throughout the project. Her sincerity, dynamism, motivation has inspired me. It was a great privilege to work under her guidance.

I express my sincere gratitude to **Dr. Navin Chandra** of Department of Statistics, Pondicherry University, Puducherry for his valuable support and **Dr. Angshuman Sarkar,** Head of the R&D Division, GSK, Bengaluru for this wonderful opportunity to work with them.

I am thankful to the Dean **Dr. Rajeswari Seshadri,** Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry for her generous support.

I am very thankful to all the faculty members of the Department, **Dr. R. Vishnu Vardhan, Dr. V.S. Vaidyanathan, Dr. Kiruthika, Dr. P. Tirupathi Rao**, **Dr. Sudesh Pundir** for their generous and moral support.

I wish to place on record the valuable support of my family members without whose support it would have been impossible to complete this project.

I am thankful to my friends for their cooperation and support in bringing out this project as a thriving one.

(KARRA SAIKIRAN)

CONTENTS

Page

1. **Abstract** 2
2. **Introduction** 2
   - 2.1 Type I Error 3
   - 2.2 Family Wise Error Rate (FWER) 3
   - 2.3 Additional Elements 6
3. **Methodology** 6
   - 3.1 Overview of the Multiple Testing Procedures (MTP's) 7
   - 3.2 Single-step Procedures 8
     - 3.2.1 Bonferroni Procedure (Non-Parametric) 8
     - 3.2.2 Sidak Procedure (Semi-Parametric) 9
     - 3.2.3 Simes Procedure (Semi-Parametric) 10
     - 3.2.4 Dunnett Procedure (Parametric) 11
     - 3.2.5 Summary 13
   - 3.3 Hierarchical Hypothesis Ordering 14
     - 3.3.1 Simple or Pre-specified Hypothesis Ordering 14
     - 3.3.2 Gatekeeping Procedures 20
     - 3.3.3 Summary 25
4. **Power Comparison of Hierarchical Methods of Multiplicity Adjustment using Simulations** 26
   - 4.1 Simulation Study 27
   - 4.2 Methodology of the Study 28
   - 4.3 Power Comparison of the Study with 10,000 Simulations 34
     - 4.3.1 Simulation Results 35
     - 4.3.2 Interpretation of the Results 35
     - 4.3.3 Overall Interpretation 38
5. **Conclusion** 38
6. **References** 39

Some Statistical Methods for Multiplicity Issues in Clinical Trials

1. Abstract

Multiplicity issues has been identified as one of the causes behind failure of confirmatory phase III trials. Both European Medical Agency (EMA) and United States (US) FDA issued draft guidance on confirmatory inferences with multiple endpoints in clinical trials to ensure strong control of the family wise error rate (FWER). Multiplicity issue arise due to multiple endpoints or multiple comparison problems like multiple dose-control comparisons, trials with multiple subgroups and, trials with interim analysis. This may lead to inflated FWER and consequently to a false positive conclusion for an ineffective treatment. There exist many different methods to account for multiplicity, for example step-down (or hierarchical / fixed sequence) testing methods, gatekeeping methods and alpha-adjustment methods. One indispensable issue requiring contemplation is that different multiplicity adjustment methods might lead to different conclusions for the same data. Hence it becomes crucial to identify the most appropriate method to account for multiplicity at the design stage and pre-specify it in protocol. In this work, we propose to evaluate the possible sources (single and multiple sources) of multiplicity in a clinical trial and respective adjustment methods through a comprehensive literature review. Further, choices of multiplicity adjustment method will be evaluated for simulated study designs.

2. Introduction

Clinical trials are essential for evaluating the safety and efficacy of new medical interventions, ranging from drugs and vaccines to medical devices and procedures. However, designing and analyzing these trials can be complex due to various factors, one of which is multiplicity.

Multiplicity in clinical trials refers to testing many comparisons at once, like different treatments or symptoms. While not inherently problematic, it can lead to misleading results if not properly managed. Special statistical methods are used to address multiplicity, ensuring reliable conclusions are drawn from the study.

Multiplicity Issues can be faced more often in confirmatory phase III trials since it is the comparison phase, trial sponsors are often interested in investigating multiple clinical objectives based on the evaluation of several endpoints, multiple dose-control comparisons, assessment of the treatment effect in two or more patient populations, etc. [1]. Regulatory agencies like the FDA and EMA recognize the importance of addressing multiplicity to ensure the reliability and validity of trial results, often requiring specific methods for multiplicity adjustment in trial designs and analyses.

Multiplicity issues can lead to inflated type I error rates, which occur when a statistically significant result is declared even though there is no true effect or difference (discussed in detail below). This inflation occurs because the more comparisons that are made, the greater the likelihood of observing a significant result by chance alone. Failing to account for multiplicity can lead to false-positive findings, where interventions appear effective when they are not.

2.1 Type I Error:

The rejection of the null hypothesis supports the study conclusion that there is a difference between treatment groups but does not constitute absolute proof that the null hypothesis is false. There is always some possibility of mistakenly rejecting the null hypothesis when it is in fact true. Such an erroneous conclusion is called a Type I error. For an endpoint, the probability of falsely rejecting its null hypothesis and, thus concluding that there is a treatment effect due to the drug on this endpoint when, in fact, there is none, is called the Type I error probability or Type I error rate for this endpoint. The significance level is denoted as alpha (α), is the threshold below which the Type I error rate should be controlled.[3]

2.2 Family Wise Error Rate (FWER):

The familywise error rate (FWER) is a measure used to control the probability of making one or more false discoveries (Type I errors) in a family of statistical tests. In the context of clinical trials, the 'family' typically refers to the collection of all hypothesis tests conducted within the study, including comparisons of multiple endpoints, treatment arms, or subgroup analyses. Controlling the FWER ensures that the overall probability of falsely rejecting a true null hypothesis across all tests in the family remains below a predetermined threshold, typically denoted by α (e.g., $\alpha = 0.05$). Researchers use various methods to control the familywise error rate to maintain the overall integrity and reliability of the study. Some common procedures include Single-Step, Stepwise or Data-Driven, Simple or Pre-Specified and Gatekeeping Procedures. These methods adjust the significance level for individual tests to reduce the inflation of Type I error that can occur when multiple tests are performed simultaneously. FWER is calculated as,

$$\mathbf{1 - (1 - \boldsymbol{\alpha})^{\boldsymbol{k}}}$$

where $'\boldsymbol{\alpha}'$ is the significance level and **'k'** is the number of tests performing under the study.

Example: In a clinical trial with a single endpoint tested at two-sided α = 0.05, the probability of finding a difference between the treatment group and a control group in favor of the treatment group when no difference exists in the population is 0.025 (a 2.5% chance). That is, there is a 97.5% chance of appropriately not finding a favorable effect if there is no true effect for this endpoint. By contrast, if there are two independent endpoints, each tested at two-sided α = 0.05, and if success on either endpoint by itself would lead to a conclusion of a drug effect, the chance of appropriately not finding a favorable effect on both endpoints together is thus 0.975 * 0.975, which is approximately 0.95, and so the probability of falsely finding a favorable effect on at least one endpoint is approximately 0.05. Thus, the overall Type I error rate in favor of the drug nearly doubles when two independent endpoints are tested. This higher-than-intended overall Type I error rate when multiple tests are conducted without adjustment is called the multiplicity problem. Thus, without correction for multiplicity, the chance of making a Type I error for this example study would rise to as high as 5% in favor of the drug, and, therefore, the overall Type I error rate would not be adequately controlled. The problem is exacerbated when more than two endpoints are considered. For example, for three independent endpoints, the Type I error rate is 1 - (0.975 * 0.975 * 0.975), which is about 7%. For ten independent endpoints, the Type I error rate is about 22%.[3]

There are several sources of multiplicity in clinical trials:

1. **Multiple Treatment Comparisons:** Multiple testing is commonly encountered in clinical trials involving several treatment groups. Examples of such situations include Phase II trials that are designed to assess the efficacy and safety profile of several doses of an experimental treatment compared to placebo or to an active control. Performing multiple comparisons of different dose levels of an experimental treatment to a control causes multiplicity problems.[1]

2. **Multiple primary endpoints**: Multiplicity is also caused by multiple criteria for assessing the efficacy or safety of an experimental treatment. The multiple criteria are required to accurately characterize various aspects of the expected therapeutic benefits. For example, the efficacy profile of cardiovascular drugs is typically evaluated using multiple outcome variables such as all-cause mortality, nonfatal myocardial infarction, or refractory angina/urgent revascularization [1]. Multiple primary endpoints occur in three ways,

    - When there are multiple primary endpoints, and each endpoint could be sufficient on its own to establish the drug's efficacy. These multiple endpoints thus correspond to multiple chances of success, and in this case, failure to adjust for multiplicity can lead to Type I error rate inflation and a false conclusion that the drug is effective [3].
    - When the determination of effectiveness depends on success on all primary endpoints, when there are two or more primary endpoints. In this setting, there are no multiplicity issues related to primary endpoints, as there is only one path that leads to a successful outcome for the trial and therefore, no concern with Type I error rate inflation [3].
    - Critical aspects of effectiveness can be combined into a single primary composite or other multicomponent endpoint, thereby avoiding multiple-endpoint related multiplicity issues. For example, in many cardiovascular studies it is usual to combine several endpoints (e.g., cardiovascular death, heart attack, and stroke) into a single composite endpoint that is primary and to consider death a secondary endpoint [3].

3. **Multiple secondary endpoints**: When an effect on the primary endpoint is shown, the secondary endpoints can be formally tested. A secondary endpoint could be a clinical effect related to the primary endpoint that extends the understanding of that effect or provide evidence of a clinical benefit distinct from the effect shown by the primary endpoint [3]. In modern clinical trials, it is more desirable to also make formal claims based on key secondary endpoints. Proper multiplicity adjustment is then also required for the secondary statistical tests to ensure control of the overall Type I error rate. For example, in oncology clinical trials, it is often desirable to characterize the clinical benefit using both progression-free survival and overall survival [1].

4. **Multiple patient populations**: The primary analysis in a clinical trial can be performed in the general population of patients as well as a prespecified target subgroup. Patients in this subgroup can be, for example, expected to experience greater treatment benefit or more likely to respond compared with the patients in the general population. Multi-population designs of this kind arise in oncology clinical trials and other therapeutic areas [1].

2.3 Additional Elements:

Scope of the Report: In this report, we will delve into various statistical methods used to address multiplicity issues in clinical trials, exploring their application, strengths and limitations.

Objectives of the Report: Our objective is to evaluate existing multiplicity adjustment methods by simulating the real-world examples of multiplicity issues in clinical trials to provide insights into best practices for addressing this challenge.

Overview of the Report Structure: The report is structured into several sections, beginning with an abstract and Introduction of multiplicity issues in clinical trials. This will be followed by a review of existing statistical methods for multiplicity adjustment, and an evaluation of their effectiveness through case studies using simulations. We will conclude with a discussion of the implications of our findings and recommendations for future research.

After discussing the significance and challenges of multiplicity issues in clinical trials, the methodology section will explore a range of statistical approaches and techniques employed to effectively manage multiplicity in trial design and analysis. This section will offer comprehensive insights into the practical application, strengths and limitations of these methods, providing a comprehensive understanding of their role in addressing multiplicity challenges within clinical research.

3. Methodology

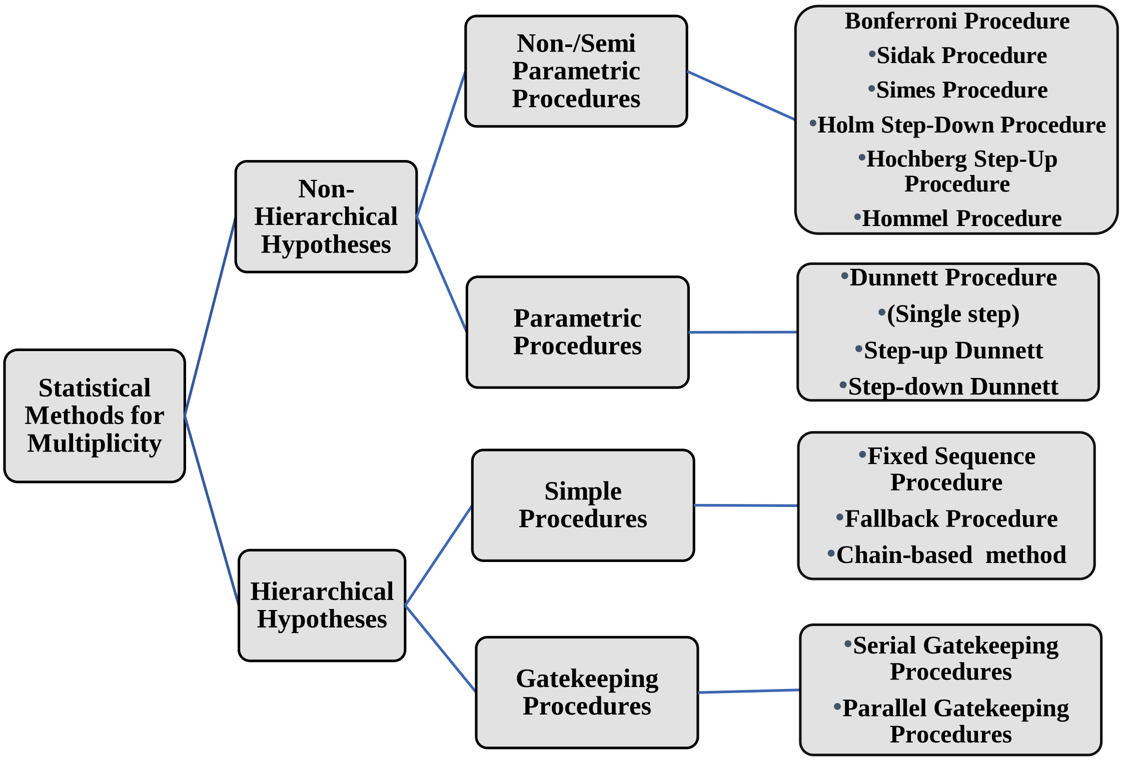
This section focuses on a range of methods used in confirmatory Phase III trials to address multiplicity issues. Various techniques have been proposed to maintain the Type I error rate when dealing with multiple comparisons. We will explore popular approaches known as multiple testing procedures (MTPs), with a particular emphasis on their implementation in R.

We will start by discussing fundamental single-step MTPs, like the Bonferroni procedure and then move on to more sophisticated data-driven stepwise MTPs, such as the Holm, Hochberg and Hommel procedures. Following that, we will delve into MTPs that rely on predetermined hypothesis testing ordering, including the fixed-sequence, fallback, and chain procedures.

Additionally, we will touch upon parametric methods for multiplicity adjustment, like the Dunnett procedure and its stepwise variations. Finally, we will introduce gatekeeping procedures designed for trials with multiple sources of multiplicity, like those with hierarchically ordered endpoints and multiple dose-placebo comparisons.

3.1 Overview of the Multiple Testing Procedures (MTP's):

Several classes of MTPs have been developed over the past 20 to 30 years that have found numerous applications in clinical trials. Selection of the most appropriate approaches to handle multiplicity in a particular setting is typically driven by several factors such as available clinical information (i.e., information on relevant logical restrictions or dependencies among the null hypotheses in a multiplicity problem) and statistical information (i.e., information on the joint distribution of the hypothesis test statistics) [1].



3.2 Single-step Procedures:

Single-step procedures are statistical methods used to address multiplicity issues in clinical trials by controlling the overall Type I error rate without requiring iterative steps. These procedures are relatively simple to implement and involve adjusting the significance level for each individual hypothesis test conducted in the trial. Here is a detailed explanation of some common single-step procedures,

3.2.1 Bonferroni Procedure (Non-Parametric)

The Bonferroni method is a single-step procedure that is commonly used, perhaps because of its simplicity and broad applicability [3]. The Bonferroni procedure is a statistical method used to adjust the significance level of individual hypothesis tests in situations where multiple comparisons are being made simultaneously. It is named after the Italian mathematician Carlo Emilio Bonferroni, who first introduced the method in the 1930s.

As we have discussed in Introduction part that, in a clinical trial or any research study where multiple comparisons are conducted, there is an increased risk of obtaining false positive results (Type I errors) due to chance alone. This risk arises because the more comparisons that are made, the greater the likelihood of finding a significant result by random variation, even if there is no true effect present.

The Bonferroni procedure addresses this issue by adjusting the threshold for statistical significance for each individual comparison. It does so by dividing the desired overall significance level (often denoted as α, typically set at 0.05) by the total number of comparisons being made and then comparing all the p-values to the adjusted significance. This can be expressed as,

$$\boldsymbol{\frac{\alpha}{m}}$$

where **'α'** is the significance level and '**m**' is the number of tests performing under the study.

Alternatively, if we want to adjust the p-value for each comparison while keeping the significance level constant, the adjusted p-value is calculated as,

$$\mathbf{p{*}m}$$

Here we compare each adjusted p-values to the desired overall significance level (0.05) and then conclude the results.

Example: Suppose you are conducting a clinical trial to test the effectiveness of a new drug on reducing symptoms of a disease. You measure 5 different symptoms, resulting in 5 separate hypothesis tests. If you set your overall significance level (alpha) at 0.05, the Bonferroni procedure would adjust this to 0.05/5 = 0.01. So, you would only declare a symptom significantly reduced if the p-value for that test is less than 0.01.

It is crucial to recognize that while the Bonferroni correction is successful in decreasing the risk of false positives, it simultaneously heightens the chance of false negatives (Type II errors), wherein genuine effects may go unnoticed. Moreover, the Bonferroni correction relies on the assumption of independence among comparisons, and its efficacy could wane if this assumption is breached. Another limitation is the conservative nature of the correction, which can make it challenging to detect true effects, particularly in studies with small sample sizes or weak treatment effects.[1]

3.2.2 Sidak Procedure (Semi-Parametric):

The Sidak procedure, named after the American statistician Jack W. Sidak, is another single-step method used to address multiplicity issues in clinical trials. Similar to the Bonferroni procedure, the Sidak method aims to control the overall Type I error rate when conducting multiple hypothesis tests simultaneously.

The Sidak procedure adjusts the significance level for each individual comparison to maintain a desired overall error rate. Instead of dividing the significance level by the number of tests (as in the Bonferroni correction), the Sidak method calculates the adjusted significance level using the formula,

$$\boldsymbol{1-(1-\alpha)^{\frac{1}{m}}}$$

where '$\boldsymbol{\alpha}$' represents the desired overall significance level (typically set at 0.05), and '$\mathbf{m}$' is the total number of comparisons being made. In simpler terms, the Sidak procedure calculates one adjusted significance level based on the total number of comparisons being made. This adjusted significance level is then compared with each individual raw p-value from the hypothesis tests.

Alternatively, Sidak procedure adjusts the p-values for each individual comparison while keeping the desired significance level constant. The formula for adjusting the p-values is,

$$\mathbf{1-(1-\boldsymbol{p}\;\boldsymbol{value})^{\boldsymbol{m}}}$$

After adjusting the p-values, they are compared with the desired significance level (often denoted as α, typically set at 0.05). If the adjusted p-value is less than or equal to α, the null hypothesis is rejected for that comparison. This process ensures that the overall probability of making a Type I error across all comparisons remains below the predetermined threshold.

Example: With the Sidak procedure, instead of simply dividing the overall significance level by the number of tests, you would adjust the significance level using the formula as, $\mathbf{1-(1-0.05)^{\frac{1}{5}}}$ **= 0.0102**. So, you would only declare a symptom significantly reduced if the p-value for that test is less than 0.0102.

Similar to the Bonferroni correction, the Sidak procedure is effective in reducing the risk of false positives (Type I error) but may increase the likelihood of false negatives (Type II error) due to its conservative nature. Additionally, like other single-step procedures, the Sidak method assumes independence among comparisons and may lose efficacy if this assumption is violated.[1]

3.2.3 Simes Procedure (Semi-Parametric):

The Simes test is also known as the Simes procedure, is a statistical method used in the context of multiple hypothesis testing. It was proposed by John Simes in 1986 as a less conservative alternative to the Bonferroni correction.

The goal of the Simes test is to control the familywise error rate (FWER), which is the probability of making at least one type I error when multiple hypotheses are tested simultaneously.

Here is how the Simes procedure works:

1. Suppose we have '**m**' null hypotheses, and we have computed '**m**' p-values corresponding to these hypotheses.

2. We then order these p-values from smallest to largest.

3. The Simes procedure rejects the first null hypothesis corresponding to the smallest p-value if it is less than or equal to **(1/m) * alpha**, where alpha is the desired overall significance level (often set to 0.05).

4. If the smallest p-value is not less than or equal to (1/m) * alpha, do not reject the null hypothesis associated with that p-value, and stop the procedure. Do not test any of the remaining hypotheses. Otherwise, proceed to second hypothesis and reject the second null hypothesis corresponding to this p-value if it is less than or equal to **(2/m) * alpha.**

5. We continue this process until you find a p-value that is not less than or equal to

(i/m) * alpha, where '**i**' is the rank of the p-value in the ordered list. At this point, you stop and do not reject any of the remaining null hypotheses.

Example: Suppose you have the following p-values for the 5 symptoms: 0.01, 0.02, 0.03, 0.04, and 0.05. You would order these from smallest to largest and compare each one to (i/5)*0.05, where i is the rank. The p-value of 0.01 (ranked 1) is less than (1/5)*0.05 = 0.01, so you would reject the null hypothesis for that symptom. The p-value of 0.02 (ranked 2) is not less than (2/5)*0.05 = 0.02, so you would stop there and not reject the null hypotheses for any of the other symptoms.

The Simes test is less conservative than the Bonferroni correction, which means it is more likely to reject false null hypotheses (i.e., it has more power). However, it also has a higher risk of making type I errors. Therefore, the choice between the Simes procedure and other methods depends on the specific context and the relative importance of avoiding type I errors versus having more power.[1]

3.2.4 Dunnett Procedure (Parametric):

The single-step Dunnett procedure can be thought of as the parametric counterpart of the Bonferroni procedure. Unlike the Bonferroni procedure, which does not depend on any distributional assumptions, the Dunnett procedure is based on the joint distribution of the test statistics associated with the hypotheses of interest. Thus, it accounts for the correlations among the test statistics. The use of the Dunnett procedure leads to more powerful inferences compared to the Bonferroni procedure.

Suppose the ‘m’ test statistics t_i follow a fully specified multivariate t distribution, and given a balanced design, it is easy to show that the test statistics are equally correlated with the common correlation coefficient of 0.5. Using this fact, Dunnett (1955) proposed to define the common one-sided critical value for the hypothesis test statistics $t_1, . . . , t_m$. This critical value is denoted by d_α(m, ν) and is found as the (1 − α)-quantile of the distribution of the maximum of t-distributed random variables with ν = (m+1)(n−1) degrees of freedom. Here ‘**m**’ represents the number of treatment groups being compared to the control group and ‘**n**’ represents the sample size of each treatment group.

The Dunnett procedure rejects the null hypothesis Hi if its test statistics is greater or equal to the common critical value, i.e., if

$$\boldsymbol{t_i \geq d_\alpha(\mathbf{m}, \nu)},\ \ \mathrm{i} = 1, . . . , \mathrm{m}.$$

For example, Consider the four hypothesis test statistics as,

$$\boldsymbol{t_1 = 2.68,\ \ t_2 = 2.23,\ \ t_3 = 3.30,\ \ t_4 = 3.00}$$

The one-sided, Dunnett-adjusted critical value is given by d_α(m, ν) with α = 0.05, m = 4, n=25, and ν = (m + 1)(n − 1) = 125 i.e.,

$$\boldsymbol{d_{0.05}(4, 125) = 2.57}\ (\text{approximately}).$$

Therefore, we reject t_3 and t_4 as the test statistic is greater than the tabulated value with (m,v) as (4,125) and alpha=0.05.

The parametric procedure defined rely on the assumption that the joint distribution of the hypothesis test statistics is fully specified. In this setting, parametric procedures provide a power advantage over p-value based procedure. In particular, the single-step Dunnett procedure corresponds to a parametric extension of the Bonferroni procedure. This single-step procedure dominates the Bonferroni procedure in terms of power, and it uses a common critical value to test all null hypotheses.[1]

3.2.5 Summary

The section discussed single-step multiplicity adjustment strategies, also known as single-step methods. These methods evaluate each null hypothesis independently, making the sequence of examination irrelevant. Single-step Multiple Testing Procedures (MTPs) are straightforward to implement and have gained significant impact in clinical applications.

Four single-step MTPs were discussed above,

- **The Bonferroni procedure:** This method manages the Family-Wise Error Rate (FWER) for any combined distribution of the marginal p-values. However, it is considered conservative. Despite its conservative nature, no single-step MTP is universally more potent than the Bonferroni procedure. More potent MTPs can only be developed if we are ready to make extra assumptions about the joint distribution of p-values linked with the null hypotheses of interest.[1]
- **The Sidak procedure:** This method is universally more potent than the Bonferroni procedure. However, its size relies on the joint distribution of the marginal p-values and can surpass the nominal level. The Sidak procedure manages the FWER when the test statistics are independent or follow a multivariate normal distribution. It offers a minor improvement over the Bonferroni procedure in general multiplicity issues. The Sidak procedure and related MTPs will not be further discussed.[1]
- **The Simes test:** This test is only applicable for testing the global null hypothesis. The Simes global test is more potent than the Bonferroni global test but does not always maintain the FWER. Its size is known to be no larger than the nominal level when the individual test statistics follow any multivariate normal distribution with non-negative correlation coefficients. The Simes global test is crucial in defining potent Simes-based stepwise procedures for addressing multiplicity in confirmatory clinical trials.[1]
- **The Dunnett test:** The Dunnett's procedure controls the Family-Wise Error Rate (FWER), similar to the Bonferroni and Sidak procedures. However, it is more powerful than the Bonferroni procedure when comparing multiple groups to a single control, as it considers the correlation between the comparisons. In summary, the Dunnett's procedure is a valuable tool in experimental design when multiple comparisons to a single control group are required. It offers a balance between statistical power and control of the FWER, making it a popular choice in many scientific fields.[1]

3.3 Hierarchical Hypothesis Ordering

3.3.1 Simple or Pre-specified Hypothesis Ordering:

Pre-specified hypothesis ordering methods are used in various fields such as artificial intelligence, machine learning, and data analysis. These methods are used to determine the order in which hypotheses are tested or evaluated. In this approach, hypotheses are arranged hierarchically or according to logical relationships, with some hypotheses considered primary or more important than others. The ordering typically reflects the study's objectives and the logical flow of the research questions.[1]

There are three methods in this pre specified hypotheses ordering which are,

I. Fixed Sequence Procedure (Non-Parametric):

In this method, hypotheses are tested in a pre-determined order. The order is usually based on the priority of the hypotheses, which can be determined by factors such as the importance of the hypothesis, the cost of testing it, or the likelihood of it being true. If a hypothesis is failed to be rejected, the process stops, and no further hypotheses are tested. If a hypothesis is rejected, the next hypothesis in the sequence is tested. This continues until a hypothesis is confirmed or all hypotheses have been tested.

In the Fixed Sequence Procedure, there is no explicit multiplicity adjustment required because hypotheses are tested sequentially, and the decision to stop testing further hypotheses is based on the outcomes of previous tests. Once a null hypothesis cannot be rejected at any step in the sequence, the testing process stops, and no further adjustments are made for multiplicity.

Here is how the fixed sequence procedure works,

i. **Pre-specified Hypothesis Order**: Before conducting the study, researchers establish a fixed sequence in which hypotheses will be tested. This sequence is typically determined based on logical or theoretical considerations relevant to the research question.

ii. **Sequential Testing**: Hypotheses are tested one after the other, following the predetermined order. Each hypothesis is subjected to statistical testing using appropriate methods, such as t-tests or ANOVA, to determine if there is sufficient evidence to reject the null hypothesis.

iii. **Stopping Rule**: If a null hypothesis cannot be rejected at any step in the sequence, the testing process stops. This means that no further hypotheses are evaluated beyond the point where the first non-rejected hypothesis occurs.

Example: Suppose a clinical trial is evaluating the effectiveness of three different treatments (A, B, and C) compared to a control group. The fixed sequence procedure might dictate testing Treatment A first, followed by Treatment B, and then Treatment C.

- If the null hypothesis for Treatment A (comparing it to the control group) is rejected, indicating a significant difference, the testing proceeds to Treatment B.
- If the null hypothesis for Treatment A is not rejected, the testing stops, and no further comparisons are made.

In summary, the fixed sequence procedure offers a structured approach to hypothesis testing, ensuring that hypotheses are evaluated in a logical order while controlling the overall Type I error rate.

However, it is essential to note that the Fixed Sequence Procedure also has limitations. For example, it may not be suitable for scenarios where the logical order of hypothesis testing is unclear or when there is a high degree of interdependence among hypotheses. Additionally, the sequential nature of the procedure may result in increased Type II error rates (false negatives) if later hypotheses are not adequately tested due to early stopping.[1]

II. Fallback Procedure (Non-Parametric):

The Fallback Procedure is a method used in hypothesis testing to control the overall Type I error rate when multiple hypotheses are tested sequentially. It incorporates the concept of weighted hypothesis testing, allowing for a flexible approach to hypothesis prioritization.

The fallback procedure offers a flexible approach to managing multiplicity, especially in scenarios where hypotheses are pre-arranged in a specific order. This method, provides an alternative to the fixed-sequence procedure, allowing for greater adaptability in multiplicity management.

In this procedure, the hypotheses H_1, H_2,….,H_m are organized in a predetermined sequence. Each hypothesis is assigned a weight w_1, w_2,….,w_m representing its relative importance. These weights are non-negative and sum up to 1, determine how the overall error rate (α) is distributed among the hypotheses. Specifically, the error rate allocated to each hypothesis H_i is calculated as αw_i, where i ranges from 1 to m.

The execution of the fallback procedure follows a defined algorithm based on the prearranged hypothesis order. This algorithm ensures that the error rate assigned to each hypothesis aligns with its weight, thus controlling the overall Type I error rate while accommodating the predetermined sequence of hypotheses.

The following algorithm is followed by fallback procedure which is,

Step 1. Test H_1 at α_1 = αw_1. If $p_1 \leq \alpha_1$, reject this hypothesis; otherwise, accept it. Go to the next step.

Steps i = 2, . . . , m − 1. Test H_i at α_i= α_i−1 + αw_i if H_i−1 is rejected and at α_i= αw_i if H_i−1 is accepted. If p_i≤ α_i, reject H_i; otherwise, accept it. Go to the next step.

Step m. Test H_m at α_m= α_m−1 + αw_m if H_m−1 is rejected and at α_m= αw_m if H_m−1 is accepted. If p_m≤ α_m, reject H_m; otherwise, accept it.

In contrast to the fixed-sequence approach, the fallback procedure maintains flexibility by allowing progression to the next hypothesis in the predetermined sequence, even if the current test yields nonsignificant results. For instance, if H_1 is not rejected, H_2 can still undergo testing at the significance level αw_2. Conversely, if H_1 is rejected, its error rate is transferred to H_2, permitting testing at a higher significance level, specifically, α(w_1+ w_2).

Example: Suppose we are conducting a clinical trial to compare the effectiveness of three different treatments (X, Y, and Z) in reducing blood pressure compared to a placebo. The hypotheses to be tested are: H_X (treatment X versus placebo), H_Y (treatment Y versus placebo), and H_Z (treatment Z versus placebo), each tested at a significance level of α = 0.05. We anticipate that treatment X will have the most significant effect, followed by treatment Y and then treatment Z.

We assign the following weights to the hypotheses: w_X = 1/2, w_Y = 1/3, and w_Z = 1/6, reflecting our expectation of their relative impact.

Initially, the significance levels assigned to H_X, H_Y, and H_Z are α/2, α/3, and α/6, respectively.

Now, let us apply the fallback procedure to a hypothetical scenario with the following p-values: p_X= 0.02, p_Y= 0.03, and p_Z= 0.06.

1. Start with testing H_X: Since $p_X \leq \alpha_1$= α/2 = 0.025, we reject H_X.
2. After rejecting H_X, the error rate from H_X is carried over to H_Y,. Test H_Y at significance level α_2= α_1+ α/3 = 0.025 + α/3 = 0.0417. Since $p_Y \leq$ 0.0417, we reject H_Y.
3. Proceed to test H_Z: At this final step, test H_Z at significance level α_3= α_2+ α/6 = 0.0417 + α/6 = α = 0.05. If $p_Z \leq$ 0.05, reject H_Z.

Even if H_Y,was not rejected in Step 2, H_Z would still be tested in Step 3 but at a lower significance level of α/6.

This example demonstrates how the fallback procedure adapts the significance levels for each hypothesis test based on the outcomes of preceding tests, ensuring effective control of the overall Type I error rate while allowing for flexible testing sequences.[1]

III. Chain based Graphical Procedure (Non-Parametric):

The chain-based graphical procedure, also called the closed testing procedure (CTP), is a highly advanced method for handling multiple comparisons in hypothesis testing. It is like a powerful tool that tackles the problem of testing many hypotheses at once while considering how they are related, how important each one is, and the order in which they are tested. It offers a comprehensive approach that considers weights assigned to hypotheses, alpha propagation rules, and transition matrices to control the familywise error rate (FWER) effectively.[1]

Here is a detailed explanation of the components involved in the chain-based graphical procedure:

1. **Hierarchical Structure:** This procedure begins by establishing a hierarchical structure among the hypotheses, often depicted graphically as a directed acyclic graph (DAG). In this graph, each hypothesis is represented as a node, and the

logical connections between hypotheses are denoted by directed edges. This hierarchical arrangement serves as the foundation for the subsequent testing process, ensuring that dependencies and logical relationships are considered during hypothesis evaluation.

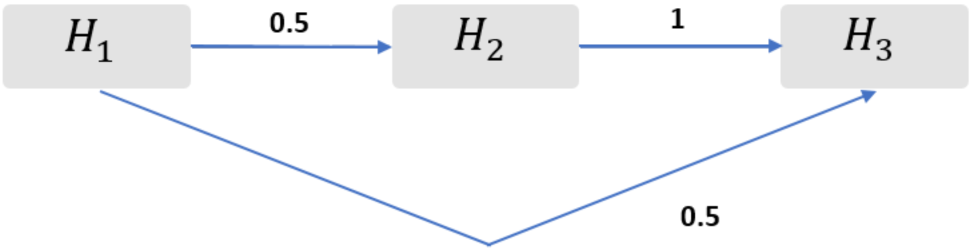
2. **Testing Paths:** Once the hierarchical structure is established, the procedure identifies all possible testing paths within the graph. A testing path represents a sequence of hypotheses that can be tested together, ensuring that the overall FWER is controlled. These paths are determined based on the logical connections between hypotheses as defined by the graph's structure.
3. **Weighted Hypotheses:** One of the key features of the chain-based graphical procedure is its ability to incorporate weighted hypotheses. Recognizing that not all hypotheses carry equal significance, researchers have the flexibility to assign weights to each hypothesis based on its relative importance in addressing the research question. These weights influence the decision-making process, allowing hypotheses with greater relevance to receive more emphasis during testing.
4. **Alpha Propagation Rules:** As the procedure progresses along testing paths defined by the hierarchical structure, it adheres to alpha propagation rules that govern the adjustment of significance levels. Alpha, the threshold for statistical significance, is propagated sequentially along the testing paths, with adjustments made based on the weights assigned to hypotheses encountered along the way. This dynamic adjustment of alpha ensures that the overall Type I error rate remains controlled as hypotheses are tested sequentially.
5. **Transition Matrix:** The transition matrix serves as a guiding framework for alpha propagation, providing systematic rules for adjusting significance levels at each step of the testing process. It is basically a matrix which tells us how the alpha is propagated from one hypothesis to another, and this propagation will be done at the initial stage of the study. By considering the weights of hypotheses and the logical connections between them, the transition matrix facilitates the precise control of Type I error rates while navigating the complex hierarchical relationships within the hypothesis set. [5]

Example: Suppose if we consider the weighted Bonferroni-Holm procedure within the chain-based graphical framework, we will illustrate how graphical tools aid in managing multiplicity issues. Each elementary hypothesis (H_1, H_2, H_3) is represented by a vertex in the graph, with associated weights denoting local significance levels. Let us assume initial significance levels allocated to H_1, H_2, and H_3 are denoted as α_1, α_2, and α_3, respectively, with the constraint that $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$.

Now, let us consider the alpha propagation rule: if H_1 is rejected, half of the initial significance level (α_1) is transferred to both H_2 and H_3. Similarly, if H_2 is rejected, the entire significance level of H_2 is transferred to H_3.

Visualizing this scenario using the weighted Bonferroni-Holm procedure graphically, we can observe the initial allocation of the overall significance level α to each hypothesis. If, for instance, H_1 is rejected, half of the initially allocated significance level ($\alpha_1$at vertex H_1) is passed on to both H_2 and H_3 (as indicated by directed edges with associated weight 0.5). Subsequently, if H_2 is rejected, the entire significance level of H_2 is fully transferred to H_3.

This graphical representation effectively illustrates how the weighted Bonferroni-Holm procedure operates within the chain-based graphical framework, facilitating a nuanced understanding of the allocation and propagation of significance levels across multiple hypotheses. [5]



Graphical illustration of the weighted Bonferroni–Holm procedure with three hypotheses.

Note: All the three pre-specified hypotheses ordering methods discussed above will be discussed in much more detail in the power comparision using simulations section by considering an example.

3.3.2 Gatekeeping Procedures

This section discusses the strategies used in clinical trials to handle the challenges posed by conducting multiple tests with various objectives, such as comparing different treatments or analyzing different subgroups. These tests often lead to what we call multiplicity issues, where the more tests we conduct, the greater the chance of finding false results just by luck.

To address this, we organize the tests into groups, or families, and test them in a specific order. This means that the results of earlier tests in one family can affect whether we proceed to test other hypotheses in subsequent families. It is like passing through a series of gates, where each gate represents a different set of tests.

Gatekeeping procedures are the main approach used to manage these complex situations. These procedures ensure that we control the overall risk of false findings while still allowing us to test multiple hypotheses. There are different types of gatekeeping procedures, depending on how the tests are organized. Some focus on tests that must pass one after the other (serial gatekeepers), while others allow tests to proceed independently (parallel gatekeepers).

In this section, we will focus on gatekeeping methods that use simple statistical techniques, like Bonferroni and Holm methods. These methods are straightforward to understand and implement, making them suitable for many situations in clinical trials. More advanced methods, based on complex mathematical models, are discussed elsewhere.

There are three methods in this gatekeeping procedures which are,

I. Serial Gatekeeping Procedure:

The serial gatekeeping procedure is a method used in clinical trials to manage the issue of multiplicity, which arises when conducting multiple statistical tests. It is designed to control the overall risk of making false discoveries while still allowing for the testing of multiple hypotheses.

Suppose you have several hypotheses you want to test in a trial. Instead of testing them all at once, you organize them into groups or families, with each family representing a different set of hypotheses. You then test these families one after the other, in a specific order.

Now, picture each family of hypotheses as a gate in a series. Before you can proceed to test the hypotheses in a particular family, you must first pass through the gates representing the earlier families. In other words, the outcome of the tests in one family influences whether you can proceed to test the hypotheses in the next family.

This sequential testing approach ensures that you maintain control over the overall risk of making false discoveries. If a hypothesis fails to meet the criteria for statistical significance in one family, you will not proceed to test hypotheses in subsequent families. This helps prevent the accumulation of false positives that can occur when conducting numerous tests simultaneously.

The serial gatekeeping procedure is particularly useful when you have hypotheses that build upon each other or when you want to prioritize certain comparisons over others. By structuring the tests in a sequential manner, you can efficiently manage the multiplicity issue without sacrificing the integrity of your trial results.

Example: Let us consider a clinical trial investigating the effectiveness of two different treatments (A and B) compared to a placebo. We want to test multiple hypotheses related to the efficacy of each treatment.

In the serial gatekeeping procedure, we organize these hypotheses into families based on their relationships. For example:

Family1 (Treatment A versus Placebo):

Hypothesis 1: Treatment A is superior to Placebo.

Hypothesis 2: Treatment A is non-inferior to Placebo.

Hypothesis 3: Treatment A has a specific effect on a subgroup.

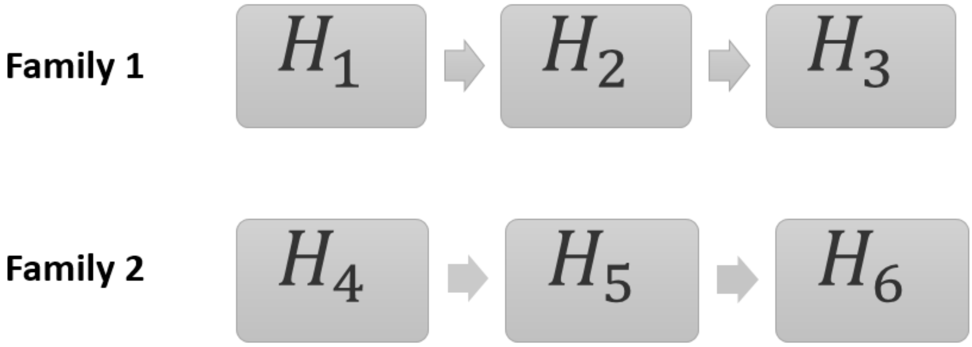
Family2 (Treatment B versus Placebo):

Hypothesis 1: Treatment B is superior to Placebo.

We begin by testing all hypotheses in Family1. If any hypothesis fails to demonstrate significance, we stop testing within Family1 and conclude that Treatment A does not have the desired effect.

If all hypotheses in Family1 show significance, we proceed to test the single hypothesis in Family2. If this hypothesis fails to show significance, we conclude that Treatment B is not superior to Placebo.

This sequential approach ensures that we maintain control over the overall Type I error rate, while also allowing us to efficiently evaluate multiple treatments in a structured manner.



From the above graph, we can clearly see that Family1 is acting as a serial gatekeeper for Family2. We proceed to Family2 only when all the hypotheses in Family1 are significant i.e., H_1, H_2, and H_3 are significant.

II. Parallel Gatekeeping Procedure:

The parallel gatekeeping procedure is another method used to address multiplicity issues in clinical trials. Unlike the serial gatekeeping procedure, which tests hypotheses sequentially, the parallel approach allows for simultaneous testing of multiple hypotheses.

In the parallel gatekeeping procedure, if any hypothesis within a family is rejected, it does not prevent us from testing hypotheses in other families. Each family of hypotheses is tested independently of the others. Therefore, if any hypothesis in Family 1 is rejected, we can still proceed to test hypotheses in Family 2.

Example: Suppose we have two families of hypotheses:

Family1 (Treatment A versus Placebo):

Hypothesis 1: Treatment A is superior to Placebo

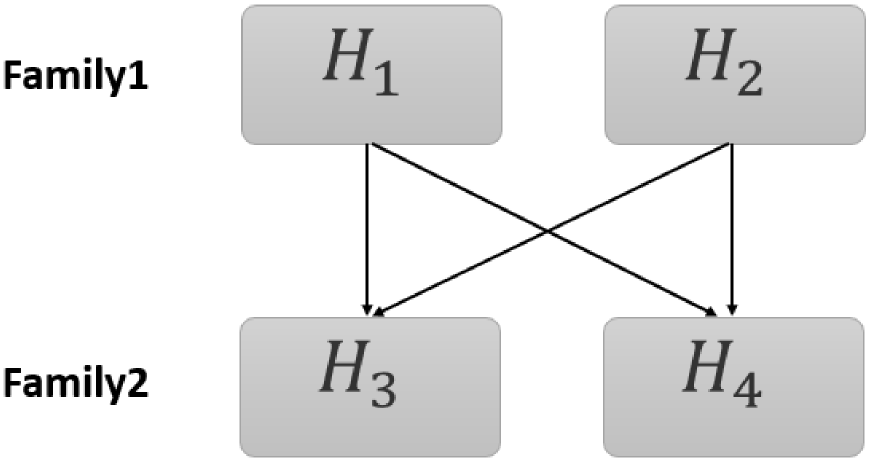
Hypothesis 2: Treatment A is non-inferior to Placebo

Family2 (Treatment B versus Placebo):

Hypothesis 1: Treatment B is superior to Placebo

In the parallel gatekeeping procedure, we would test all hypotheses in both families simultaneously. If, for example, Hypothesis 1 in Family1 is rejected (indicating that Treatment A is indeed superior to Placebo), we can still proceed to test Hypothesis 1 in Family2 (comparing Treatment B to Placebo) without any constraints imposed by the rejection in Family1.

This approach allows for more flexibility in the testing process and ensures that the evaluation of hypotheses in one family does not hinder the evaluation of hypotheses in other families.



From the above graph we can clearly see that Family1 is acting as a Parallel gatekeeper to Family2. If at least one of the hypothesis is significant from Family1 i.e., H_1 or H_2 we can proceed to test Family2.

III. Multi-Sequence Gatekeeping Procedure:

The multi-sequence gatekeeping procedure is a sophisticated method used in clinical trials to address multiplicity issues when testing multiple hypotheses in a hierarchical order. Unlike simpler procedures like Bonferroni correction, the multi-sequence gatekeeping approach allows for more complex testing strategies that consider the interdependence between different families of hypotheses.

In the multi-sequence gatekeeping procedure, the testing of hypotheses is organized into multiple sequences, each representing a distinct set of related hypotheses. The key principle is that the testing within each sequence is contingent upon the outcomes of tests in preceding sequences.

Example: Consider a clinical trial evaluating the efficacy of three different treatments (A, B, and C) compared to a control (Placebo) for a certain medical condition. The trial is designed to test multiple hypotheses organized into two sequences:

Sequence 1 (Comparing Treatments A, B, and C to Placebo):

Hypothesis 1: Treatment A is superior to Placebo

Hypothesis 2: Treatment B is superior to Placebo

Hypothesis 3: Treatment C is superior to Placebo

Sequence 2 (Comparing Treatments A, B, and C to Each Other):

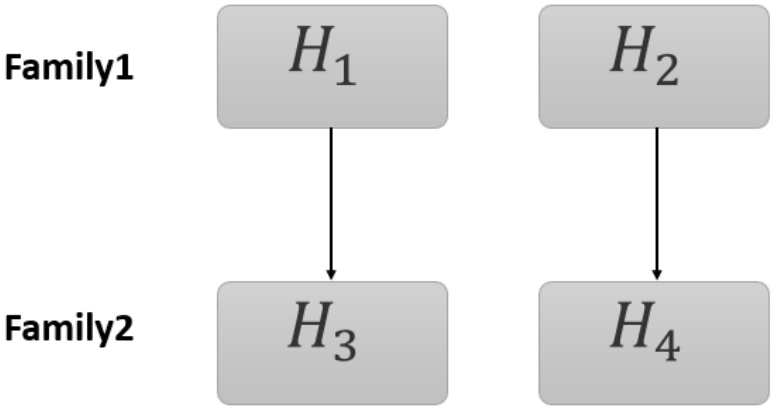
Hypothesis 4: Treatment A is superior to Treatment B

Hypothesis 5: Treatment A is superior to Treatment C

Hypothesis 6: Treatment B is superior to Treatment C

The testing begins with Sequence 1. Each hypothesis in this sequence is tested sequentially. If any hypothesis within Sequence 1 is rejected (let us say H_1), the testing proceeds to Sequence 2. In Sequence 2, only the hypotheses corresponding to the rejected hypotheses in Sequence 1 are tested. So, if H_1 is rejected, H4 is tested; if $H_2$is rejected, H_5 is tested and if H_3 is rejected, H_6 is tested.

However, if none of the hypotheses in Sequence 1 are rejected, the testing does not proceed to Sequence 2, and the hypotheses in Sequence 2 are not tested.



From the above graph we can see that Family1 is acting as a multi-Sequence gatekeeper to Family2, where if H_1 is significant then only we proceed to test H_3, similarly if H_2 is significant then only we proceed to test H_4.

3.3.3 Summary:

The multiple testing procedures discussed in this section depend on a predetermined sequence of hypotheses and offer an alternative to hypothesis testing based on empirical data when the null hypotheses follow a natural order.

- **Fixed-Sequence:** The fixed-sequence procedure exhibits optimal performance under conditions where the treatment effect shows a consistent trend, such as a monotonic increase or decrease over time or dosage. However, if the assumption of monotonicity is violated, the fixed-sequence approach may yield misleading or inaccurate outcomes.[1]
- **Fallback:** The fallback procedure was introduced as a flexible substitute for the fixed-sequence method. Like its counterpart, it evaluates hypotheses in a predetermined order. However, it differs by not exhausting the entire α-level at each step, allowing for a fallback strategy if a null hypothesis remains unrejected.[1]
- **Chain Procedure:** Both fixed-sequence and fallback procedures belong to a broader category known as chain procedures. These methods provide flexibility in decision-making by allowing customization of hypothesis weights and alpha propagation rules. Importantly, the chain procedures discussed here are nonparametric, meaning they do not rely on assumptions about the underlying distribution of test statistics.[5]

- **Serial Gatekeeping Procedure:** The serial gatekeeping procedure involves testing hypotheses sequentially, where the acceptance or rejection of hypotheses in a particular family depends on the outcome of significance tests carried out in the preceding families. In other words, hypotheses are grouped into families, and the testing progresses through these families in a predefined order. This procedure ensures that the overall Type I error rate is controlled while allowing for a hierarchical testing approach.
- **Parallel Gatekeeping Procedure:** The parallel gatekeeping procedure allows for more flexibility by testing hypotheses in multiple families simultaneously. This approach is useful when hypotheses across different families are not interdependent and can be tested concurrently. It provides a less restrictive framework compared to the serial procedure while still controlling the overall Type I error rate.
- **Multi-Sequence Gatekeeping Procedure:** The multi-sequence gatekeeping procedure extends the concept of sequential testing further by allowing for multiple testing sequences based on the outcomes of preceding hypotheses. Hypotheses are organized into sequences, and the testing progresses through these sequences sequentially. However, unlike the serial procedure, the multi-sequence approach allows for branching, where the testing can proceed along different sequences based on the rejection or acceptance of hypotheses at each step. This flexibility makes it suitable for complex testing scenarios where hypotheses may have different dependencies and hierarchical structures.

4. Power Comparison of Hierarchical Methods of Multiplicity Adjustment using Simulations

In this section, we shall delve into the practical application and performance evaluation of hierarchical methods outlined in Section 3.3.1 and 3.3.2. Through the utilization of simulated data, we aim to elucidate the efficacy and utility of these methods in real-world scenarios.

Furthermore, we aim to conduct power comparisons for all of these hierarchical methods, considering various sample sizes and weights. This evaluation will be based on 10,000 simulations, allowing for a robust analysis of the methods performance under different conditions. Subsequently, we will analyze the results, explaining the strengths and weaknesses of each method in comparison.

4.1 Simulation Study:

Objective: The aim of this study is to demonstrate the non-inferiority of the immune response and evaluate safety of XXX investigational vaccine in adults 50-59 years of age (YOA), including those who are at increased risk of XXX-lower respiratory tract disease (LRTD), versus adults ≥60 YOA, where vaccine efficacy against XXX-LRTD is being assessed in another clinical study.

The End points and the Hypotheses are defined as,

1. **Null hypothesis 1 (H_1):** The anti-XXX-A GMT ratio (OA-XXX Group over HA-XXX Group) is >1.5 or the SRR difference (OA-XXX Group – HA-XXX Group) is >10% at 1 month post XXX vaccine administration.
2. **Null hypothesis 2 (H_2):** The anti-XXX-B GMT ratio (OA-XXX Group over HA-XXX Group) is >1.5 or the SRR difference (OA-XXX Group – HA-XXX Group) is >10% at 1 month post XXX vaccine administration.
3. **Null hypothesis 3 (H_3):** The anti-XXX-A GMT ratio (OA-XXX Group over AIR-XXX Group) is >1.5 or the SRR difference (OA-XXX Group – AIR-XXX Group) is >10% at 1 month post XXX vaccine administration.
4. **Null hypothesis 4 (H_4):** The anti-XXX-B GMT ratio (OA-XXX Group over AIR-XXX Group) is >1.5 or the SRR difference (OA-XXX Group – AIR-XXX Group) is >10% at 1 month post XXX vaccine administration.

Here the XXX Investigational vaccine has two components A and B, where this vaccine's GMT ratio and SRR is to be compared among OA and HA at 1 month post vaccine.

Where GMT is Geometric Mean Titer Ratio.

SRR is Seroconversion Rate Ratio.

OA is Older Adults (>60 YOA).

HA is Healthier Adults (50 to 59 YOA).

AIR is At Increased Risk (50 to 59 YOA).

Note: Clinical team has identified a Hierarchy among four Hypotheses H_1, H_2, H_3, H_4

Since we have 4 hierarchical hypotheses, the use of some specified Multiplicity adjustment methods needs to be done in order to control the inflation of Family Wise Error Rate which are,

1. Fixed Sequence Procedure
2. Fallback Procedure
3. Chain Procedure
4. Parallel gatekeeping procedure
5. Multi-Sequence gatekeeping procedure

Note: Since Fixed Sequence, Parallel gatekeeping and multi-Sequence gatekeeping procedures does not require or consider any weights, we consider the initial weights and alpha levels for Fallback and Chain Procedure, which are taken as,

i. The desired alpha or the significance level is **0.025**

ii. Let us consider three sets of Hypothesis weights

   i. $\boldsymbol{w_1}$ **= 1,** $\boldsymbol{w_2}$ **= 0,** $\boldsymbol{w_3}$ **= 0 and** $\boldsymbol{w_4 = 0}$.

   ii. $\boldsymbol{w_1}$ **= 3/4,** $\boldsymbol{w_2}$ **= 1/4,** $\boldsymbol{w_3}$ **= 0 and** $\boldsymbol{w_4 = 0}$.

   iii. $\boldsymbol{w_1}$ **= 1/2,** $\boldsymbol{w_2}$ **= 1/4,** $\boldsymbol{w_3}$ **= and** $\boldsymbol{w_4}$ **= 1/8.**

iii. Initial significance levels for the four hypotheses by considering the second set of weights are then given by,

$\boldsymbol{\alpha_1}$ = 0.025*0.75 = **0.01875,** $\boldsymbol{\alpha_2} =$ 0.025*0.25 = **0.00625,** $\boldsymbol{\alpha_3 = 0}$ **and** $\boldsymbol{\alpha_4 = 0}$**.**

4.2 Methodology of the study:

1. **Study Design**

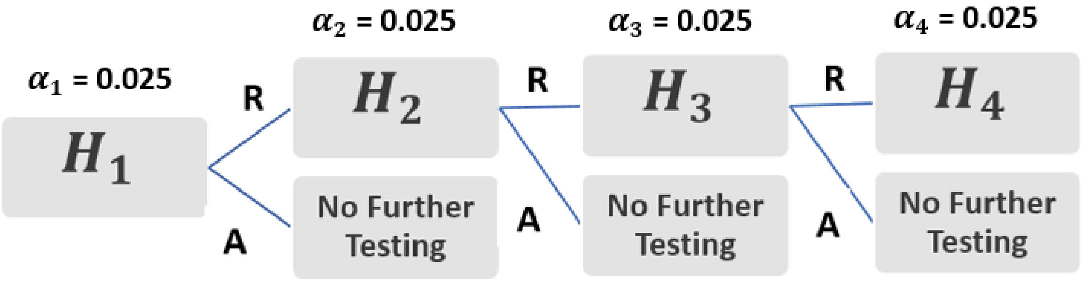
We conducted a simulation study to compare the power of hierarchical multiplicity adjustment methods: fixed sequence, fallback, chain-based graphical procedure, parallel gatekeeping and multi-sequence gatekeeping procedure. This study aimed to evaluate the performance of these methods under various scenarios commonly encountered in clinical trial settings.

2. **Data Generation**

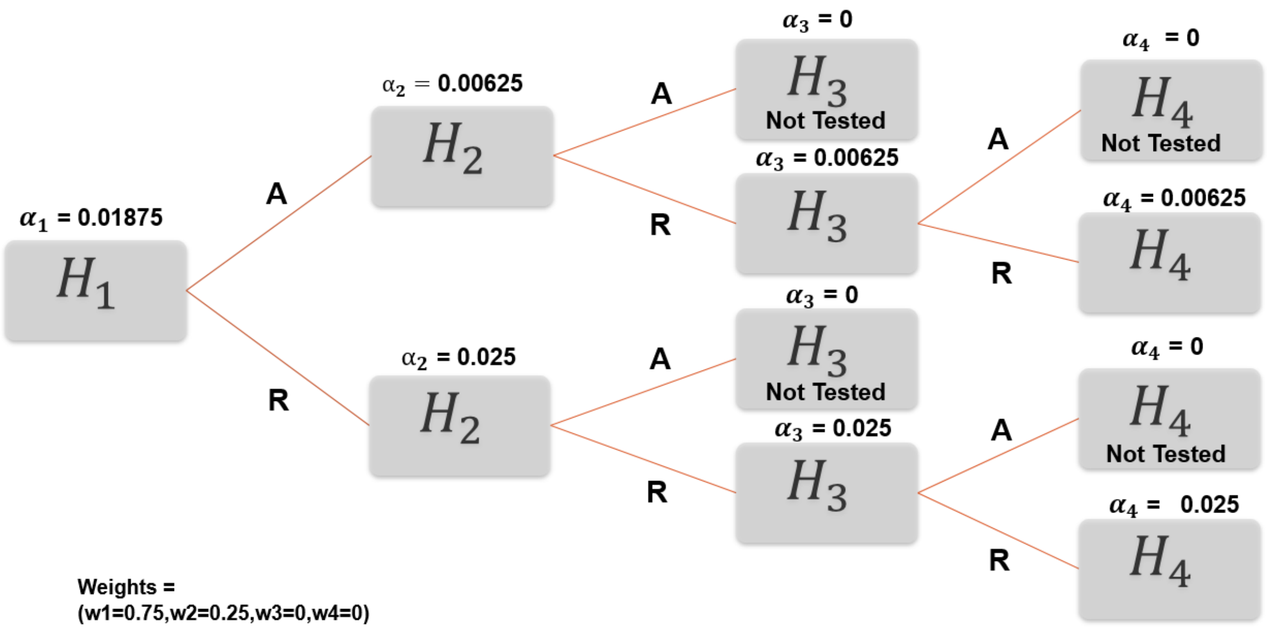
To replicate the structure of a clinical trial data was generated using R programming language. We considered four hypotheses representing different treatment comparisons. Each hypothesis was associated with specific mean treatment effects and standard deviations. Data generation was performed to ensure realistic distributions of treatment effects and variability within treatment groups.

3. **Multiplicity Adjustment Methods**

- **Fixed Sequence**: In this method, hypotheses were tested sequentially according to a predetermined order. The significance level for each test remained constant throughout the sequence. Here is the idea or the graphical overview about how this procedure works.



- **Fallback**: Similar to the fixed sequence method, but with a fallback strategy in case a null hypothesis was not rejected. If a hypothesis was not rejected, the remaining hypotheses were tested at a reduced significance level to maintain overall error control. Here is how the fallback Procedure works with the initial weights considered above,

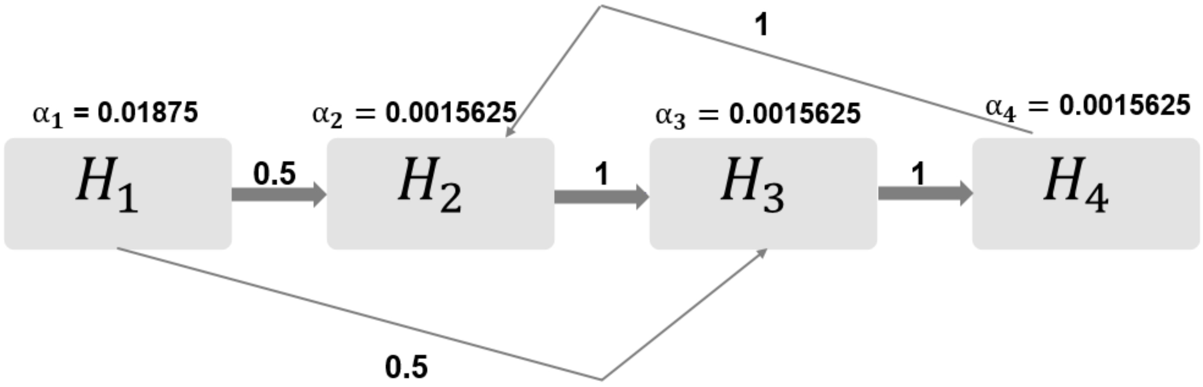


- **Chain-based Graphical Procedure**: This method allowed for more flexible decision-making based on the outcomes of previous tests. Hypothesis testing decisions depended on predefined decision rules and the hierarchical structure of the hypotheses.

  Chain procedures are similar to the fallback procedure in the sense that weights are pre-assigned to each hypothesis in a multiplicity problem, and it applies propagation rules to ensure that the Type I error rate is controlled.

  The main difference with the fallback procedure lies in the fact that chain procedures support more flexible α propagation rules, e.g., after each rejection, the error rate can be transferred simultaneously to several hypotheses.

  Here is the overview of how the chain-based procedure works,

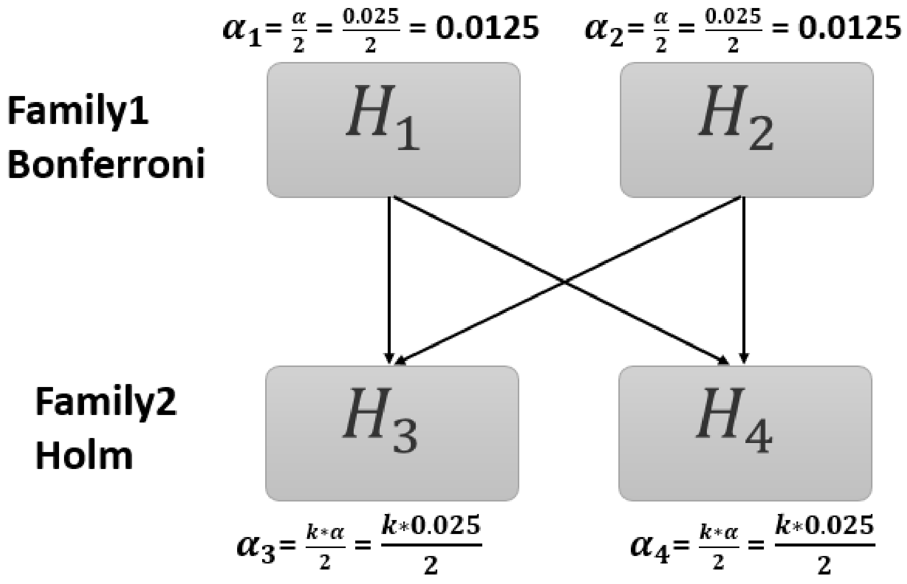


Here the weights and alpha propagation rules are decided at the initial stage of the study that is, the first hypothesis is tested at 0.01875 and if it is rejected half of the significance level is transferred to both second and third hypotheses i.e., α_2 = α_1/2 + α_2*w_2 = 0.0015625. Similarly, if second hypothesis is rejected, entire significance is transferred to third hypothesis and same goes for fourth.

Note: If second hypothesis is failed to be rejected then no transfer of significance level is done to third hypothesis then third hypothesis significance level will be α_3= α_3*w_3 = 0. Therefore, no testing will be done since the significance level is zero in this case. Similarly for fourth hypothesis as well. [1]

- **Parallel and multi-sequence gatekeeping Procedure:** We have already discussed about these methods with examples above in the section 3.32. Both procedures are valuable tools in controlling the Type I error rate in complex multiple testing scenarios, but they differ in their approach to determining the testing sequence and the handling of subsequent hypotheses based on previous outcomes.

Overview of Parallel Gatekeeping Strategy in the study


$\alpha_1 = \frac{\alpha}{2} = \frac{0.025}{2} = 0.0125$
$\alpha_2 = \frac{\alpha}{2} = \frac{0.025}{2} = 0.0125$
Family1
Bonferroni
H_1
H_2
Family2
Holm
H_3
H_4
$\alpha_3 = \frac{k*\alpha}{2} = \frac{k*0.025}{2}$
$\alpha_4 = \frac{k*\alpha}{2} = \frac{k*0.025}{2}$

Where K is the number of hypotheses rejected in Family1 (k = 1, 2).

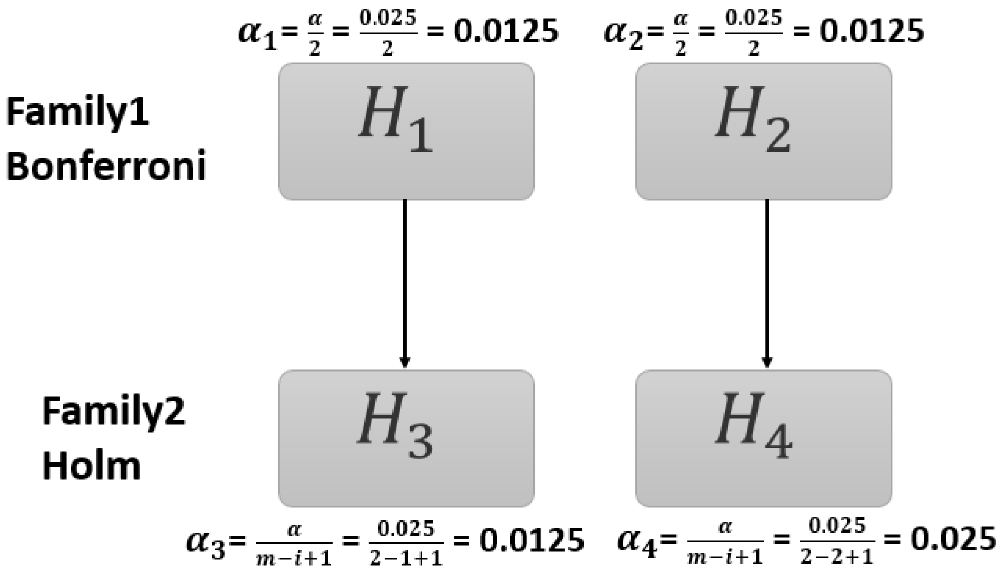
Another critical requirement is that the individual procedures must be separable. A separable procedure can be likened to a "generous" approach, allowing for the transfer of the error rate to the next family even if certain null hypotheses are accepted within the current family. It is important to note that the final family in the sequence does not necessarily need to be separable, as there are no subsequent families to which the error rate needs to be transferred. The Bonferroni procedure is recognized for its separability and holm is a non-separable method. [1]

The Decision rules for the parallel gatekeeping strategy are,

Step 1: Test H_1 and H_2 at α = 0.025 using the Bonferroni procedure. If at least one hypothesis is rejected, go to Step 2. Otherwise, accept all hypotheses and stop.

Step 2: Let k be the number of hypotheses rejected in Step 1 (k = 1, 2). Test H_3 and H_4 at kα/2 using the Holm procedure.

Overview of the multi-sequence gatekeeping Procedure



$\alpha_1 = \frac{\alpha}{2} = \frac{0.025}{2} = 0.0125$
$\alpha_2 = \frac{\alpha}{2} = \frac{0.025}{2} = 0.0125$
Family1
Bonferroni
H_1
H_2
Family2
Holm
H_3
H_4
$\alpha_3 = \frac{\alpha}{m-i+1} = \frac{0.025}{2-1+1} = 0.0125$
$\alpha_4 = \frac{\alpha}{m-i+1} = \frac{0.025}{2-2+1} = 0.025$

Where 'm' is the number of tests and 'i' is the number of tests currently performing (i=1,2)

Similar to Parallel gatekeeping strategy, multi-sequence gatekeeping also performs using a separable method for Family1 which is Bonferroni and non-separable method for Family2 which is Holm's method.

The Decision rules for this procedure are,

Family1

i. Reject H_1 if $p_1 \leq \alpha_1$; Otherwise accept H_1 and proceed to next hypothesis.

ii. Reject H_2 if $p_2 \leq \alpha_2$; Otherwise accept H_2 and stop further testing.

Family2

iii. If H_1 is significant then proceed to test H_3 at $p_3 \leq \alpha_3$; If H_1 is not significant do not test H_3.

iv. If H_2 is significant then proceed to test H_4 at $p_4 \leq \alpha_4$; If H_2 is not significant do not test H_4.

4. **Power Calculation**

Power for each hypothesis under each multiplicity adjustment method was calculated as the proportion of simulations where the null hypothesis was correctly rejected. A significance level of $\alpha = 0.025$ was used for all tests. Power calculations were performed separately for each hypothesis and aggregated across simulations for each method.

5. **Simulation Procedure**

Simulations were conducted using the R programming language. For each scenario, 10,000 datasets were generated to ensure robustness of results. Hypothesis tests were conducted according to the specified multiplicity adjustment method for each dataset.

6. **Statistical Analysis**

Simulation results were analyzed using descriptive statistics to summarize the power of each method across scenarios. Mean power and standard deviation were calculated to assess the consistency and variability of results. Statistical significance of differences in power between methods was assessed using appropriate inferential test, which is t-test.

7. **Ethical Considerations**

As this study involved only simulated data, no ethical approval was required. However, ethical principles of data privacy and confidentiality were upheld throughout the study.

8. **Statistical Software**

Data generation, analysis, and visualization were performed using R. Standard R packages were used for statistical analysis, and custom scripts were developed for simulation procedures and result visualization.

4.3 Power Comparison of the Study with 10,000 Simulations:

In this section we will simulate a dataset and do the power comparison of these five hierarchical methods and interpret the results based on the method that is performing well under different sample sizes and weights.

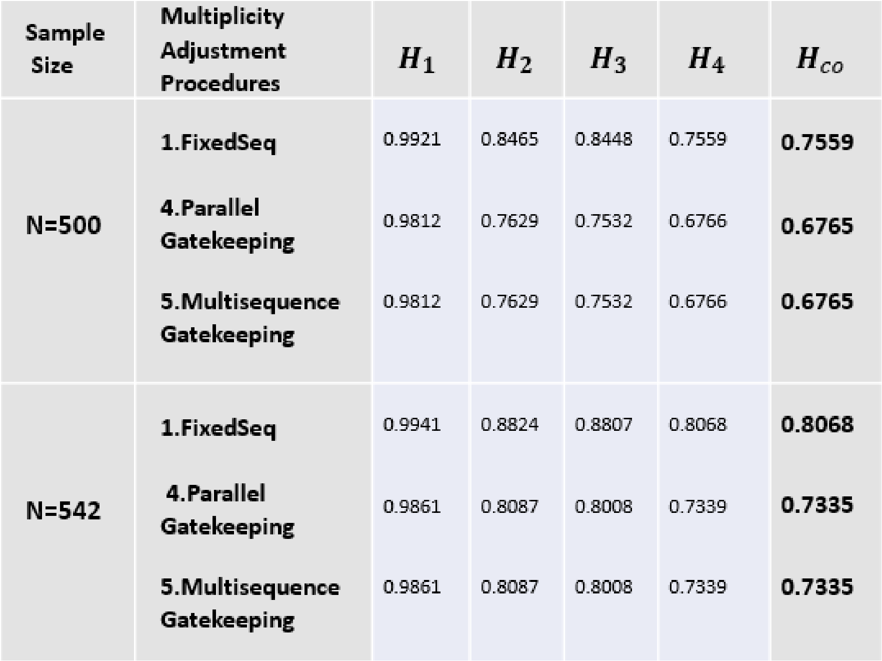
We simulated the data according to our study which includes pre vaccination and post vaccination of two different components A and B of XXX vaccine. Utilizing the mean and covariance structure derived from prior research, we modeled the data to follow a Multivariate Log Normal Distribution with a moderate correlation structure.

Upon simulating the data and obtaining the necessary parameters, we performed 10,000 simulations for sample sizes of 500 and 542, utilizing three different sets of hypothesis weights: (0.75, 0.25, 0, 0), (0.50, 0.25, 0.125, 0.125), and (1, 0, 0, 0).

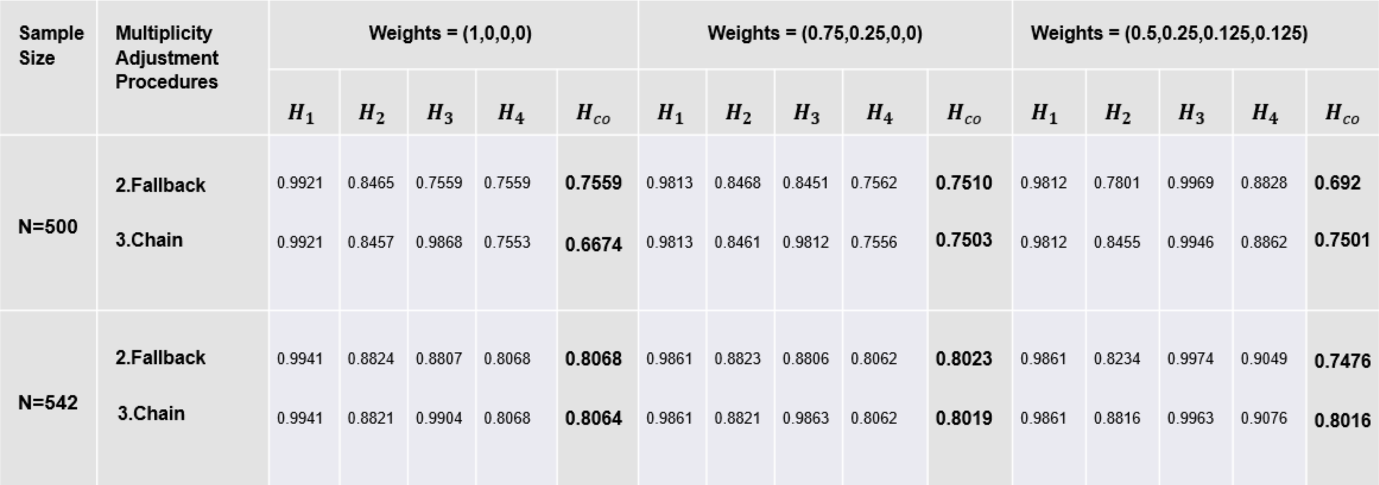
Subsequently, we applied each hierarchical multiple testing procedure to analyze the simulated data. Power calculation for each method was performed, indicating the proportion of times all four null hypotheses were correctly rejected given that each null hypothesis is false.

Upon completion of the power analysis, we further scrutinized the performance of each method under varying conditions of sample sizes and weights, seeking to discern the most robust approach for our study's objectives.

4.3.1 Simulation Results:



Sample Size	Multiplicity Adjustment Procedures	H_1	H_2	H_3	H_4	H_{co}
N=500	1.FixedSeq	0.9921	0.8465	0.8448	0.7559	**0.7559**
	4.Parallel Gatekeeping	0.9812	0.7629	0.7532	0.6766	**0.6765**
	5.Multisequence Gatekeeping	0.9812	0.7629	0.7532	0.6766	**0.6765**
N=542	1.FixedSeq	0.9941	0.8824	0.8807	0.8068	**0.8068**
	4.Parallel Gatekeeping	0.9861	0.8087	0.8008	0.7339	**0.7335**
	5.Multisequence Gatekeeping	0.9861	0.8087	0.8008	0.7339	**0.7335**



Sample Size	Multiplicity Adjustment Procedures	Weights = (1,0,0,0)					Weights = (0.75,0.25,0,0)					Weights = (0.5,0.25,0.125,0.125)				
		H_1	H_2	H_3	H_4	H_{co}	H_1	H_2	H_3	H_4	H_{co}	H_1	H_2	H_3	H_4	H_{co}
N=500	2.Fallback	0.9921	0.8465	0.7559	0.7559	**0.7559**	0.9813	0.8468	0.8451	0.7562	**0.7510**	0.9812	0.7801	0.9969	0.8828	**0.692**
	3.Chain	0.9921	0.8457	0.9868	0.7553	**0.6674**	0.9813	0.8461	0.9812	0.7556	**0.7503**	0.9812	0.8455	0.9946	0.8862	**0.7501**
N=542	2.Fallback	0.9941	0.8824	0.8807	0.8068	**0.8068**	0.9861	0.8823	0.8806	0.8062	**0.8023**	0.9861	0.8234	0.9974	0.9049	**0.7476**
	3.Chain	0.9941	0.8821	0.9904	0.8068	**0.8064**	0.9861	0.8821	0.9863	0.8062	**0.8019**	0.9861	0.8816	0.9963	0.9076	**0.8016**

4.3.2 Interpretation of Results:

This table summarizes the average power of each hierarchical method across all simulations for different combinations of sample sizes, hypothesis weights, and correlation structure, providing valuable insights into their relative performance.

We can interpret the results from the above table by comparing the power of different methods also considering the weights of the hypotheses.

Sample Sizes 500 and 542

Hypothesis Weights (0.75, 0.25, 0, 0), (0.50,0.25,0.125,0.125), (1,0,0,0)

Fixed-Sequence, Parallel and Multi-Sequence Gatekeeping: (Table1)

i. As we know that these procedures do not consider weights, so, the interpretation of these methods is done irrespective of assigned weights. From the above table H_{co} from table1 represents the Overall Power of the Co-Primary end points which is calculated as, rejecting all hypotheses provided all the null hypotheses are false.

ii. So, in these methods we basically compare the overall power, which is in this case 75.5% and 80% respectively for fixed sequence method, 67% and 73% respectively for Parallel gatekeeping procedure and multi-sequence gatekeeping procedure for the sample sizes of 500 and 542, which tells us that all the procedures are performing well with more than 70% co-primary power when the sample size is 542, it indicates a high likelihood of correctly rejecting the null hypotheses when they are false.

Fallback: (Table2)

i. The fallback procedure also demonstrates a moderate power, indicating its flexibility in adjusting the significance levels based on the rejection status of previous hypotheses. If we consider the first set of weights, we can see that the weight is totally onto first hypothesis, which indicates the importance of the particular hypothesis, so we consider the power of that individual hypothesis here but not the overall co-primary power. Therefore, the power here is 99%, which states that it signifies an extremely high probability of correctly rejecting the null hypotheses when they are actually false. Since the weights of the remaining hypotheses are zero, we can see the reduction in individual power of the hypothesis.

ii. For the second set of weights (0.75,0.25,0,0), we can see that 75% of the weight is on first hypothesis and 25% on second hypothesis. Therefore, the interpretation of power should focus on the first and second hypotheses only. Since there are no weights assigned to the third and fourth hypotheses, they do not contribute to the overall power calculation.

The first hypothesis has more than 95% chance of being correctly detected if the alternative hypothesis is true for both the sample sizes.

The second hypothesis has more than 84% chance of being correctly detected if the alternative hypothesis is true for both the sample sizes.

iii. For the third set of weights, the weights are almost equally shared among the hypotheses. Therefore, we consider the overall co-primary power of this set of weights which is 69.2% and 74.7% respectively.

Chain-Based: (Table2)

i. The chain-based graphical procedure exhibits the highest power, suggesting that it effectively utilizes the assigned weights to reject the null hypotheses.

As we have discussed above, the first set of weights gives the importance to first hypothesis; therefore, the power of the individual hypothesis is more than 99% for both the sample sizes indicating that it has high probability of correctly rejecting the null hypotheses when they are actually false.

ii. If we consider the second set of weights the individual powers of both the hypotheses are more than 95% and 84% respectively for both the sample sizes, it indicates a greater percentage of chance for correctly rejecting the null hypothesis when it is false.

iii. If we consider the last set of weights, where the weights are almost equally shared, in this case we have the powers 75% and 80% respectively, indicating a good overall co-primary power for both the sample sizes.

Note: We can see that the procedure has a higher overall power when the sample size is larger (542) compared to when it is smaller (500). This suggests that increasing the sample size improves the procedure's ability to detect significant effects across all hypotheses.

4.3.3 Overall Interpretation:

The choice of hierarchical multiple testing procedure should consider not only the sample size but also the assigned weights to the hypotheses.

i. The fixed-sequence procedure is suitable when there is a strong belief in the order of hypotheses and when the first hypothesis is of primary interest.

ii. The fallback procedure offers flexibility in adjusting significance levels based on the rejection status of previous hypotheses, making it suitable for scenarios where the order of hypotheses is less clear or when there are equal weights assigned to all hypotheses.

iii. The chain-based graphical procedure provides a comprehensive approach that considers both the hierarchy of hypotheses and the assigned weights, making it robust across different scenarios.

iv. The parallel gatekeeping procedure approach is beneficial when hypotheses within each family are of equal importance and can be tested independently of one another. It provides a straightforward method for controlling the overall Type I error rate while testing multiple hypotheses concurrently.

v. The Multi-sequence gatekeeping procedure is useful when hypotheses are logically ordered across families, and the outcome of tests in one family influences the testing strategy in subsequent families. It offers a flexible approach for addressing complex hypothesis structures and dependencies among hypotheses.

5. Conclusion

In this scenario, our objective was to conduct a comprehensive power comparison study of hierarchical multiple testing procedures utilizing simulated data. We sought to assess how different methods performed by systematically adjusting both the sample sizes and the weights assigned to each test. This allowed us to determine the relative effectiveness of each method across a range of scenarios. The findings of this study yield crucial insights into the efficacy and constraints of hierarchical methods within multiple testing contexts.

The interpretation of our results highlights the significance of accounting for both sample size and hypothesis weights when selecting an appropriate hierarchical multiple testing procedure. Each method exhibits distinct strengths and limitations, underlying the necessity for an important choice based on the specific attributes of the study and the hypotheses under investigation.

Moreover, our analysis reveals a noteworthy trend: an increase in sample size corresponds to an increase in test power. This observation underscores the crucial role of sample size determination in optimizing the statistical power of multiple testing procedures. Researchers should carefully consider sample size allocation to ensure adequate power for detecting meaningful effects while maintaining efficiency and resource utilization.

Furthermore, our study emphasizes the dynamic interplay between sample size, hypothesis weights, and method performance. By systematically exploring various combinations of these factors, we gain understanding of how different hierarchical methods respond to different experimental conditions. Such insights can inform researchers decision-making processes, guiding the selection of the most suitable approach for their specific research objectives and constraints.

In summary, our power comparison study contributes valuable insights to the field of hierarchical multiple testing, emphasizing the importance of methodological considerations and empirical evaluations in statistical practice. By clarifying how methods work in different situations, we help researchers choose wisely, making their statistical analyses more dependable and accurate.

6. References

1. Durkalski, Valerie. (2006). Analysis of Clinical Trials Using SAS: A Practical Guide. Alex Dmitrienko, Geert Molenberghs, Christy Chuang-Stein, and Walter Offen. Journal of the American Statistical Association. 101. 858-858. 10.2307/27590761
2. Dressler, E. (2019). Clinical Trial Optimization Using R.: Alex Dmitrienko and Erik Pulkstenis. Boca Raton, FL: Chapman &amp; Hall/CRC Press, 2019, 319 pp., ISBN: 9780367261252. *The American Statistician*, *73*(2), 210–211. https://doi.org/10.1080/00031305.2019.1603479
3. Guidance, D. (2017). Multiple Endpoints in Clinical Trials Guidance for Industry.
4. Dmitrienko, A., Tamhane, A. C., & Bretz, F. (2009). Multiple testing problems in pharmaceutical statistics. In Chapman and Hall/CRC eBooks. https://doi.org/10.1201/9781584889854.
5. Bretz, F., Maurer, W., Brannath, W., & Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine*, *28*(4), 586–604. https://doi.org/10.1002/sim.3495

6. Charles W. Dunnett, (1955)., "A multiple comparison procedure for comparing several treatments with a control", Journal of American Statistical Association", Vol.50, Issue 272, PP-1096-1121
7. Simes.R.J. (1986), "An improved Bonferroni Procedure for multiple tests of significance", Biometrika, Vol.73, No.3, PP.751-754