Introduction to Multiplicity in Clinical Trials

Frank Bretz, Xiaolei Xun (Novartis)

Tutorial at IMPACT Symposium III

November 20, 2014 – Cary NC

NOVARTIS

Outline

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

- Introduction
  - Type I Error Rate Inflation
  - Sources of Multiplicity
  - Dealing with Multiplicity
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusion

NOVARTIS

Type I Error Rate Inflation

Simple example with two hypotheses

- Assume that we test a single null hypothesis at significance level $\alpha = 0.05$,
  - What is the maximum Type I error rate?
- If we have two null hypotheses and do two independent tests, each at level $\alpha = 0.05$,
  - What is the probability of rejecting at least one true null hypothesis?

$$\begin{aligned}\Pr(\text{reject at least one true null}) &= 1 - \Pr(\text{reject neither true null}) \\ &= 1 - 0.95^2 \\ &= 0.0975\ (> 0.05)\end{aligned}$$

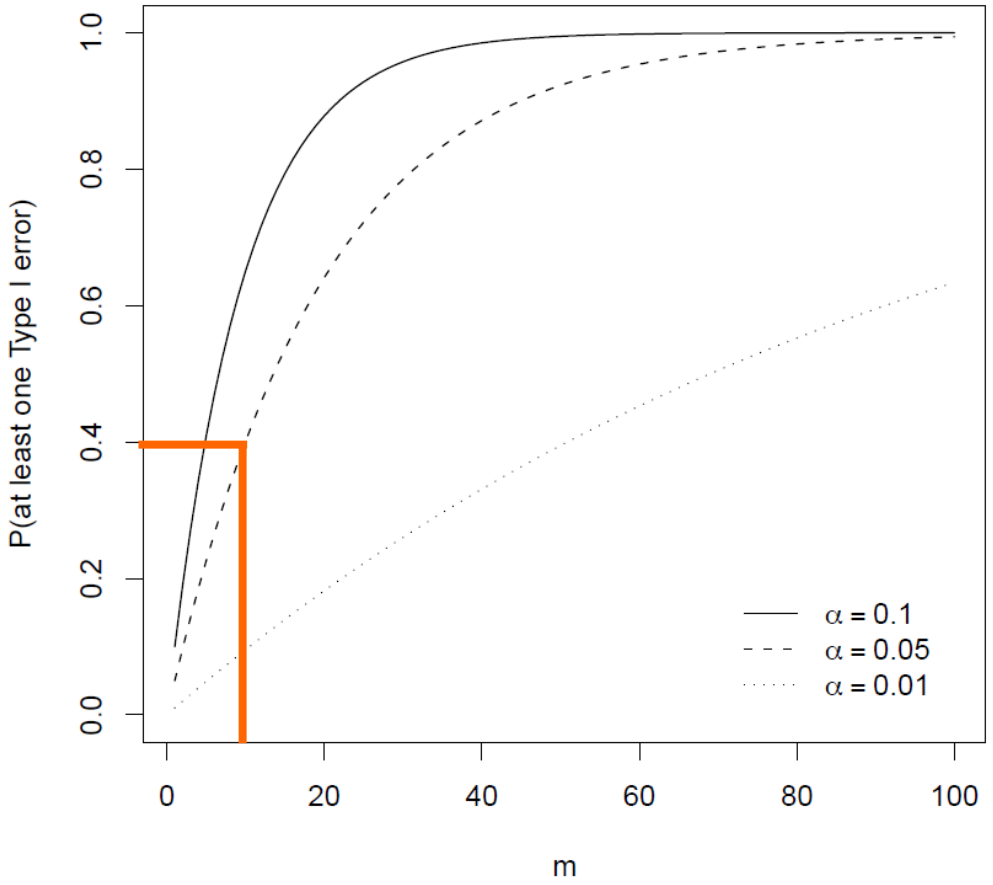
  - The Type I error rate is almost doubled
- One possible solution: Test each hypothesis at level $\alpha/2 = 0.025$ (Bonferroni test, see later). Then,

$$\Pr(\text{reject at least one true null}) = 0.0494\ (< 0.05)$$

Type I Error Rate Inflation

More than two hypotheses

Probability of at least one Type I error
for different number of hypotheses m and significance levels α



- Probability for Type I error increases with larger values of m and α
- Example:
  For $m = 10$ and $\alpha = 0.05$, the probability of at least one Type I error is 40.1%
- For large m we almost surely reject incorrectly at least one null hypothesis

NOVARTIS

Sources of Multiplicity

Overview

- Multiple test problems are very common in clinical trials
- Example applications include the comparsion 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

Dealing with Multiplicity

- Reducing the degree of multiplicity by
  - Addressing a limited number of questions only
  - Minimizing number of variables, using composite endpoints, summary statistics, ...
  - Prioritizing questions

- If multiplicity still persists
  - Multiplicity adjustment should always be considered
  - Regulatory guidance (see Appendix) requires a description of the multiplicity adjustment in Phase III study protocols
  - If not thought necessary, explain why

- Introduction
- Common Multiple Test Procedures
  - Basic concepts
  - Procedures by
    - Bonferroni, Holm
    - Simes, Hochberg
    - Dunnett, stepwise Dunnett
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

- Introduction
- Common Multiple Test Procedures
  - Basic concepts
  - Procedures by
    - Bonferroni, Holm
    - Simes, Hochberg
    - Dunnett, stepwise Dunnett
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

Basic Concepts

Notation

- Assume a "family" of m inferences
- Parameters of interest are $\theta_1, \ldots, \theta_m$
- Individual null hypotheses

$$H_1: \theta_1 = 0, \ldots, H_m: \theta_m = 0$$

- Example:
  - Comparison of m treatments with a control therapy
  - Then, $\theta_i = \mu_i - \mu_0$ are the m treatment effect differences of interest, where
    - μ_i denotes the effect for treatment $i = 1, \ldots, m$
    - μ_0 denotes the effect for the control therapy

Basic Concepts

Family-wise error rate (FWER)

- Need to extend the usual Type I error rate concept when testing a family of null hypotheses $H_1, \ldots, H_m$
- A multiple test procedure is said to control the FWER at level α (in the strong sense) if

$$\Pr(\text{reject at least one true null}) \leq \alpha$$

under any configuration of true/false null hypotheses

Basic Concepts

Adjusted p-values

- Adjusted p-values extend ordinary (i.e. unadjusted) p-values by adjusting them for a given multiple test procedure
  - Adjusted p-values can be compared directly with the significance level α, while controlling the FWER
- Formally, the adjusted p-value is the smallest significance level at which a given hypothesis is significant as part of the multiple test procedure
- Example: Bonferroni method

$$p_i \leq \alpha/m \quad \Leftrightarrow \quad q_i = \min(mp_i, 1) \leq \alpha$$

where p_i is the ordinary and q_i the adjusted p-value for $i = 1, \ldots, m$

Basic Concepts

Single step and stepwise test procedures

- Single step methods

  The rejection or non-rejection of a single hypothesis does not depend on the decision on any other hypothesis.

  Examples: Bonferroni, Simes, Dunnett, …

- Stepwise methods

  The rejection or non-rejection of a particular hypothesis may depend on the decision on other hypotheses.

  Examples: Holm, Hochberg, stepdown Dunnett, …

- Introduction
- Common Multiple Test Procedures
  - Basic concepts
  - Procedures by
    - Bonferroni, Holm
    - Simes, Hochberg
    - Dunnett, stepwise Dunnett
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

Bonferroni Method

Overview

- Use α/m for all inferences; for $i = 1, \ldots, m$:

$$\text{Reject } H_i \text{ if } p_i \leq \alpha/m$$

- Example: With $m = 3$, p-values must be less than $0.05/3 = 0.0167$ in order to be "significant"

- With adjusted p-values $q_i = \min(mp_i, 1)$,

$$\text{Reject } H_i \text{ if } q_i \leq \alpha$$

  - Note that $mp_i > 1$ is possible and we thus need to truncate the adjusted p-avlues at 1, resulting in the minimum expression

- Both rejection rules above lead to the same test decisions

Bonferroni Method

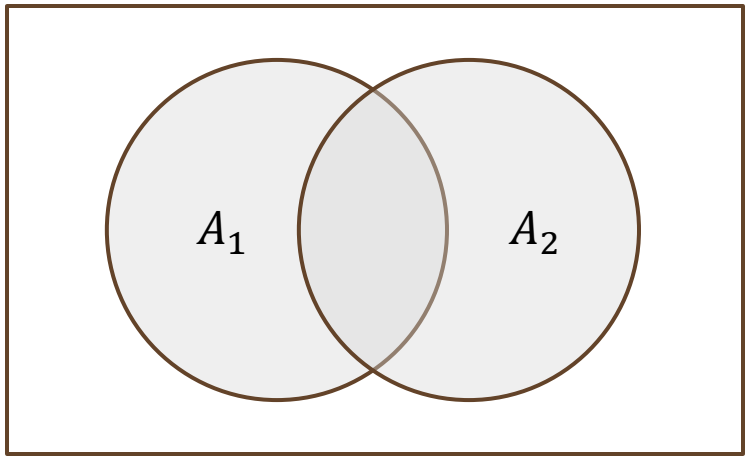
Rationale

- The Bonferroni method follows from the Boole's inequality

$$\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i)$$

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

$$\Pr(A_1 \cup A_2) \leq \Pr(A_1) + \Pr(A_2)$$



- For $m = 2$,

$$\begin{aligned}\text{FWER} &= \Pr(p_1 \leq \alpha/2 \text{ or } p_2 \leq \alpha/2 \,|H_1, H_2 \text{ are true}) \\ &\leq \Pr(p_1 \leq \alpha/2 \,|H_1 \text{is true}) + \Pr(p_2 \leq \alpha/2 \,|H_2 \text{is true}) \\ &= 2\alpha/2 = \alpha\end{aligned}$$

NOVARTIS

Bonferroni Method

Properties

- The Bonferroni method is a single step procedure
- It 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 Procedure

Simplistic explanation

- Assume p-values 0.0121, 0.0142, 0.0191, 0.1986
- Applying Bonferroni, we use $0.05/4 = 0.0125$ and reject H_1
- However, having rejected H_1 using $0.05/4$, you no longer believe that all four null hypotheses can be true
- You now think only H_2, H_3, H_4 can be true
- So, test H_2 using $0.05/3 = 0.0167$, rather than $0.05/4$

Holm Procedure

Overview

- Let $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered unadjusted p-values with associated null hypotheses $H_{(1)}, \ldots, H_{(m)}$
- Then we have the following stepwise procedure:
  - If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$ and continue; else stop
  - If $p_{(2)} \leq \alpha/(m-1)$, reject $H_{(2)}$ and continue; else stop
  - ...
  - If $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(i)}$ and continue; else stop
  - ...
  - If $p_{(m)} \leq \alpha$, reject $H_{(m)}$

Holm Procedure

Properties

- The Holm procedure is a stepwise procedure that is more powerful than the Bonferroni method
  - Bonferroni uses the same threshold α/m for all hypotheses
  - Holm uses the larger thresholds $\alpha/(m-i+1)$
- Sometimes called “stepdown Bonferroni” procedure
- The Holm procedure can be improved by accounting for correlations (e.g. stepdown Dunnett test; see later)

Holm Procedure

Adjusted p-Values

- With $p_{(1)} \leq \cdots \leq p_{(m)}$, define adjusted p-values using
  - $\tilde{q}_{(1)} = mp_{(1)}$
  - $\tilde{q}_{(2)} = \begin{cases} (m-1)p_{(2)}, & \text{if } (m-1)p_{(2)} > q_{(1)} \\ q_{(1)}, & \text{otherwise} \end{cases}$
  - ...
  - $\tilde{q}_{(m)} = \begin{cases} p_{(m)}, & \text{if } p_{(m)} > q_{(m-1)} \\ q_{(m-1)}, & \text{otherwise} \end{cases}$
- Formula for adjusted p-values:

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

- Introduction
- Common Multiple Test Procedures
  - Basic concepts
  - Procedures by
    - Bonferroni, Holm
    - Simes, Hochberg
    - Dunnett, stepwise Dunnett
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

Simes Method

Overview

- The Simes method tests the global null hypothesis

$$H = H_1 \cap H_2 \cap \cdots \cap H_m : \theta_1 = \theta_2 = \cdots = \theta_m = 0$$

- It uses all ordered p-values $p_{(1)}, \ldots, p_{(m)}$, not just $p_{(1)}$

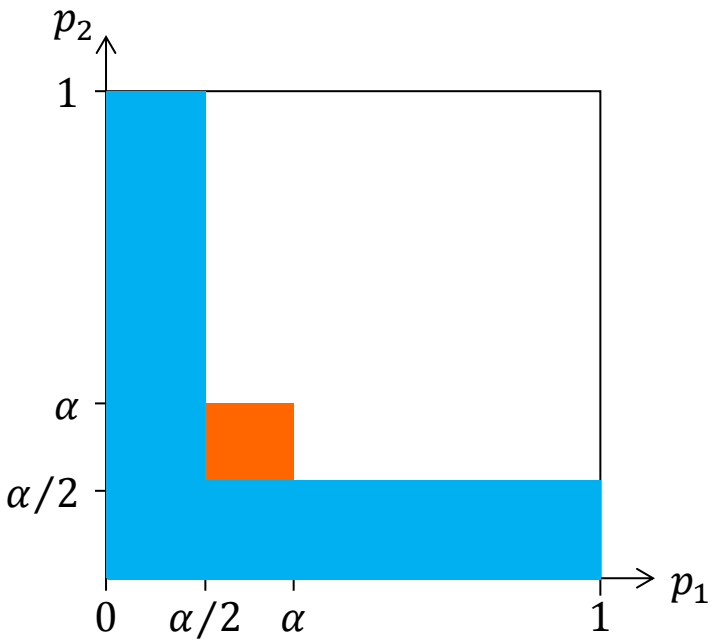
Reject H if $p_{(i)} \leq i\alpha/m$ for at least one i

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

Simes Method

Comparison with Bonferroni method (for $m = 2$)

- Bonferroni rejects H, if $p_{(1)} \leq \alpha/2$
- Simes rejects H, if $p_{(1)} \leq \alpha/2$ or $p_{(2)} \leq \alpha$
- Under independence of p_1 and p_2,
  - $\Pr(\text{Bonferroni rejects}) = 1 - (1 - \alpha/2)^2 = \alpha - (\alpha/2)^2 < \alpha$
  - $\Pr(\text{Simes rejects}) = 1 - (1 - \alpha/2)^2 + (\alpha/2)^2 = \alpha$



- Simes is more powerful than a global test based on Bonferroni
- Simes assumes non-negative correlations between p-values, Bonferroni does not

Hochberg Procedure

Overview

- The Hochberg procedure is a stepwise version of the Simes method, using the same thresholds as Holm:
  - If $p_{(m)} \leq \alpha$, reject $H_{(1)}, \ldots, H_{(m)}$ and stop; else continue
  - If $p_{(m-1)} \leq \alpha/2$, reject $H_{(1)}, \ldots, H_{(m-1)}$ and stop; else continue
  - ...
  - If $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(1)}, \ldots, H_{(i)}$ and stop; else continue
  - ...
  - If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$

- Adjusted p-values are

$$q_{(m)} = p_{(m)}$$
$$q_{(i)} = \min\left[(m-i+1)p_{(i)}, q_{(i+1)}\right], \text{ for } i = m-1, \ldots, 1$$

Hochberg Procedure

Properties

- The Hochberg procedure is sometimes called “stepup Simes” procedure
- It is more powerful than the 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
  - For example, Hommel procedure based on the closed test procedure (see later)

- Introduction
- Common Multiple Test Procedures
  - Basic concepts
  - Procedures by
    - Bonferroni, Holm
    - Simes, Hochberg
    - Dunnett, stepwise Dunnett
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

Dunnett Test

Comparing several treatments with a control

- 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

Dunnett Test

Linear model and hypotheses

- Consider the unbalanced one-way layout

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

where

- Y_{ij} denotes observation $j = 1, \dots, n_i$ in group $i = 0,1, \dots, m$
- μ_i the effect of treatment group i
- ε_{ij} are independent and identically normally distributed with mean 0 and variance σ^2, i.e. $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$

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

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

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

NOVARTIS

Dunnett test

Individual test statistics

- Consider the m pairwise t-tests

$$t_i = \frac{\hat{\mu}_i - \hat{\mu}_0}{\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_0}}}, \qquad i = 1, \ldots, m$$

  where $\hat{\mu}_i$ and $\hat{\sigma}$ are the ordinary least squares of μ_i and σ, respectively

- Note that $t_i \sim t_\nu$ under H_i, where t_ν denotes the univariate t-distribution with $\nu = \sum_i n_i - m - 1$ degrees of freedom

- Furthermore, $(t_1, \ldots, t_m)$ follows the m-variate t-distribution with ν degrees of freedom and correlations

$$\rho_{ij} = \sqrt{\frac{n_i}{n_i+n_0}}, \sqrt{\frac{n_j}{n_j+n_0}}, \qquad i, j = 1, \ldots, m$$

Dunnett test

Rejection rule

- For the m individual null hypotheses,

$$\text{Reject } H_i \text{ if } t_i \geq c_{m,1-\alpha}$$

- The quantile $c_{m,1-\alpha}$ is computed such that

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

  where $(t_1, \ldots, t_m)$ follows the m-variate t-distribution with ν degrees of freedom and correlations ρ_{ij}, for $i, j = 1, \ldots, m$

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

Dunnett test

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 (not in this tutorial)
- 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

Overview

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

- Then we have the following stepwise procedure:
  - If $t_{(1)} \geq c_{m,1-\alpha}$, reject $H_{(1)}$ and continue; else stop
  - If $t_{(2)} \geq c_{m-1,1-\alpha}$, reject $H_{(2)}$ and continue; else stop
  - ...
  - If $t_{(i)} \geq c_{m-i+1,1-\alpha}$, reject $H_{(i)}$ and continue; else stop
  - ...
  - If $t_{(m)} \geq c_{1,1-\alpha}$, reject $H_{(m)}$

  where $c_{m-i+1,1-\alpha}$ denotes the $1-\alpha$ quantile of the distribution of the maximum of $m-i+1$ t-distributed random variables and is computed from the corresponding multivariate t-distribution

Stepwise Dunnett test

Properties

- For the stepwise Dunnett test, the quantiles change as hypotheses are rejected
  - For example, if $H_{(1)}$ is rejected, then the quantile $c_{m-1,1-\alpha}$ is computed from a $(m-1)$-variate t-distribution
- The stepwise Dunnett test is better than the single step Dunnett test
  - It can be shown that $c_{m,1-\alpha} \geq c_{m-1,1-\alpha} \geq \cdots \geq c_{1,1-\alpha}$, where $c_{1,1-\alpha} = t_{\nu,1-\alpha}$ is the quantile from the univariate t-distribution with ν degrees of freedom
  - The Dunnett test uses $c_{m,1-\alpha}$ for all comparisons
- 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, similar to Hochberg (not in this tutorial)

Summary

	Correlations		
	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

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
  - Fixed Sequence Procedure
  - Fallback Procedure
  - Numerical Example
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

NOVARTIS

COPD Example

Background

- Double-blind, parallel-group study to show that drug B is better than drug A in patients with chronic obstructive pulmonary disease (COPD)
- Primary endpoint: FEV1 (forced expiratory volume in one second)
  - Continuous variable, where larger values indicate better efficacy
- Secondary endpoint: Time to exacerbation
  - Time until the event is of interest has been observed

COPD Example

Background (continued)

- There are two hypotheses corresponding to the two endpoints, thus a multiple test procedure is needed
- All of the previous multiple tests could be applied, but do not reflect the relative importance of the two endpoints
  - For example, the Bonferroni test would treat FEV1 and time-to-exacerbation as equally important
- Note that the previous stepwise procedures (Holm, Hochberg, ...) use a data-driven order of hypotheses
  - Here we need a multiple test procedure that specifies the order of the hypotheses based on clinical importance (and not based on data)

Hierarchical Test Procedures

Overview

- If the hierarchy of hypotheses is specified before data is observed, one can apply a hierarchical test procedure
- Two hierarchical test procedures will be introduced
  - Fixed sequence procedure
  - Fallback procedure

Hierarchical Test Procedures

Fixed sequence procedure – General description

- Fixed sequence procedures test hierarchically ordered hypotheses in sequence at level α until first non-rejection
- Assume m hierarchically ordered hypotheses

$$H_1 \rightarrow H_2 \rightarrow \cdots \rightarrow H_m$$

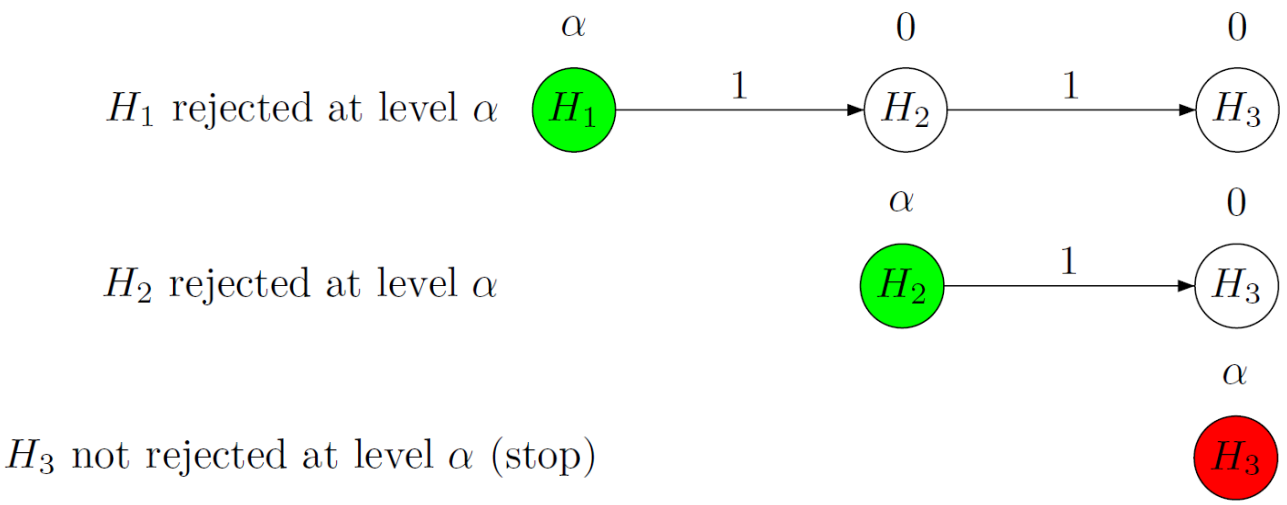
with unadjusted p-values $p_1, p_2, \ldots, p_m$

- We have the following fixed sequence procedure:
  - If $p_1 \leq \alpha$, reject H_1 and continue; else stop
  - If $p_2 \leq \alpha$, reject H_2 and continue; else stop
  - ...
  - If $p_i \leq \alpha$, reject H_i and continue; else stop
  - ...
  - If $p_m \leq \alpha$, reject H_m

Hierarchical Test Procedures

Fixed sequence procedure – Example with $m = 3$ hypotheses

- Assume $H_1 \rightarrow H_2 \rightarrow H_3$
  - That is, H_1 is more important than H_2, and H_2 is more important than H_3
- We have the following fixed sequence procedure for example:



Note: Green = rejection; red = no rejection (and stop)

NOVARTIS

Hierarchical Test Procedures

Fixed sequence procedure – Properties

- Adjusted p-values are given by

$$q_i = \max\{p_1, \ldots, p_i\}, \qquad i = 1, \ldots, m$$

- Advantages
  - Simple procedure, each test is performed in sequence at level α
  - It is optimal when hypotheses early in the sequence are associated with large effects and performs poorly otherwise
- Disadvantages
  - Once a hypothesis is not rejected, no further testing is permitted
- Great care is advised when specifying the sequence of hypotheses

Hierarchical Test Procedures

Fallback procedure – General description

- Fallback procedures test hierarchically ordered hypotheses in sequence as the fixed sequence procedure, but splits the level α between the hypotheses

- Assume m hierarchically ordered hypotheses

$$H_1 \rightarrow H_2 \rightarrow \cdots \rightarrow H_m$$

with unadjusted p-values $p_1, \ldots, p_m$ and $\alpha = \alpha_1 + \cdots + \alpha_m$

- Then the fallback procedure tests H_i at level α_i', where for $i = 2, \ldots, m$

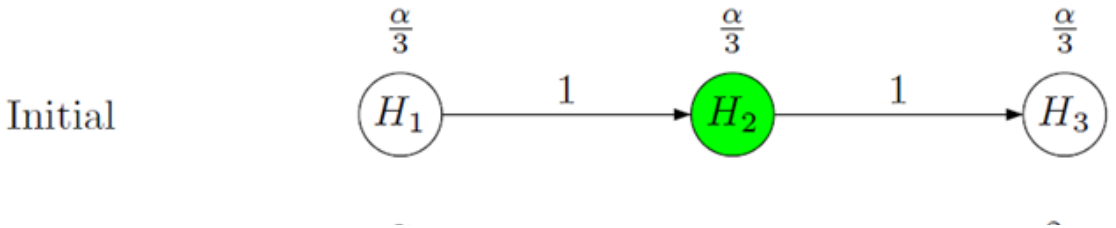
$$\alpha_i' = \begin{cases} \alpha_i, & \text{if } H_{i-1} \text{ is not rejected} \\ \alpha_i + \alpha_{i-1}', & \text{otherwise} \end{cases}$$

and $\alpha_1' = \alpha_1$

Hierarchical Test Procedures

Fallback procedure – Example with $m = 3$ hypotheses

- Assume $H_1 \rightarrow H_2 \rightarrow H_3$, and split the significance level as $\alpha_1 = \alpha_2 = \alpha_3 = \alpha/3$
- Following the fallback procedure, we could have for example:



Note: Green = rejection; red = no rejection (and stop)

NOVARTIS

Hierarchical Test Procedures

Fallback procedure – Properties

- The fixed sequence procedure is obtained as special case from the fallback procedure by setting $\alpha_1 = \alpha$ and $\alpha_i = 0$ for $i > 1$

- In contrast to the fixed sequence procedure, the fallback procedure tests all hypotheses in the pre-specified sequence even if the initial hypotheses are not rejected

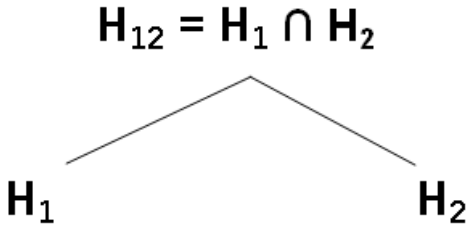
- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

NOVARTIS

Closed Test Procedure (CTP)

Operational definition for $m = 2$ null hypotheses

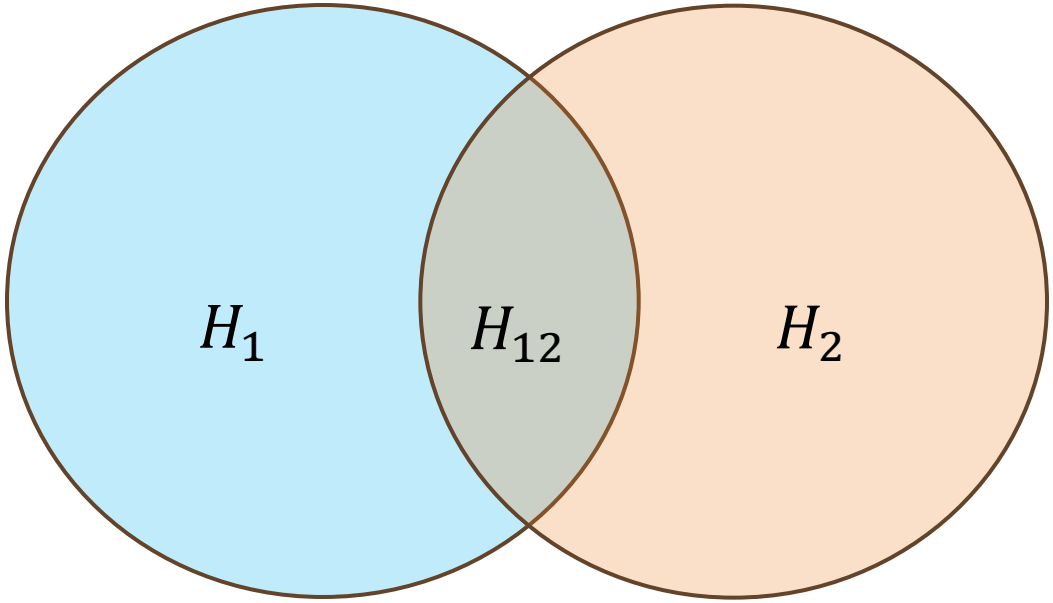
- Schematic diagram for $m = 2$ null hypotheses H_1, H_2



- Rejection rule: Reject H_1 (H_2) while controlling the FWER at α, if H_1 (H_2) and H_{12} are rejected, each at local level α
- Operationally
  - Test H_{12} at local level α (using a suitable test): If rejected, proceed; otherwise stop
  - Test H_1 and H_2 each at local level α: Reject H_1 (H_2) overall if H_{12} and H_1 (H_2) are rejected locally

Closed Test Procedure

Venn-type diagram for $m = 2$ null hypotheses

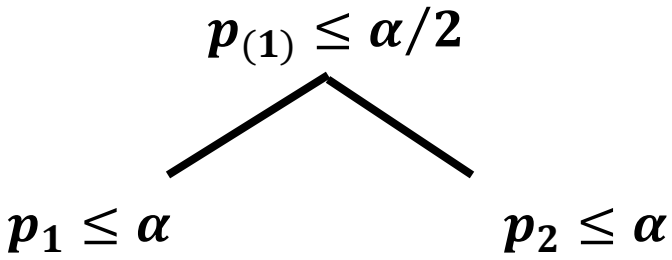


- Different parts indicate different null hypotheses as shown above
- Question: How do we test them?
  - Test H_{12} using Bonferroni, Simes, Dunnett, etc. at level α
  - Test H_1, H_2 each using a level α test

CTP Using Bonferroni

Holm procedure

- Using Bonferroni to test H_{12}, reject if $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$, i.e., if $p_{(1)} \leq \alpha/2$



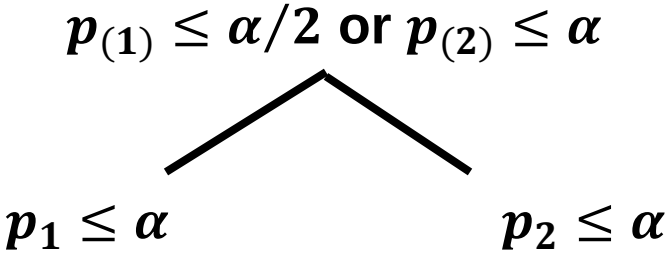
- If we fail to reject H_{12}, stop as neither H_1 or H_2 can be rejected according to the CTP
- If we reject H_{12}, then
  - $H_{(1)}$ is rejected automatically as $p_{(1)} \leq \alpha/2 < \alpha$
  - we only need to test $H_{(2)}$ at level α, i.e., reject $H_{(2)}$ if $p_{(2)} \leq \alpha$
- This results exactly in the Holm procedure

NOVARTIS

CTP Using Simes

Hochberg procedure

- Using Simes to test H_{12}, reject if $p_{(1)} \leq \alpha/2$ or $p_{(2)} \leq \alpha$
- If we fail to reject H_{12}, stop
- If we reject H_{12} because $p_{(2)} \leq \alpha$, then $H_{(1)}, H_{(2)}$ are rejected automatically as $p_{(1)} \leq p_{(2)} \leq \alpha$, and stop



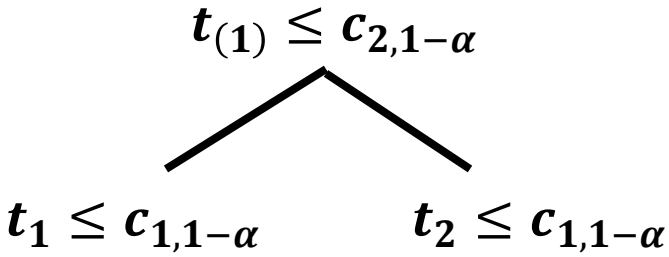
- If we reject H_{12} because $p_{(1)} \leq \alpha/2$ but $p_{(2)} > \alpha$, we then reject $H_{(1)}$ but fail to reject $H_{(2)}$ and stop
- This results exactly in the Hochberg procedure for $m = 2$
  - For $m > 2$ the Hochberg procedure is less powerful the CTP using Simes tests (Hommel procedure)

NOVARTIS

CTP Using Dunnett

Stepwise Dunnett test

- Using Dunnett test to test H_{12}, reject if $t_1 \leq c_{2,1-\alpha}$ or $t_2 \leq c_{2,1-\alpha}$, i.e., if $t_{(1)} \leq c_{2,1-\alpha}$



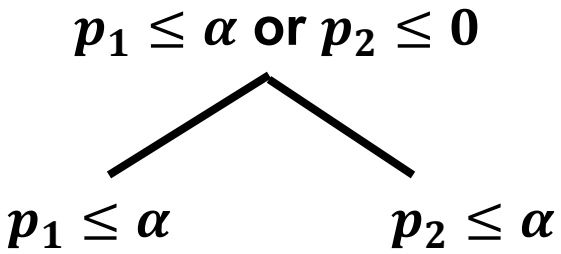
- If we fail to reject H_{12}, stop
- If we reject H_{12}, then
  - $H_{(1)}$ is rejected automatically as $t_{(1)} \leq c_{2,1-\alpha} \leq c_{1,1-\alpha}$
  - we only need to test $H_{(2)}$ at level α, i.e., reject $H_{(2)}$ if $t_{(2)} \leq c_{1,1-\alpha}$
- This results exactly in the stepdown Dunnett procedure

NOVARTIS

CTP Using Weighted Bonferroni (1)

Fixed sequence procedure

- Two ordered hypothese $H_1 \rightarrow H_2$
- Using weighted Bonferroni test to test H_{12}, reject if $p_1 \leq \alpha$ or $p_2 \leq 0$
- If we fail to reject H_{12}, stop
- If we reject H_{12}, then
  - H_1 is rejected automatically as $p_1 \leq \alpha$
  - we only need to test H_2 at level α, i.e., reject H_2 if $p_2 \leq \alpha$
- This results exactly in the fixed sequence procedure

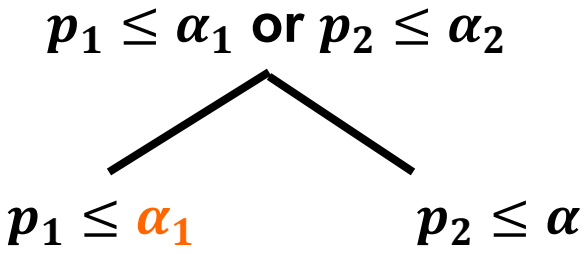


NOVARTIS

CTP Using Weighted Bonferroni (2)

Fallback procedure

- Two ordered hypothese $H_1 \rightarrow H_2$
- Using weighted Bonferroni test to test H_{12}, reject if $p_1 \leq \alpha_1$ or $p_2 \leq \alpha_2$
  - Weights α_1 and α_2 are such that $\alpha_1 + \alpha_2 = \alpha$

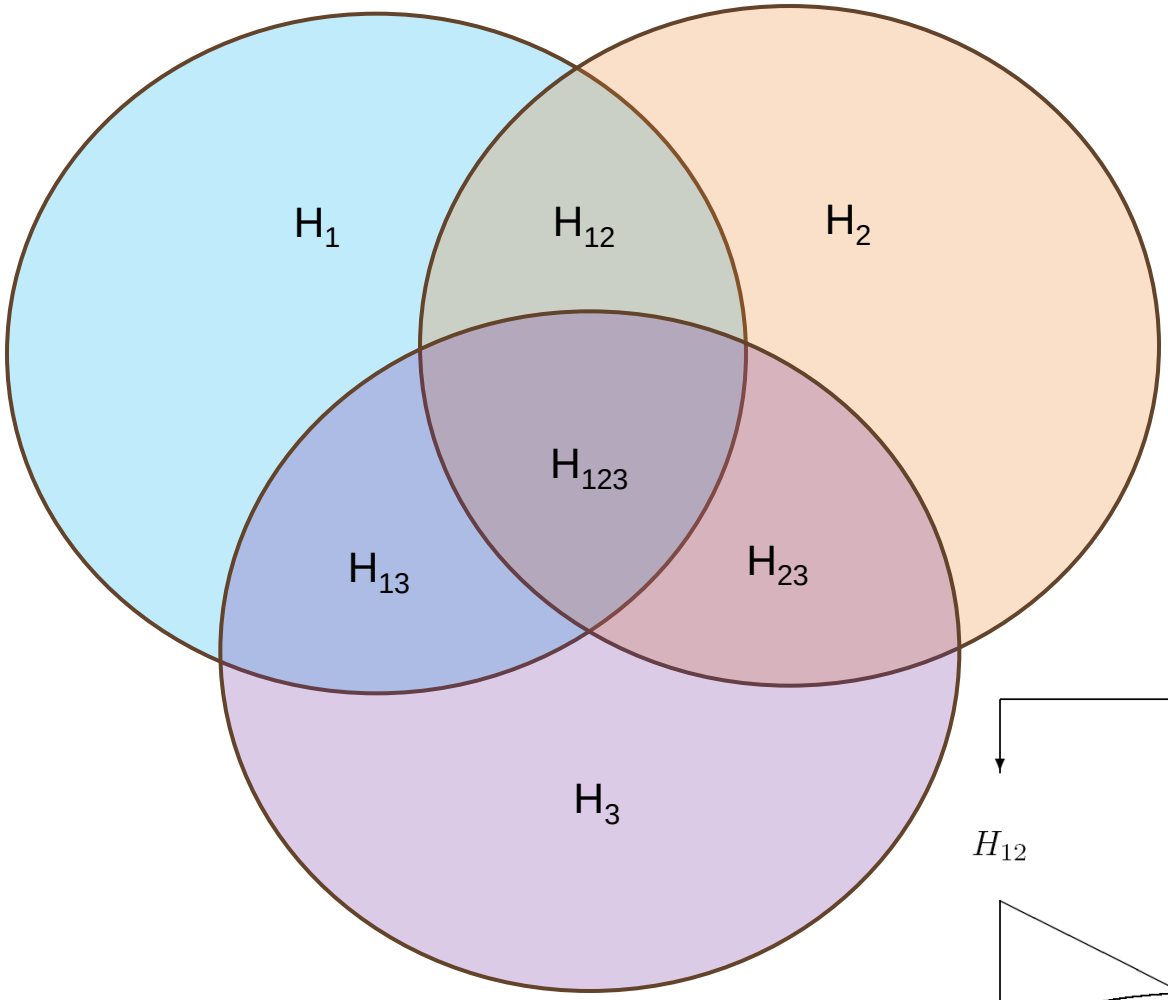


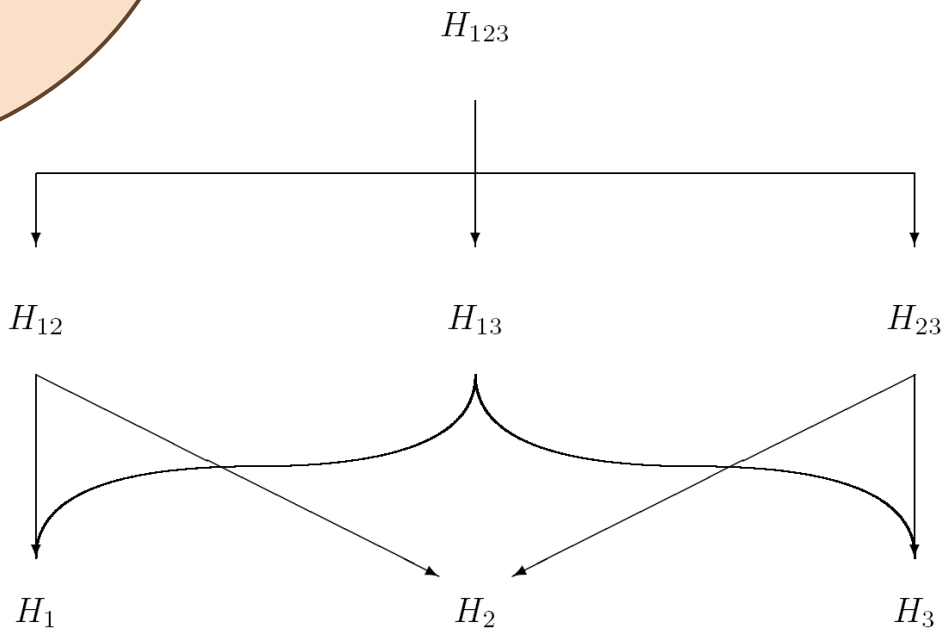
- If we fail to reject H_{12}, stop
- If we reject H_{12}, then we test H_2 at level α, i.e., reject H_2 if $p_2 \leq \alpha$
  - H_1 is tested at α_1 level instead of α
- This results exactly in the fallback procedure

NOVARTIS

Closed Test Procedure

Venn-type diagram for $m = 3$ null hypotheses





Closed Test Procedure

Formal definition for m null hypotheses

- For $m > 2$ many intersection hypotheses have to be tested
- CTP considers all intersection hypotheses

$$H_J = \bigcap_{i \in J} H_i \,, \qquad J \subseteq \{1, \ldots, m\}$$

  - Any suitable test can be used to test H_J at local level α
- An individual H_i is rejected at level α if all hypotheses H_J formed by intersection with H_i are rejected at local level α

Summary

- CTP is a general principle to construct powerful multiple test procedures
- In a CTP, one rejects an individual null hypothesis H_i at overall level α by rejecting all intersection null hypotheses $H_J \subseteq H_i$, including $J = \{i\}$
- Many common multiple test procedures are CTP, including
  - Holm, Hochberg, step-down Dunnett, ...
- CTPs satisfy certain optimality criteria and there is no reason why not to use a CTP
- The number of intersection hypotheses is $2^m - 1$
  - For large m, this number increases rapidly and CTPs are in general difficult to apply

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example extended
- Summary and Conclusions

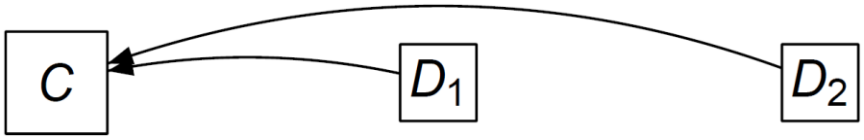
- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example extended
- Summary and Conclusions

COPD Example extended

Multiple endpoints and multiple doses

- Objective: Show that a new drug is better than a control drug in patients with COPD for two endpoints
  - Primary endpoint: FEV1 (forced expiratory volume in one second)
    - Continuous variable, where larger values indicate better efficacy
  - Secondary endpoint: Time to exacerbation
    - Time until the event of interest has been observed

- New drug is available at two doses D_1, D_2 that are compared with the control C

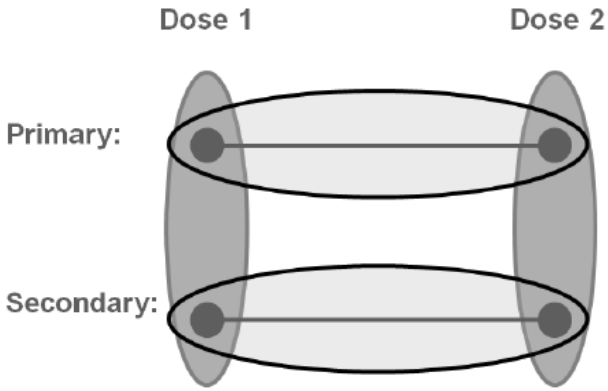


NOVARTIS

COPD Example extended

Multiple endpoints and multiple doses

- Two sources of multiplicity
  - Comparing two doses with control for each of two endpoints
- Resulting in four hypotheses of interest
  - Two primary hypotheses H_1, H_2 (comparing D_1, D_2 with C for FEV1)
  - Two secondary hypotheses H_3, H_4 (comparing D_1, D_2 with C for time to exacerbation)
- Note that the four hypotheses are not equally important
  - The secondary hypotheses H_3 (H_4) should be tested, only if the corresponding primary hypotheses H_1 (H_2) is rejected



- Need for suitable multiple test procedures

NOVARTIS

Graphical Approach

Heuristics

- As before,
  - Null hypotheses $H_1, \ldots, H_m$
  - Initial allocation of the significance level $\alpha_1 + \cdots + \alpha_m = \alpha$
  - Unadjusted p-values $p_1, \ldots, p_m$
- α–propagation

  If a hypothesis H_i can be rejected at level α_i (i.e. $p_i \leq \alpha_i$), propagate its level α_i to the remaining, not yet rejected hypotheses (according to a prefixed rule) and continue testing with the updated α levels

Graphical Approach

Conventions

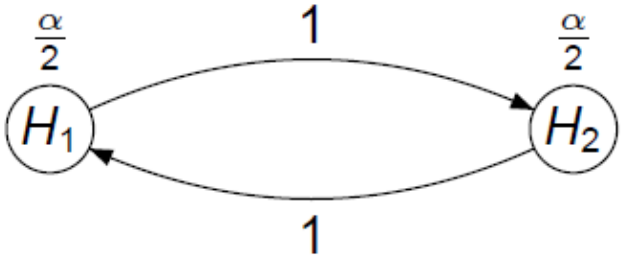
1. Hypotheses $H_1, \ldots, H_m$ represented as nodes

2. Split of significance level α as weights $\alpha_1, \ldots, \alpha_m$

Bonferroni

3. "α propagation" through weighted, directed edges

Holm



NOVARTIS

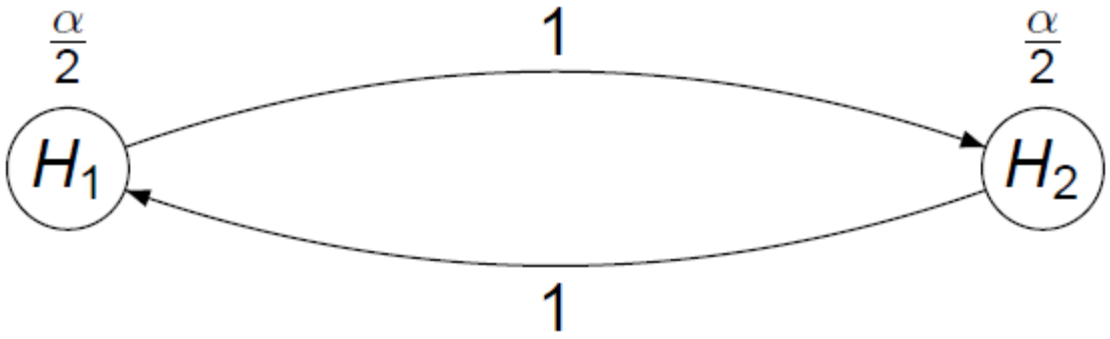
- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example extended
- Summary and Conclusions

Graphical Approach

Bonferroni test and Holm procedure: m=2

- Bonferroni: no α–propagation, i.e. no edges between nodes

- Holm: includes α–propagation and is thus more powerful



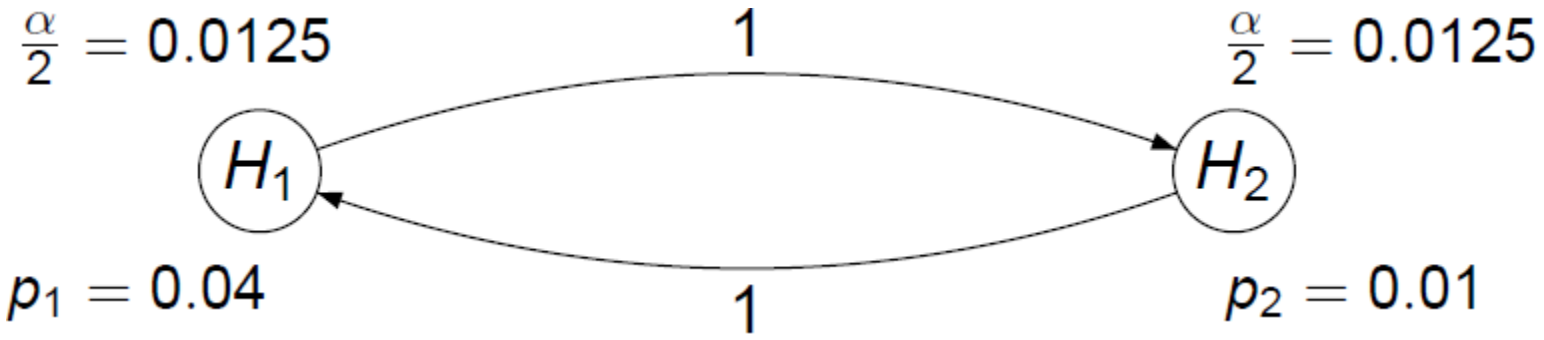
NOVARTIS

Graphical Approach

Holm procedure: Example with $\alpha = 0.025$

Test H_1 at level $\alpha/2$

Test H_2 at level $\alpha/2$

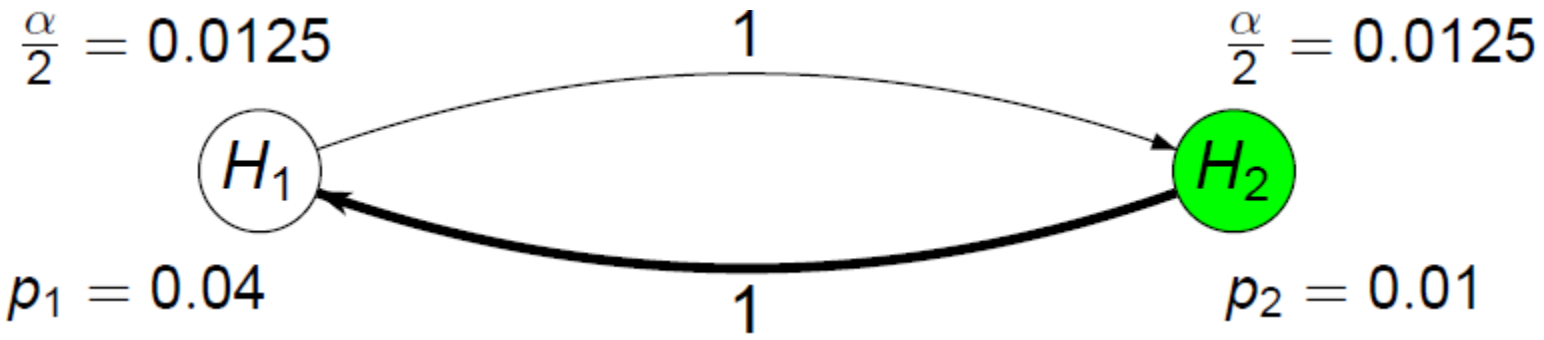


NOVARTIS

Graphical Approach

Holm procedure: Example with $\alpha = 0.025$

$p_2 < \alpha/2 \Longrightarrow$ reject H_2

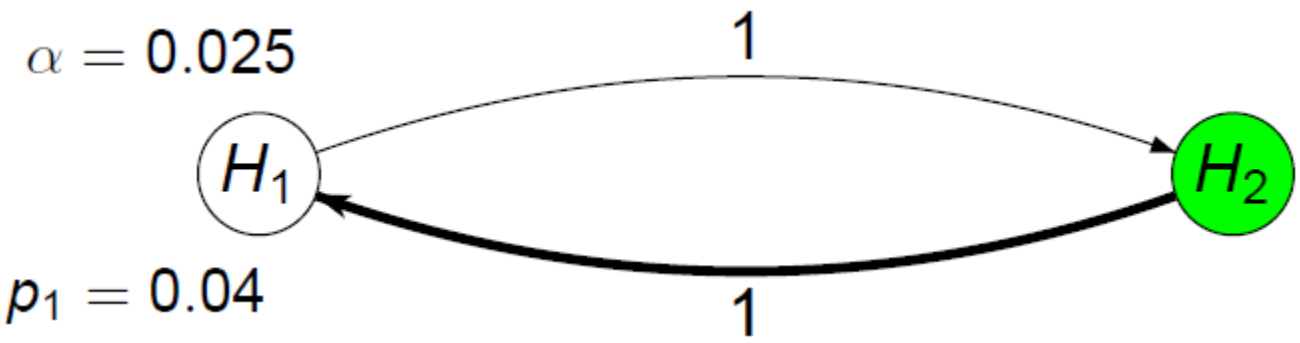


NOVARTIS

Graphical Approach

Holm procedure: Example with $\alpha = 0.025$

Propagate $\alpha/2$



NOVARTIS

Graphical Approach

Holm procedure: Example with $\alpha = 0.025$

Remove node for H_2

$\alpha = 0.025$

H_1

$p_1 = 0.04$

NOVARTIS

Graphical Approach

Holm procedure: Example with $\alpha = 0.025$

Test H_1 at level α

$p_1 > \alpha \Longrightarrow$ retain H_1 and stop

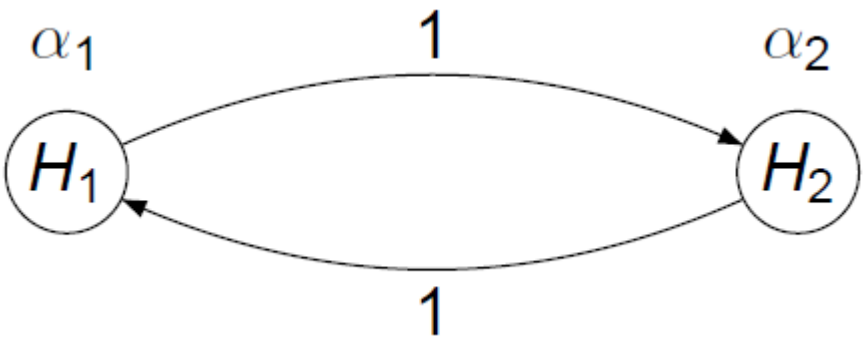
$\alpha = 0.025$

$p_1 = 0.04$

Graphical Approach

Weighted Holm procedure

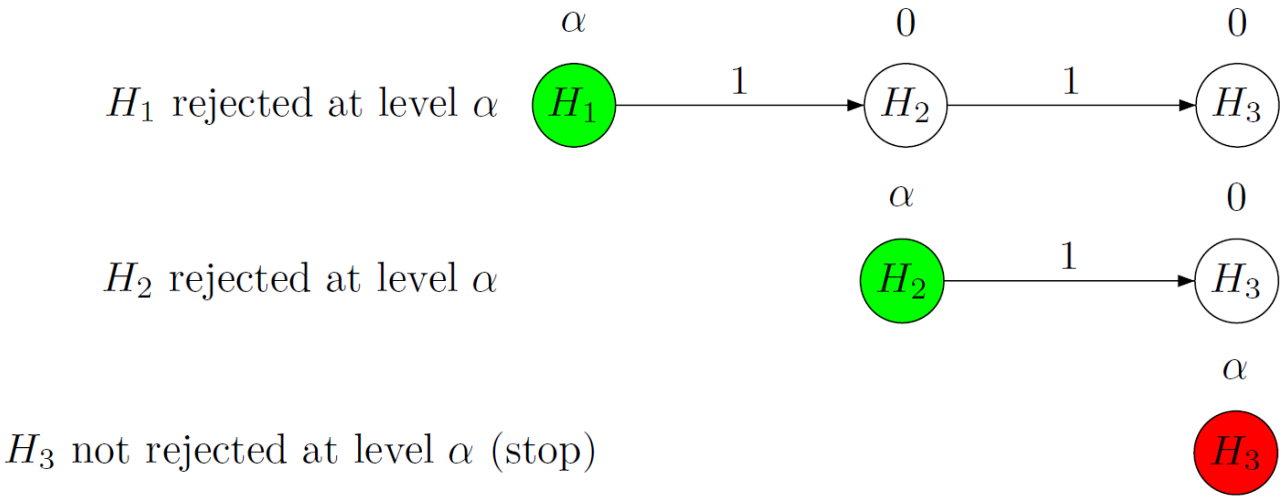
- Use α_1, α_2 with $\alpha_1 + \alpha_2 = \alpha$ instead of $\alpha_1 = \alpha_2 = \alpha/2$



Graphical Approach

Fixed sequence procedure: Example with $m = 3$ hypotheses

- Assume $H_1 \rightarrow H_2 \rightarrow H_3$
  - That is, H_1 is more important than H_2, and H_2 is more important than H_3
- Then we could have, for example, the following fixed sequence procedure:



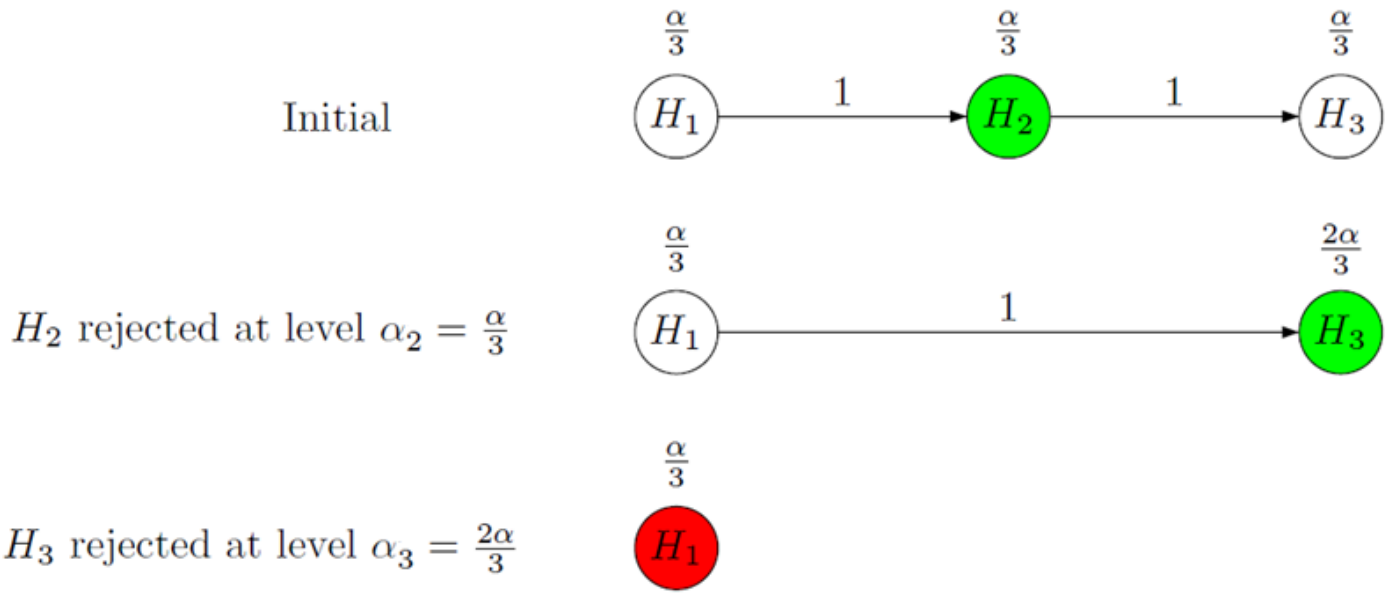
Note: Green = rejection; red = no rejection (and stop)

NOVARTIS

Graphical Approach

Fallback procedure: Example with $m = 3$ hypotheses

- Assume $H_1 \rightarrow H_2 \rightarrow H_3$, and split the significance level as $\alpha_1 = \alpha_2 = \alpha_3 = \alpha/3$
- Then we could have, for example, the following fallback procedure:



NOVARTIS

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example extended
- Summary and Conclusions

Graphical Approach

Formal definition

- Define
  - Initial levels $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ with $\sum_{i=1}^{m} \alpha_i = \alpha \in (0,1)$
  - $m \times m$ transition matrix $\boldsymbol{G} = (g_{ij})$

    where g_{ij} is the fraction of the level of H_i that is propagated to H_j with $0 \leq g_{ij} \leq 1$, $g_{ii} = 0$, and $\sum_{j=1}^{m} g_{ij} \leq 1, \ \forall i = 1, \ldots, m$

- $(\boldsymbol{G}, \boldsymbol{\alpha})$ determine a graph with an associated multiple test

Graphical Approach

Update algorithm

Set $J = \{1, \ldots, m\}$

1. Select a j such that $p_j \leq \alpha_j$

   If no such j exists, stop; otherwise reject H_j

2. Update the graph:

$$J \quad \rightarrow J \setminus \{j\}$$

$$\alpha_\ell \quad \rightarrow \begin{cases} \alpha_\ell + \alpha_j g_{j\ell}, & \ell \in J \\ 0, & \text{otherwise} \end{cases}$$

$$g_{\ell m} \rightarrow \begin{cases} \frac{g_{\ell m} + g_{\ell j} g_{jm}}{1 - g_{\ell j} g_{j\ell}}, & \ell, m \in J, \ell \neq m, g_{\ell j} g_{j\ell} < 1 \\ 0, & \text{otherwise} \end{cases}$$

3. Go to Step 1

NOVARTIS

Graphical Approach

Main result

- The initial levels $\boldsymbol{\alpha}$, the transition matrix $\boldsymbol{G}$, and the algorithm define a unique sequentially rejective test procedure that controls the FWER at level α

- Remarks:
  - Any multiple test procedure derived and visualized by a graph $(\boldsymbol{G}, \boldsymbol{\alpha})$ is based on the closed test principle
  - The graph $(\boldsymbol{G}, \boldsymbol{\alpha})$ and the algorithm define weighted Bonferroni tests for each intersection hypothesis in a CTP
  - The algorithm defines a shortcut for the resulting CTP, which does not depend on the rejection sequence

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example extended
- Summary and Conclusions

COPD Example Revisited

Background

- Recall the study objective is to demonstrate that either dose D_1 or D_2 of a new drug is better than control C in COPD patients for two endpoints:
  - Primary endpoint: FEV1
  - Secondary endpoint: Time to exacerbation
- There is a natural order in that a primary endpoint is more important than a secondary endpoint
  - Thus, we would like to test the primary null hypothesis first; only if that is rejected, we test the secondary hypothesis
- Both doses are equally important
  - Thus, both doses are simultaneously tested against the control

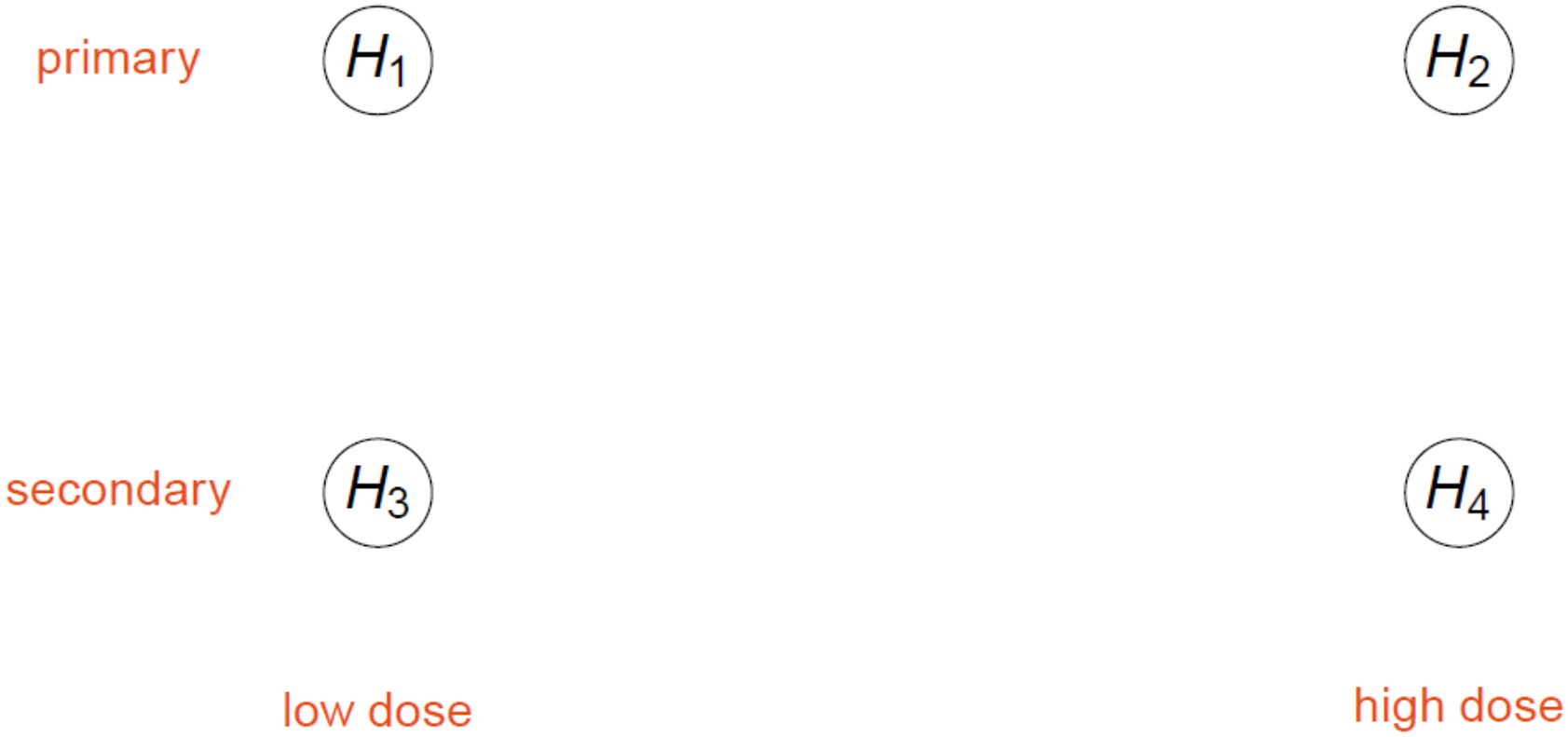
COPD Example Revisited

Background (continued)

- We have four hypotheses corresponding to the two doses and the two endpoints; a multiple test procedure is needed
- Standard multiple test procedures could be applied, but do not reflect the relative importance of the two endpoints
  - For example, the Bonferroni test would treat FEV1 and time-to-exacerbation as equally important and doesn't reflect the relative order desired
- We need a multiple test procedure that reflects the relative importance and order of the hypotheses based on clinical importance

COPD Example Revisited

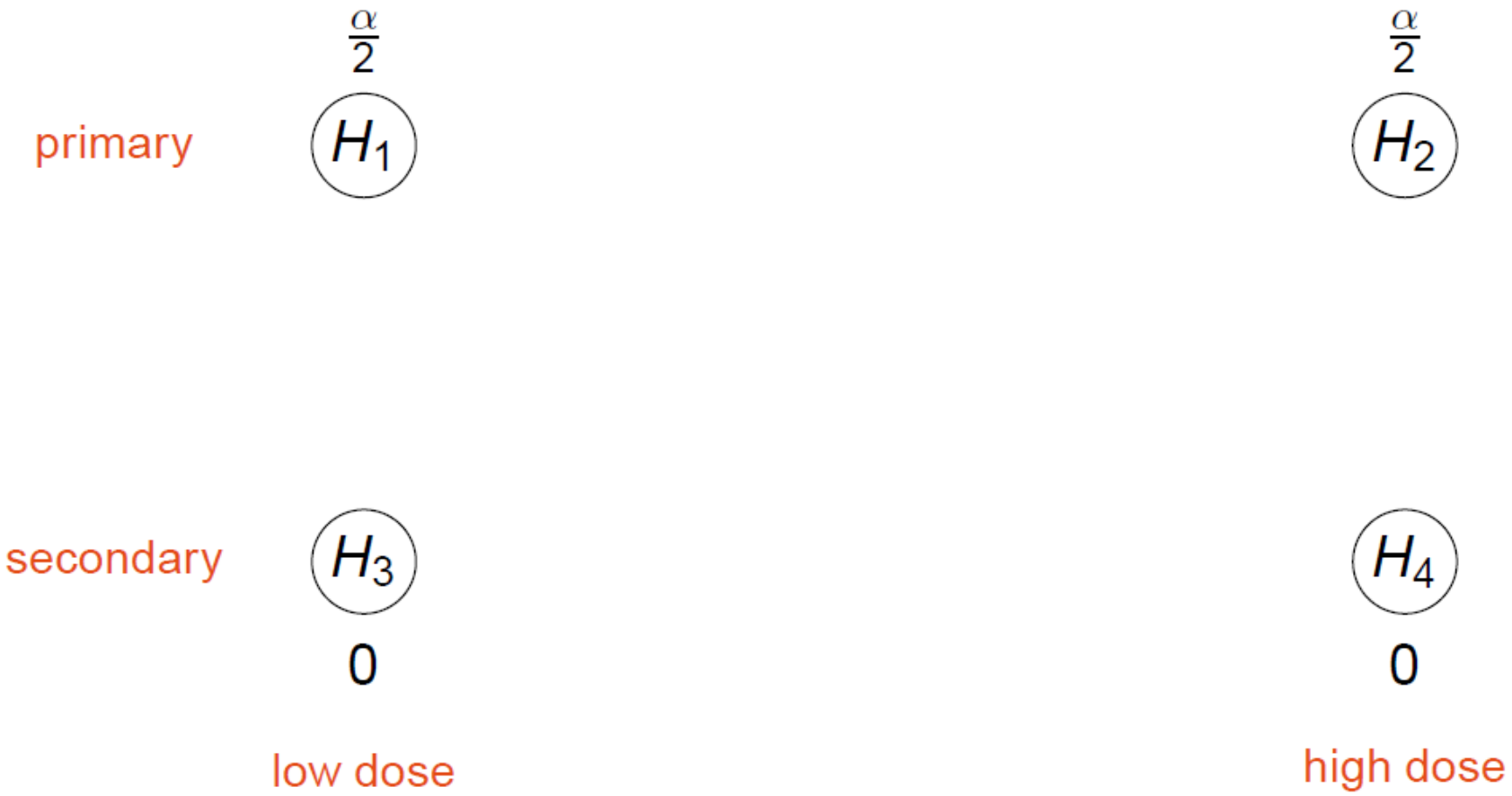
Building a multiple test procedure: Hypotheses



NOVARTIS

COPD Example Revisited

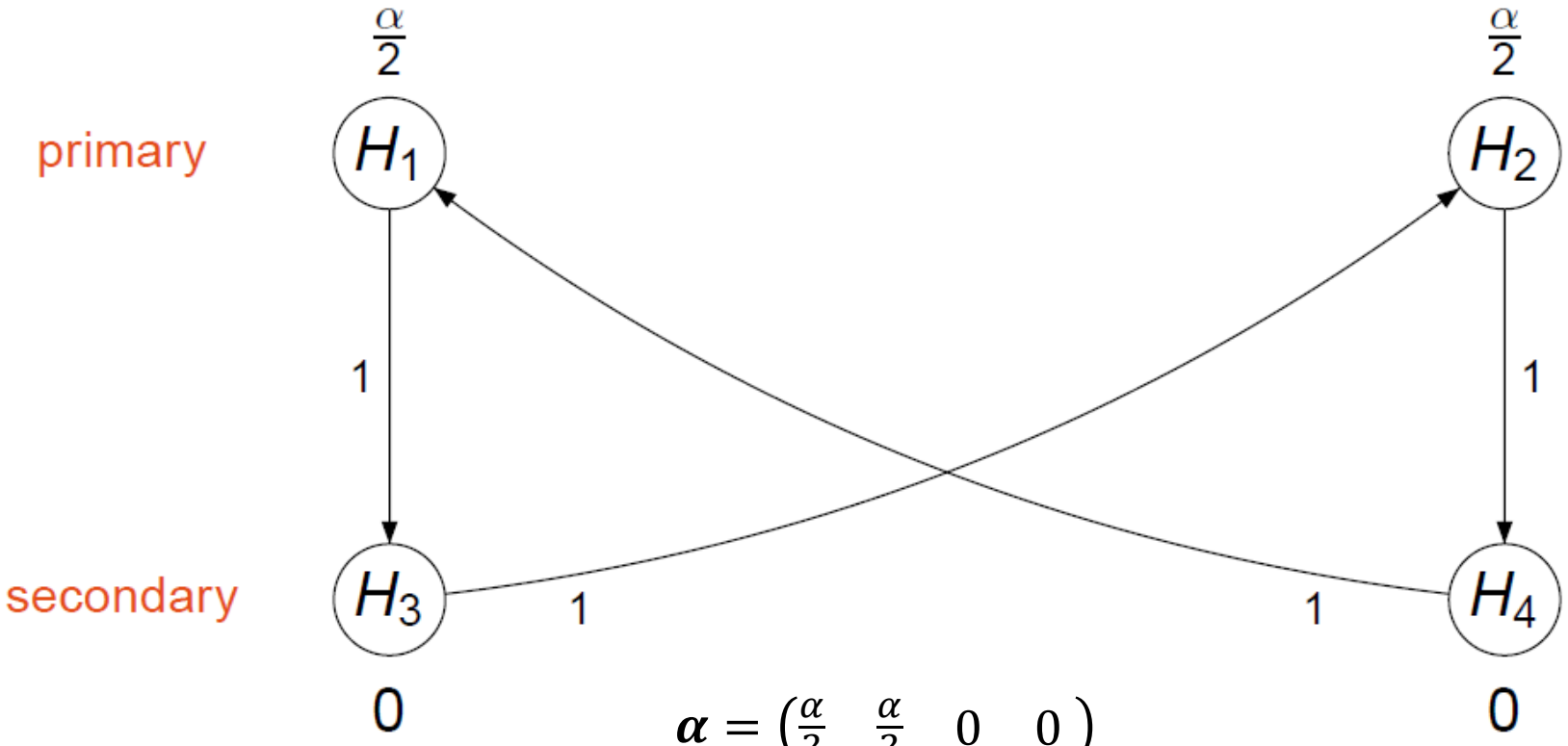
Building a multiple test procedure: Initial levels α



NOVARTIS

COPD Example Revisited

Building a multiple test procedure: α–propagation



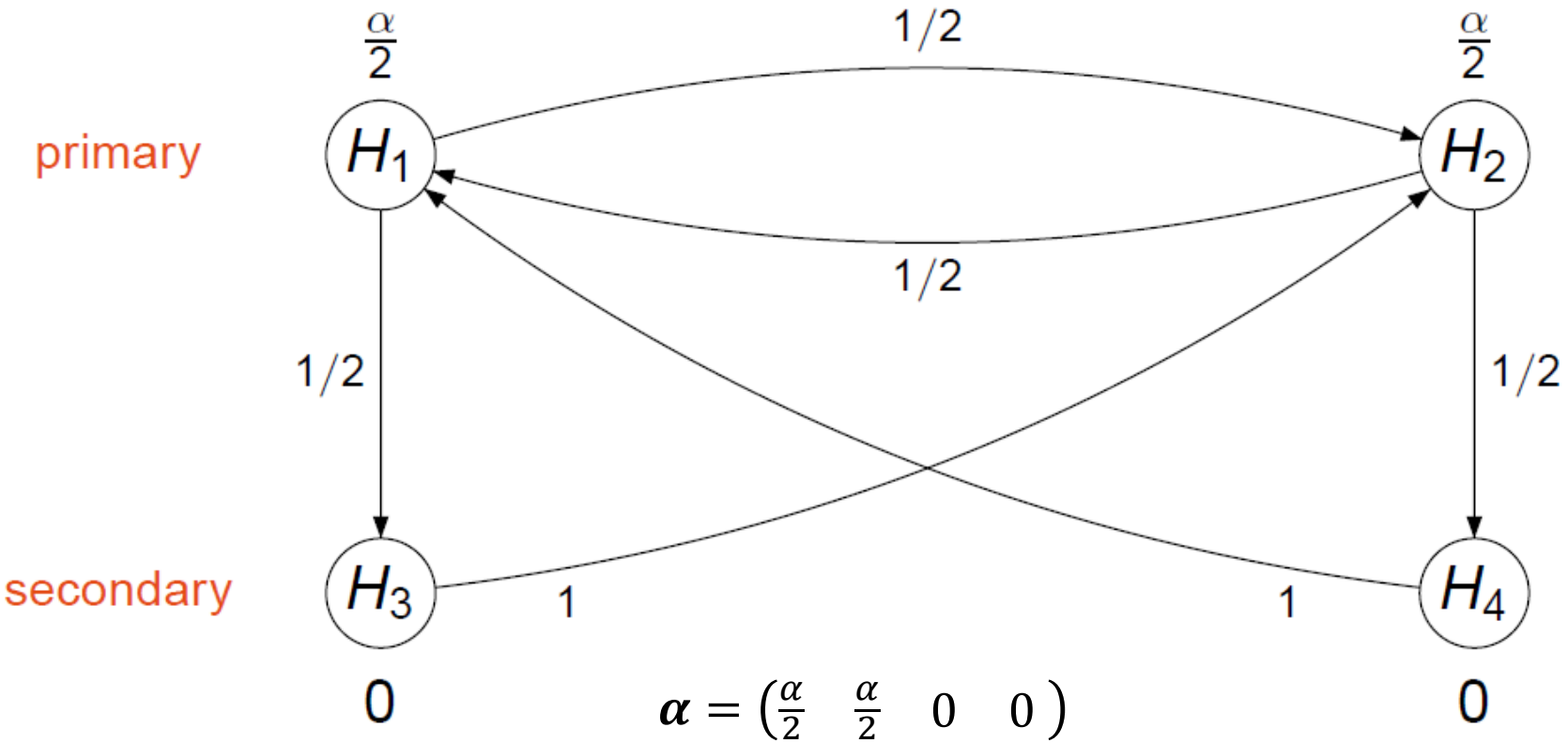
$$\boldsymbol{\alpha} = \left(\frac{\alpha}{2} \quad \frac{\alpha}{2} \quad 0 \quad 0\right)$$

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

NOVARTIS

COPD Example Revisited

Building a multiple test procedure: Alternative α–propagation



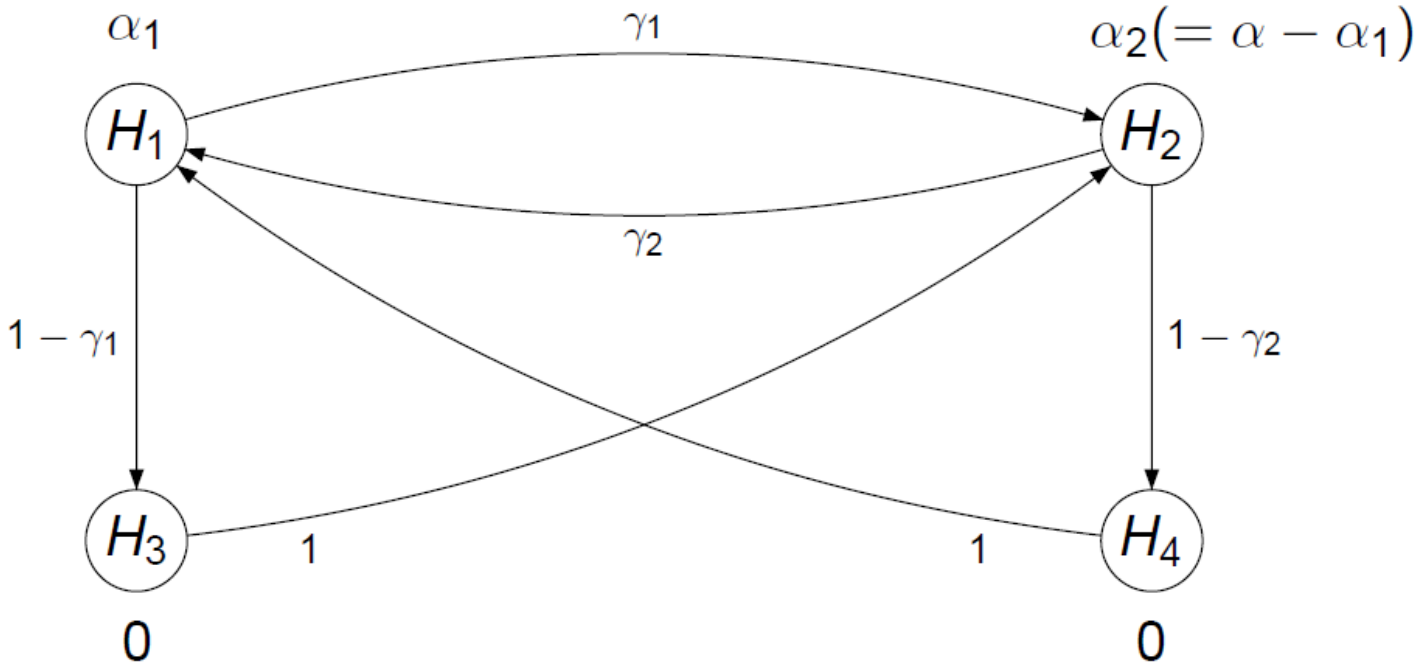
$$\boldsymbol{\alpha} = \left(\frac{\alpha}{2} \quad \frac{\alpha}{2} \quad 0 \quad 0\right)$$

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

NOVARTIS

COPD Example Revisited

Building a multiple test procedure: General solution



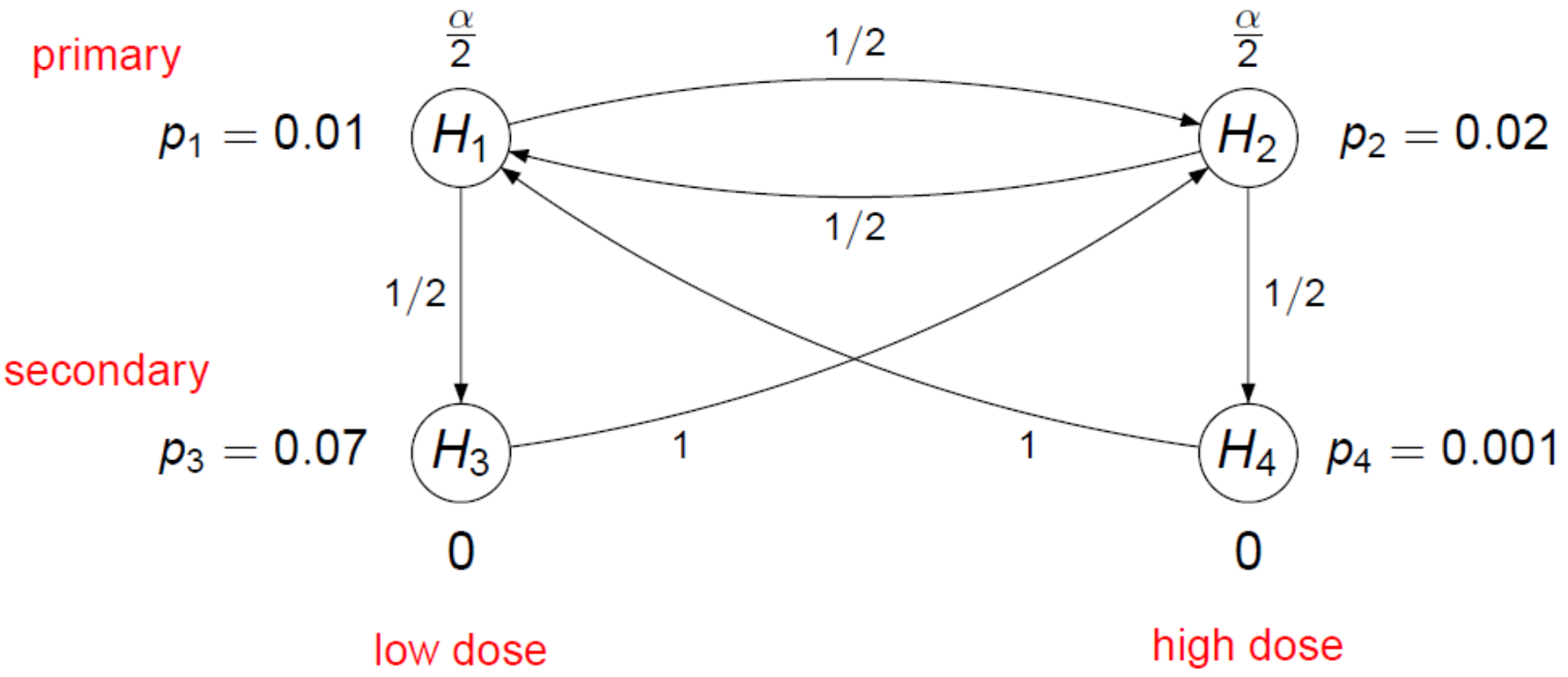
$$\boldsymbol{\alpha} = (\alpha_1 \quad \alpha_2 \quad 0 \quad 0)$$

$$\boldsymbol{G} = \begin{pmatrix} 0 & \gamma_1 & 1-\gamma_1 & 0 \\ \gamma_2 & 0 & 0 & 1-\gamma_2 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- Resulting graph depends on only three parameters α_1, γ_1, and γ_2 that can be finetuned based on:
  - further clinical considerations, or
  - assumptions about effect sizes, correlations, ...

COPD Example Revisited

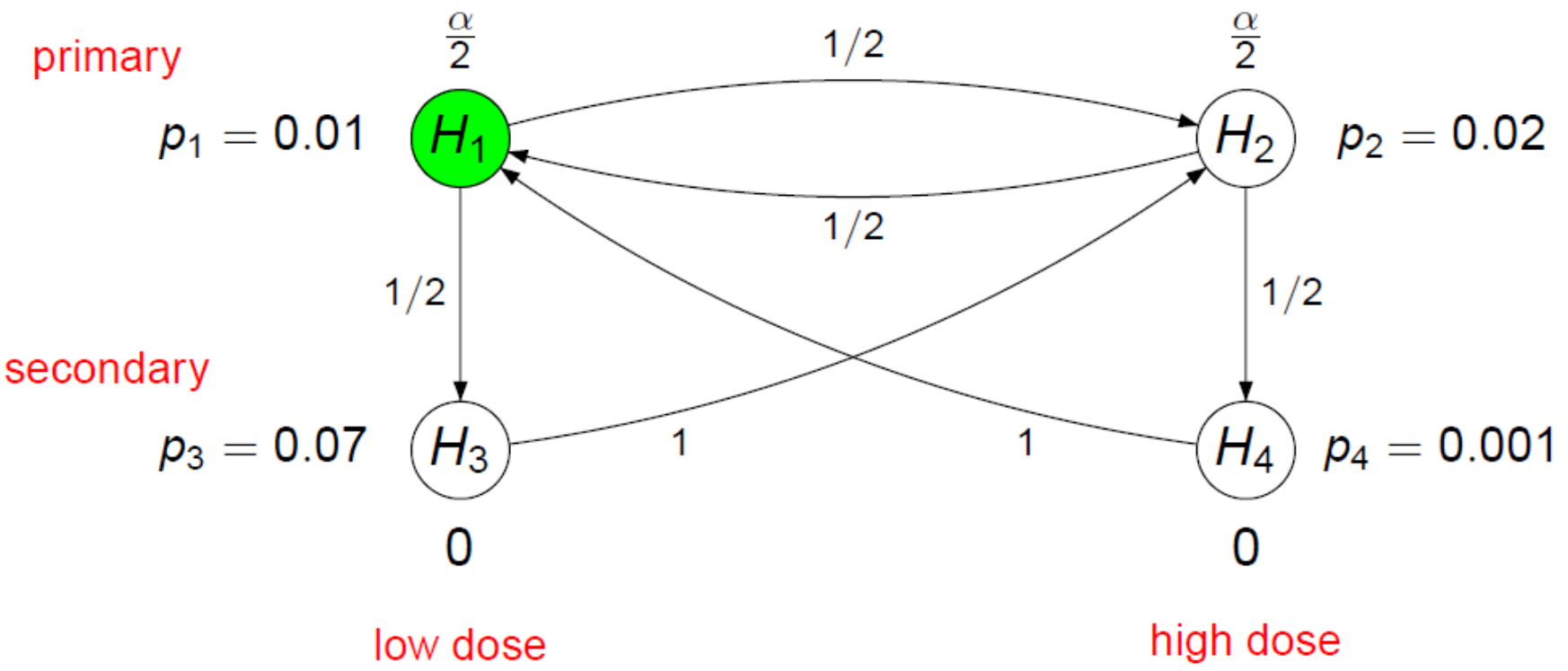
Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

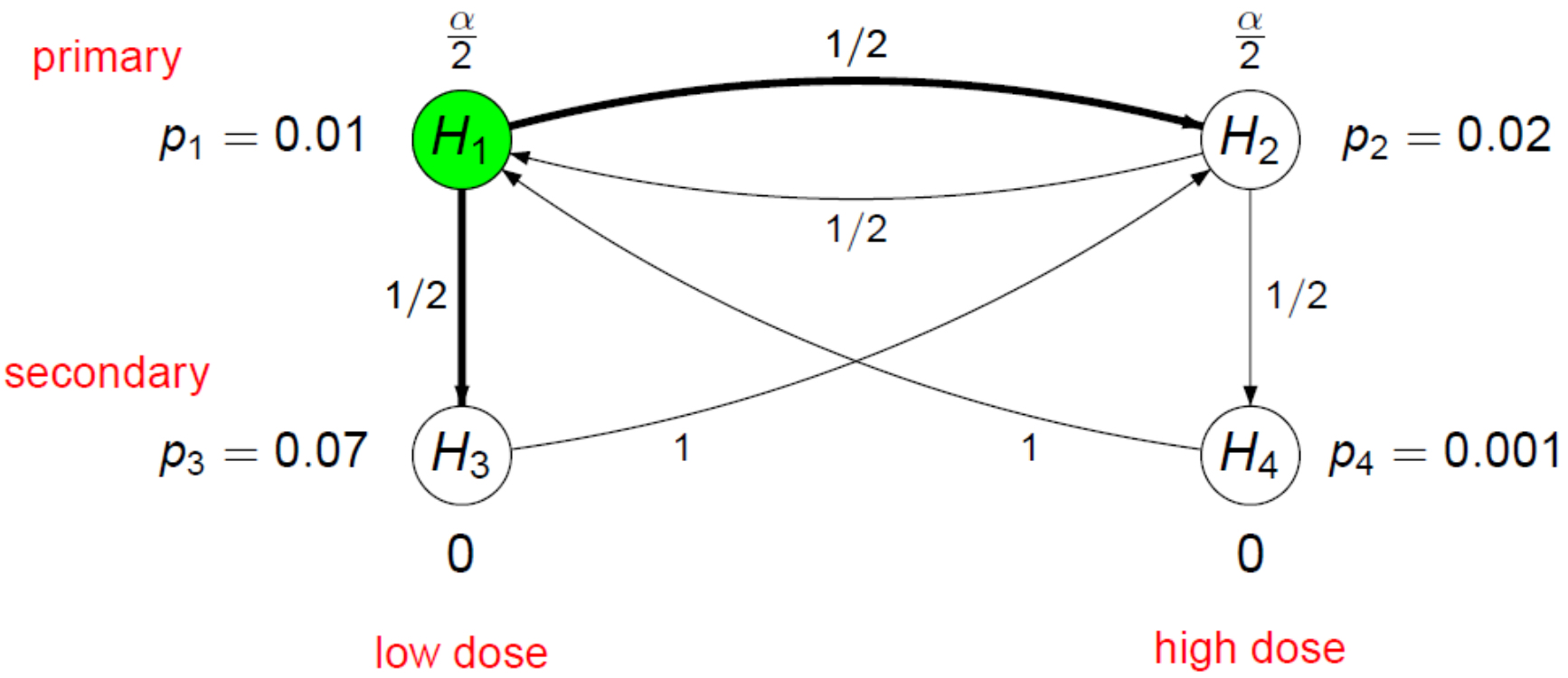
Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

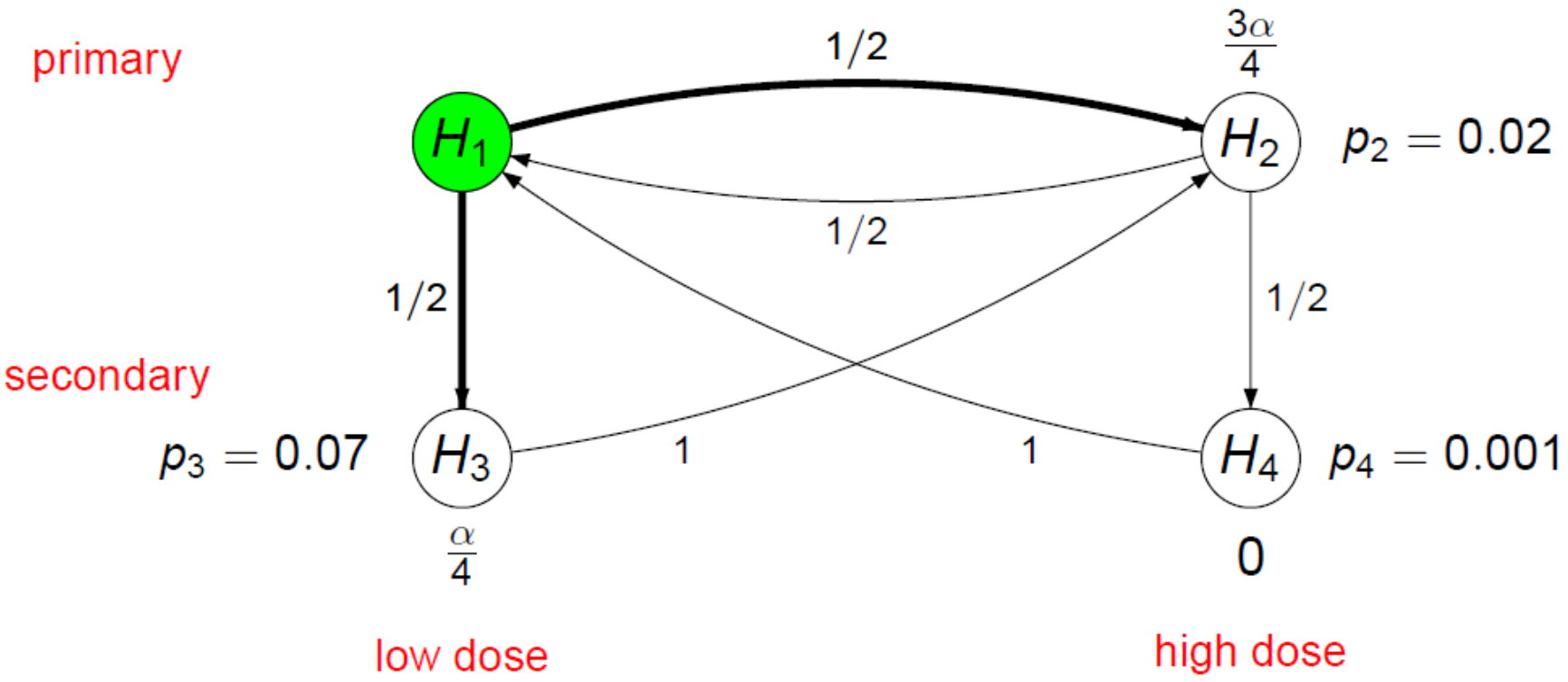
Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

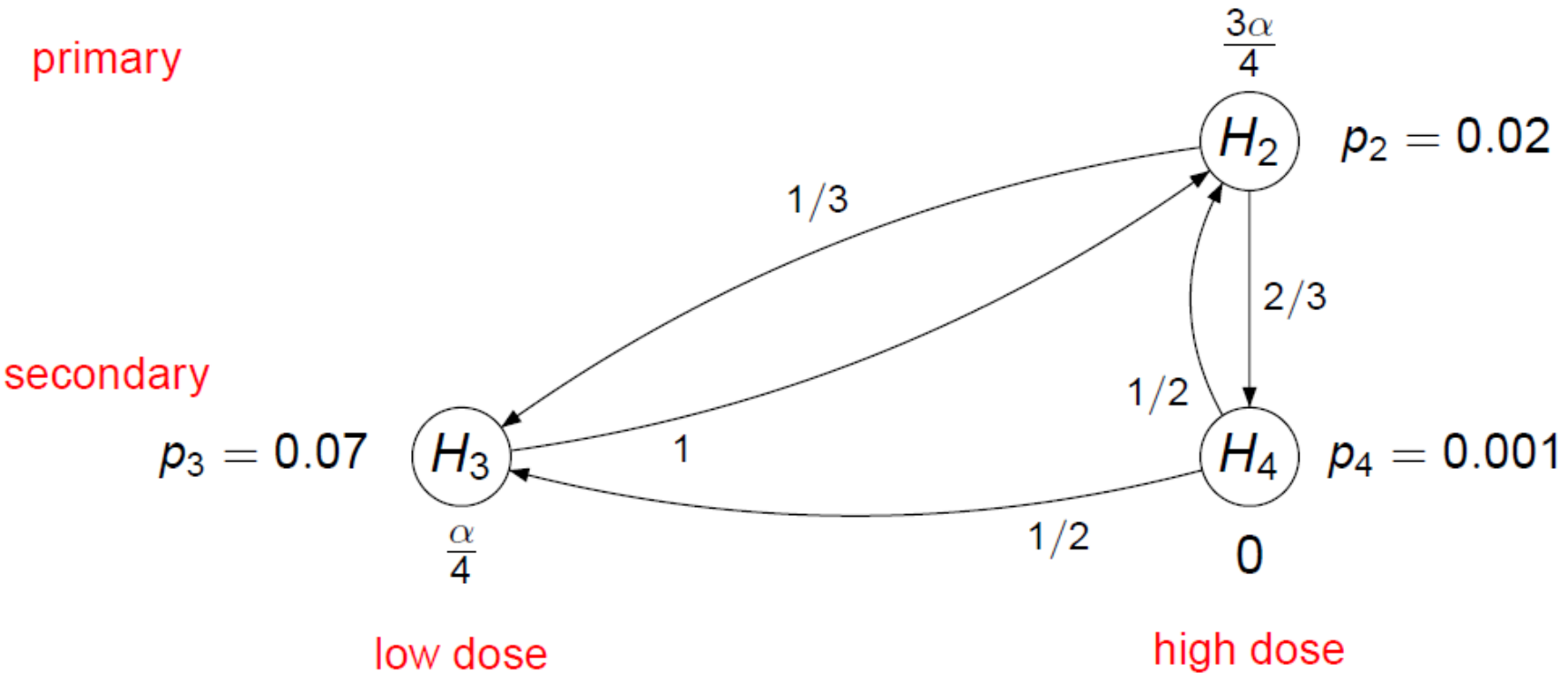
Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

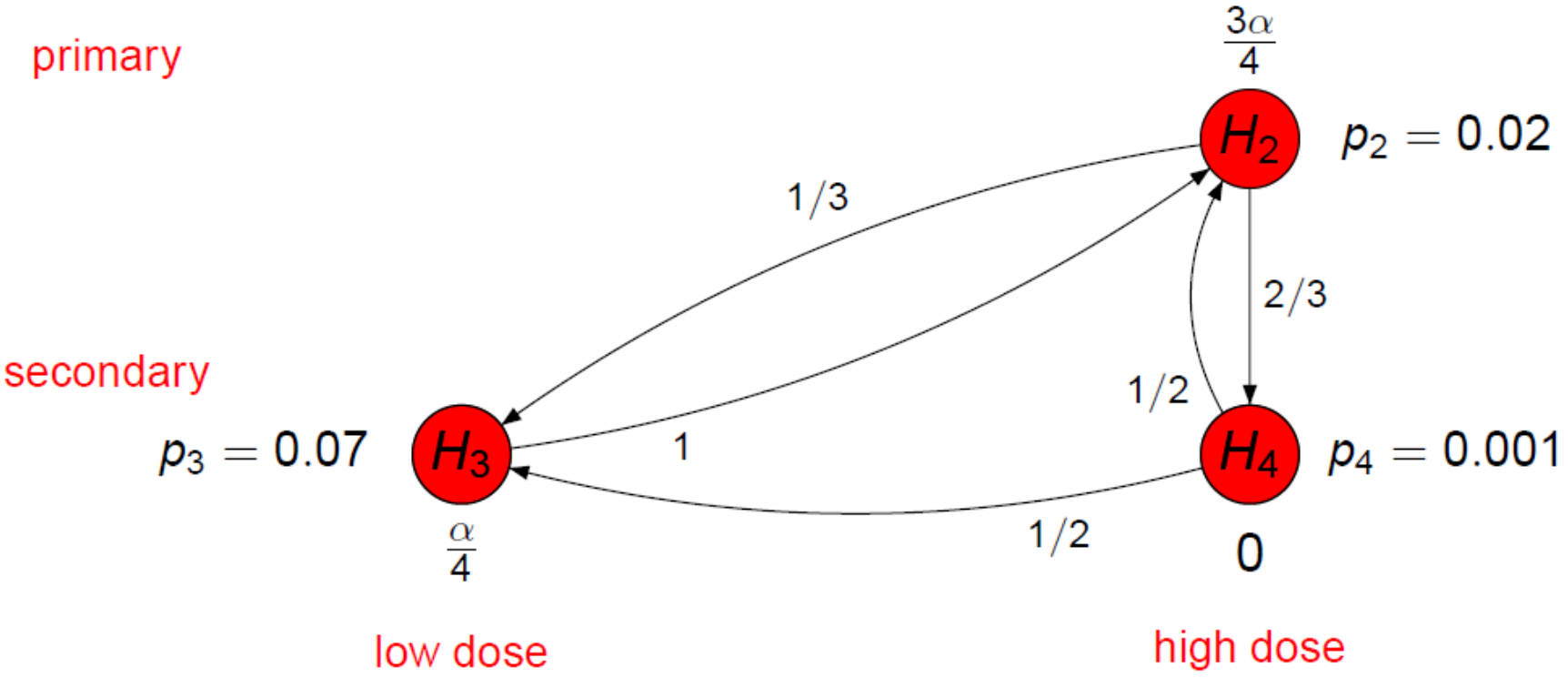
Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

Numerical example with $\alpha = 0.025$



NOVARTIS

COPD Example Revisited

SAS: Main function

```
/* h: indicator whether a hypothesis is rejected (= 1) or not (= 0) (1 x n vector)
   a: initial significance level allocation (1 x n vector)
   w: weights for the edges (n x n matrix)
   p: observed p-values (1 x n vector) */

START mcp(h, a, w, p);
    n = NCOL(h);
    mata = a;

    crit = 0;
    DO UNTIL(crit = 1);
        test = (p < a);
        IF (ANY(test)) THEN DO;
            rej = MIN(LOC(test#(1:n)));
            h[rej] = 1;
            w1 = J(n, n, 0);
            DO i = 1 TO n;
                a[i] = a[i] + a[rej]*w[rej,i];
                IF (w[i,rej]*w[rej,i]<1) THEN DO j = 1 TO n;
                    w1[i,j] = (w[i,j] + w[i,rej]*w[rej,j])/(1 - w[i,rej]*w[rej,i]);
                END;
                w1[i,i] = 0;
            END;
            w = w1; w[rej,] = 0; w[,rej] = 0;
            a[rej]  = 0;
            mata = mata // a;
        END;
        ELSE crit = 1;
    END;

    PRINT h; PRINT (ROUND(mata, 0.0001)); PRINT (ROUND(w,0.01));
FINISH;
```

COPD Example Revisited

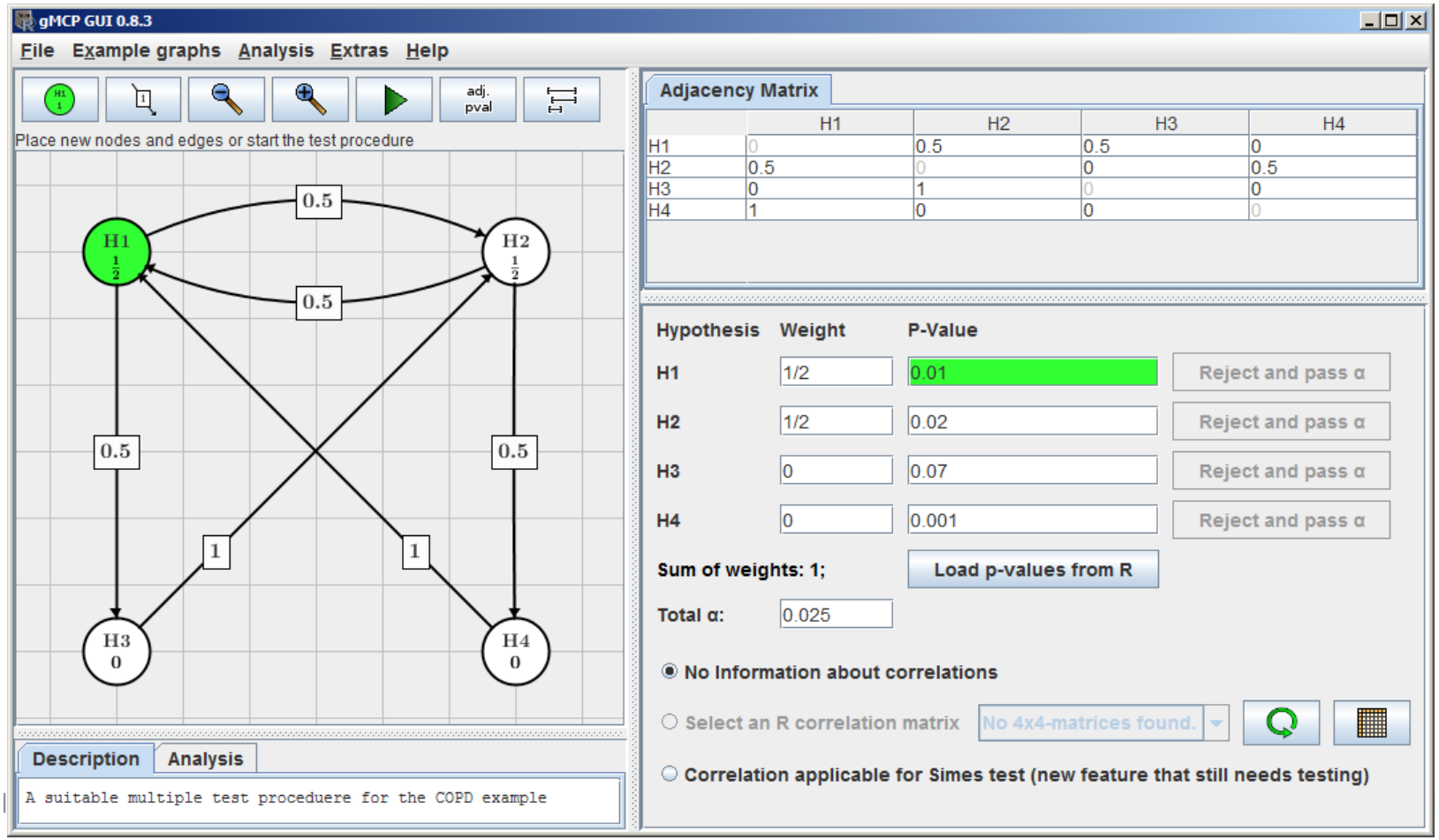
SAS: Example call

```
START mcp(h, a, w, p);
    ...
FINISH;

/*** Numerical example ***/
h = {0      0      0    0    };
a = {0.0125 0.0125 0    0    };
w = {0      0.5    0.5  0    ,
     0.5    0      0    0.5  ,
     0      1      0    0    ,
     1      0      0    0    };
p = {0.01   0.02   0.07 0.001};

RUN mcp(h, a, w, p);
QUIT;
```


COPD Example Revisited

R: gMCP package

- Open source package at http://cran.r-project.org/web/packages/gMCP/
- Provide graphical user interface (GUI) within R through JAVA



Summary

- Proposed graphical approach offers the possibility to
  - Tailor advanced multiple test procedures to structured families of hypotheses,
  - Visualize complex decision strategies in an efficient and easily communicable way, and
  - Ensure strong FWER control
- Approach covers many common multiple test procedures as special cases
  - Holm, fixed sequence, fallback, gatekeeping, ...

- Introduction
- Common Multiple Test Procedures
- Hierarchical Test Procedure
- Closed Test Procedure
- Graphical Approach
- Summary and Conclusions

NOVARTIS

Summary

- Multiplicity raises challenging problems which affect almost every decision throughout drug development
- Closed test procedure is a general principle to construct powerful multiple test procedures; many common procedures are CTPs
- For structured hypotheses, one can apply the graphical approach, which is based on CTPs
  - Reflect the difference in importance as well as the relationship between the various study objectives
  - Are often applied to clinical trials with structured families of hypotheses and several levels of multiplicity

Summary

- It is critical to choose the suitable method for a particular problem
- There are different types of multiplicity problems that need other methods than those described here, such as:
  - Safety data analyses
  - Large-scale testing in genetics, proteomics etc.
  - Post-hoc analyses / data snooping

Q & A

Any questions?

NOVARTIS

Appendix

Main References

- Alosh, M., Bretz, F., and Huque, M. (2014) Advanced multiplicity adjustment methods in clinical trials. Statistics in Medicine 33(4), 693-713.

- Bretz, F., Maurer, W. and Maca J. (2014) Graphical Approaches to Multiple Testing. To appear as Chapter 14 in: *Clinical Trial Biostatistics and Biopharmaceutical Applications* (ed: Walter Young and Ding-Geng (Din) Chen), Taylor and Francis, Boca Raton

- Bretz, F., Hothorn, T., and Westfall, P. (2010). *Multiple Comparisons with R*. Chapman and Hall, Boca Raton.

- Dmitrienko, A., Tamhane, A. C. and Bretz, F. (Eds.) (2009). *Multiple Testing Problems in Pharmaceutical Statistics.* Chapman & Hall/CRC Biostatistics Series, Boca Raton.

- Westfall, P., Tobias, R. and Wolfinger, R. (2011). *Multiple Comparisons and Multiple Tests Using SAS.* SAS Press, Cary, NC.

Appendix

Regulatory Guidelines

- ICH E9 (1998) on “Statistical principles for clinical trials”
- CPMP (2002) Points to consider on “Multiplicity issues in clinical trials”
- FDA draft guidance for industry on “Multiple endpoint analyses” expected for 2014

