

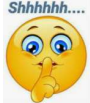


## **Graphical Approaches to Multiple Test Problems**

Ekkehard Glimm, Frank Bretz (Novartis) & Dong Xi (Gilead)  
Basel Biometric Society – March 29, 2022

# Our housekeeping

---

- Event will be recorded
- Presentations & recording will be made available to the audience
- Please mute yourself  unless you are speaking
- Q & A:
  - After each presentation we will have time for questions
  - Please enter your questions into the **chat** during the talk or raise your hand during the Q&A to ask your question.
  - In case of any problems during the event, please contact [bibiana.blatna@novartis.com](mailto:bibiana.blatna@novartis.com)



# Disclaimer

---

The views expressed in this presentation are those of the presenters and do not necessarily represent the views of, and should not be attributed to, the presenters' affiliations.

# Agenda

---

**14:00 – 14:45**

**Introduction to multiple testing**

*Dong Xi*

**14:45 – 16:15**

**Graphical approaches to multiple testing**

*Frank Bretz*

Break

**16:30 – 17:30**

**Extensions to group sequential designs**

*Ekkehard Glimm*

**17:30 – 18:00**

**Extensions to pooled analyses from two studies**

*Dong Xi*

# Learning objectives

---

- Learn about advanced problems of multiplicity in drug development
- Get familiar with the closed test procedure, a general construction method for multiple test problems
- Be able to tailor advanced multiple test procedures to given study objectives, and to visualize and implement the graphical approaches

# Agenda

---

**14:00 – 14:45**

**Introduction to multiple testing**

*Dong Xi*

14:45 – 16:15

Graphical approaches to multiple testing

*Frank Bretz*

Break

16:30 – 17:30

Extensions to group sequential designs

*Ekkehard Glimm*

17:30 – 18:00

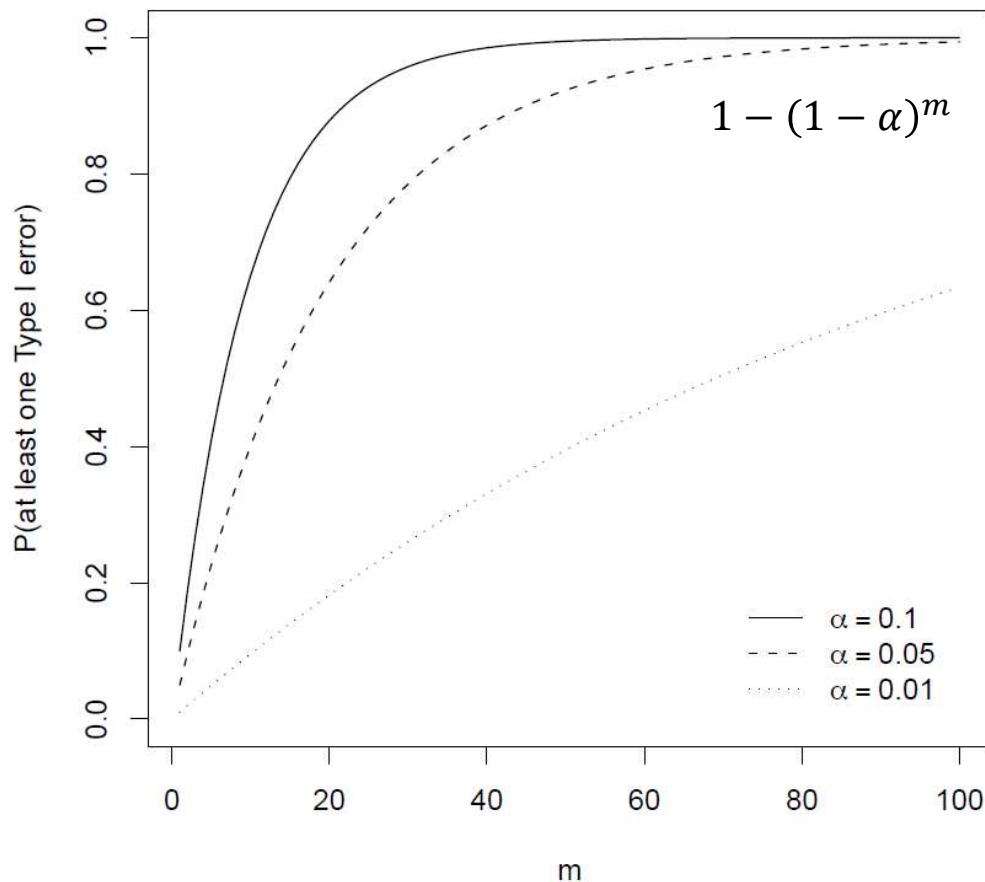
Extensions to pooled analyses from two studies

*Dong Xi*

# Type I error rate inflation

Test  $m$  independent hypotheses

Probability of at least one Type I error  
for different numbers of hypotheses  $m$



- Probability of making Type I error increases as  $m$  or  $\alpha$  increases
- For large  $m$  we almost surely reject incorrectly at least one of the true null hypotheses

# Sources of multiplicity

---

- Multiple test problems are very common in clinical trials, such as the comparison of a new treatment with
  - Several other treatments
  - A control for more than one endpoint
  - A control for more than one population
  - A control repeatedly in time
- Clinical trials often face several sources of multiplicity at the same time
- Target: To control the familywise error rate (FWER)  
 $\Pr(\text{reject at least one true null}) \leq \alpha$  under any configuration of true/false null hypotheses



# Common multiple test procedures

	Correlations		
	Without		With
Single Step	Bonferroni	Simes	Dunnett
Stepwise	Holm	Hochberg	Stepdown Dunnett

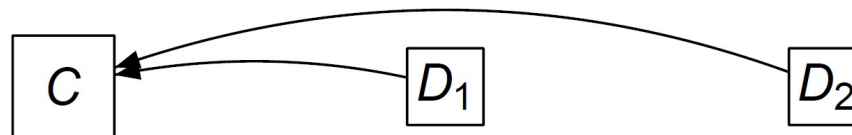
- All these methods treat the hypotheses as equally important
- Remarks on the performance of the procedures
  - Stepwise methods are more powerful than single step methods
    - Single step methods use the same critical values for all hypotheses whereas stepwise methods use different critical values
  - Simes-based methods are more powerful than Bonferroni-based methods
  - Accounting for correlations could lead to more powerful procedures

# An advanced clinical trial example in COPD

Late phase development of a new compound: *Background*

---

- **Objective:** Show that a new drug is better than a control drug in patients with chronic obstructive pulmonary disease (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**  $C$

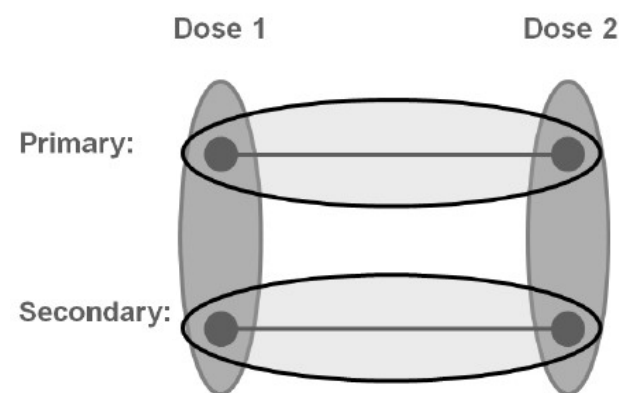


# An advanced clinical trial example in COPD

Late phase development of a new compound: *Hypotheses*

---

- 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 hypothesis  $H_3$  ( $H_4$ ) should be tested, only if the corresponding primary hypothesis  $H_1$  ( $H_2$ ) is rejected



# An advanced clinical trial example in COPD

*Late phase development of a new compound: Summary*

---

- Need for suitable multiple test procedures
- 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, in contrast to their relative order
- We need a multiple test procedure that reflects the relative importance of the hypotheses, as driven by clinical considerations

# Summary

---

- Testing multiple hypotheses may lead to an inflation of the Type I error rate
  - That is, testing individual hypothesis at level  $\alpha$  leads to overall Type I error rate larger than  $\alpha$
- Multiple test problems are very common in clinical trials and multiplicity adjustment should always be considered
- Common multiple test procedures treat all hypotheses equally and do not address the underlying structure of the test problem

# Notation

---

- Assume a “family” of  $m$  inferences

- Parameters of interest are  $\theta_1, \dots, \theta_m$

- Individual null hypotheses

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

- Individual test statistics  $t_1, \dots, t_m$  with unadjusted p-values  $p_1, \dots, p_m$

- Ordered p-values  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$

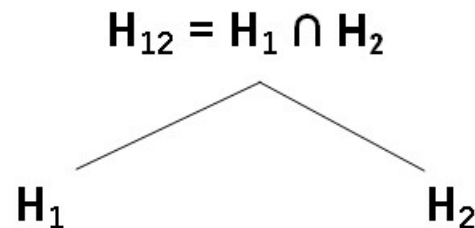
- Ordered null hypotheses according to ordered p-values  $H_{(1)}, \dots, H_{(m)}$

# Closed test procedure (CTP)

Operational definition for  $m = 2$  null hypotheses

---

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



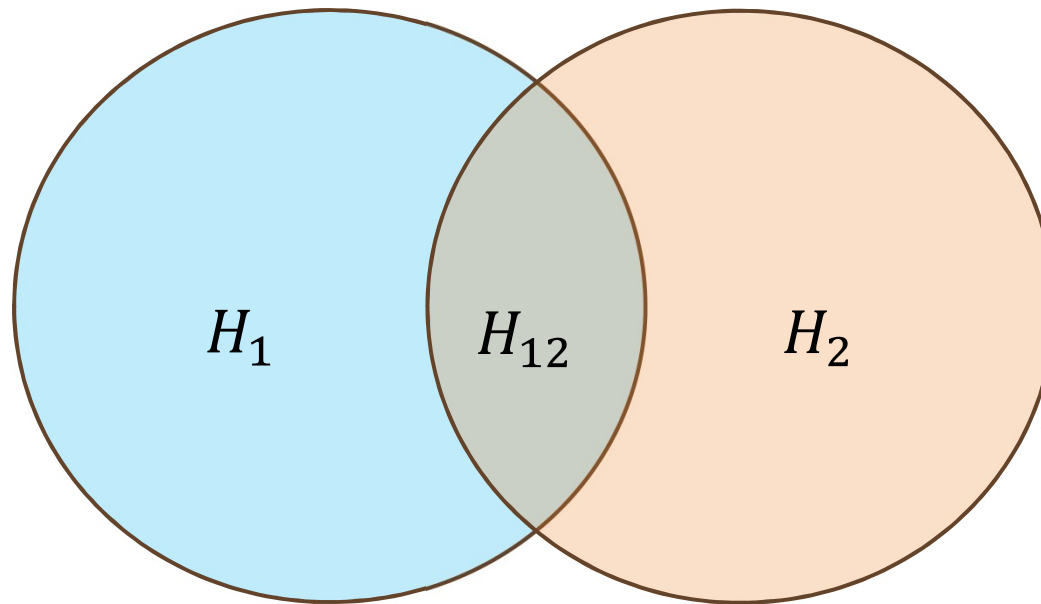
- **Rejection rule:** Reject  $H_1$  ( $H_2$ ), only if both  $H_1$  ( $H_2$ ) and  $H_{12}$  are rejected, each at local level  $\alpha$
- Operationally
  - Test  $H_{12}$  at local level  $\alpha$  (using a suitable test): If rejected, proceed; otherwise stop
  - Test  $H_1$  and  $H_2$  each at local level  $\alpha$ : Reject  $H_1$  ( $H_2$ ) overall if  $H_{12}$  and  $H_1$  ( $H_2$ ) are rejected locally
- This controls FWER as

$$P(\text{at least one rejection}) \leq P(\text{reject the global null}) \leq \alpha$$

# 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  $\alpha$
  - Test  $H_1, H_2$  each using a level  $\alpha$  test

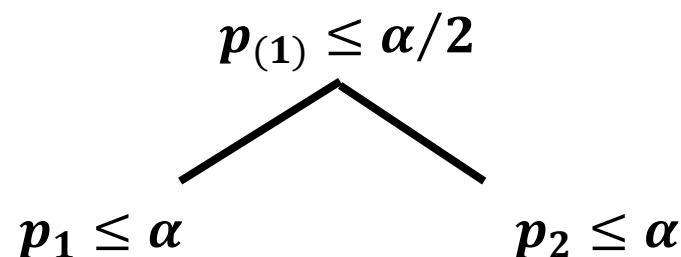


# CTP using Bonferroni

## *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  $\alpha$ , i.e., reject  $H_{(2)}$  if  $p_{(2)} \leq \alpha$
- This results exactly in the Holm procedure

# CTP using Simes

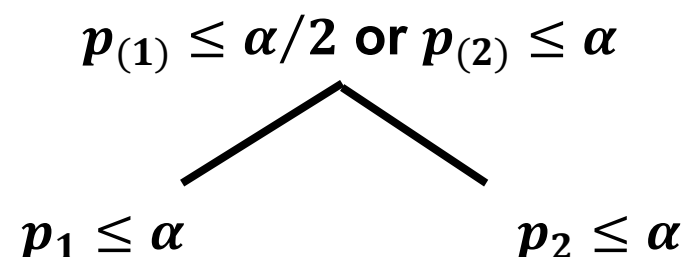
## Hochberg procedure

---

- 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 than the CTP using Simes tests

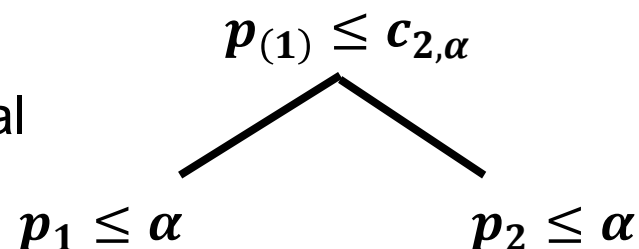
# CTP using Dunnett

## Stepwise Dunnett test

---

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

- $c_{2,\alpha}$  ( $\alpha/2 \leq c_{2,\alpha} \leq \alpha$ ) denotes the critical value for the Dunnett test to compare two treatment with a control



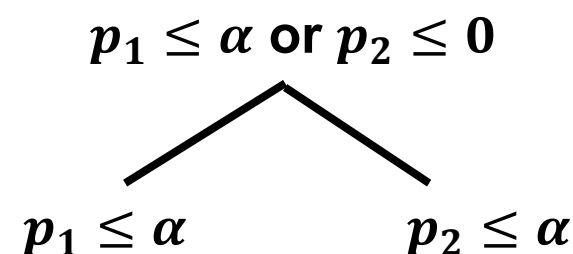
- If we fail to reject  $H_{12}$ , stop
- If we reject  $H_{12}$ , then
  - $H_{(1)}$  is rejected automatically as  $p_{(1)} \leq c_{2,\alpha} \leq \alpha$
  - we only need to test  $H_{(2)}$  at level  $\alpha$ , i.e., reject  $H_{(2)}$  if  $p_{(2)} \leq \alpha$
- This results exactly in the stepwise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

# CTP using weighted Bonferroni (1)

*Fixed sequence procedure*

---

- Two ordered hypotheses  $H_1 \rightarrow H_2$
- 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



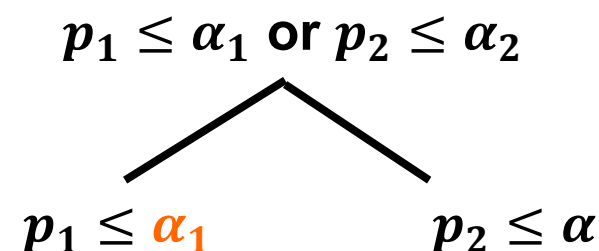
- If we reject  $H_{12}$ , then
  - $H_1$  is rejected automatically as  $p_1 \leq \alpha$
  - we only need to test  $H_2$  at level  $\alpha$ , i.e., reject  $H_2$  if  $p_2 \leq \alpha$
- This results exactly in the fixed sequence procedure

# CTP using weighted Bonferroni (2)

## Fallback procedure

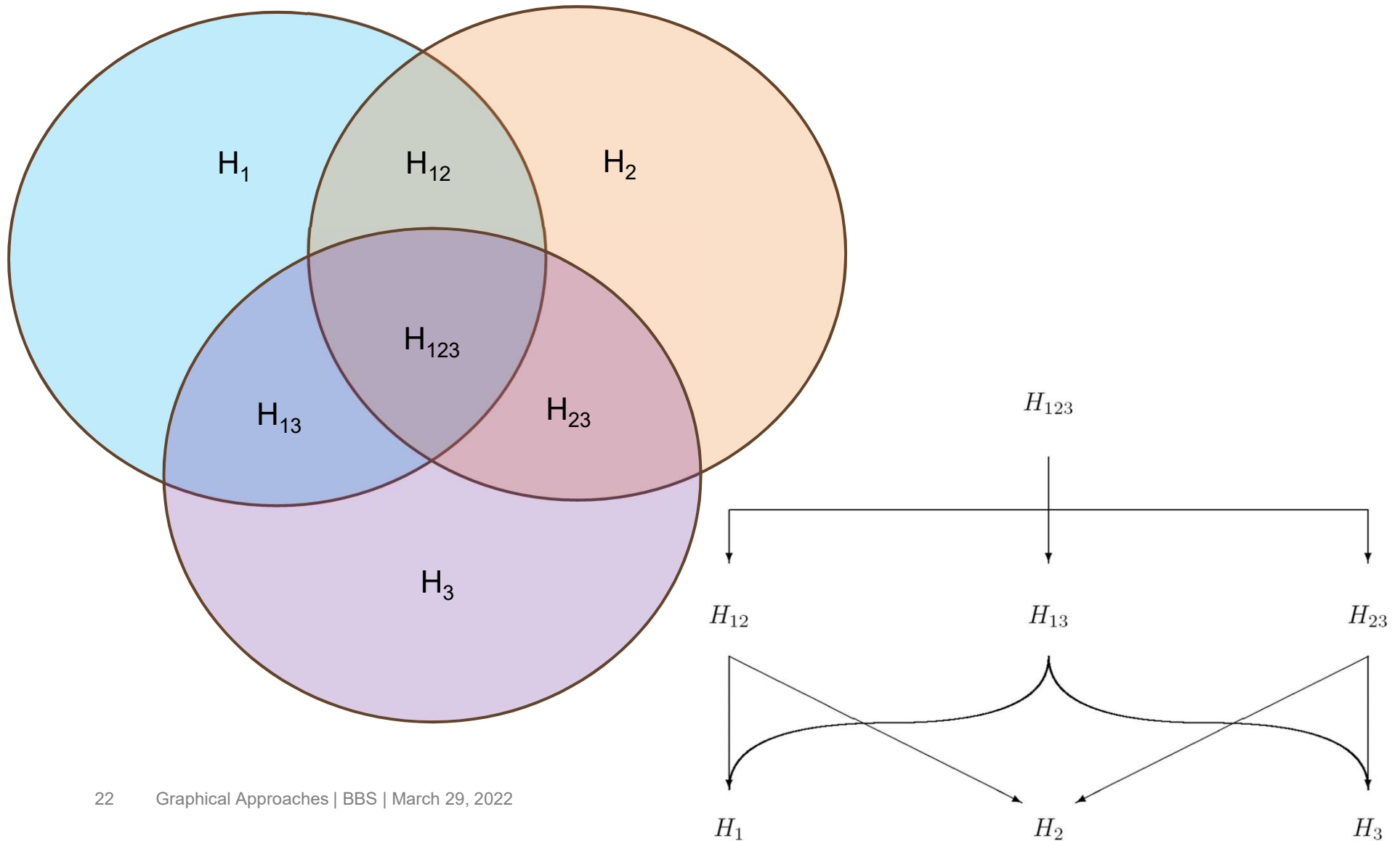
---

- Two ordered hypotheses  $H_1 \rightarrow H_2$
- Using weighted Bonferroni test to test  $H_{12}$ , reject if  $p_1 \leq \alpha_1$  or  $p_2 \leq \alpha_2$ 
  - For the weights,  $\alpha_1 + \alpha_2 = \alpha$
- If we fail to reject  $H_{12}$ , stop
- If we reject  $H_{12}$ 
  - Because  $p_1 \leq \alpha_1$ , then  $H_1$  is rejected automatically and  $H_2$  is tested at level  $\alpha$
  - Because  $p_2 \leq \alpha_2$ , then  $H_2$  is rejected at level  $\alpha$  and  $H_1$  is tested at level  $\alpha_1$
- This results exactly in the fallback procedure



# 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

$$H_J = \bigcap_{i \in J} H_i, \quad J \subseteq \{1, \dots, m\}$$

- Any suitable test can be used to test  $H_J$  at local level  $\alpha$
- An individual  $H_i$  is rejected at level  $\alpha$  if all hypotheses  $H_J$  formed by intersection with  $H_i$  are rejected at local level  $\alpha$
- This controls FWER as
$$P(\text{at least one rejection}) \leq P(\text{reject the global null}) \leq \alpha$$
- CTPs satisfy certain optimality criteria and there is no reason why not to use a CTP

# 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  $\alpha$  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, ...
- The number of intersection hypotheses is  $2^m - 1$ 
  - For large  $m$ , this number increases rapidly and CTPs are in general difficult to apply



Q & A

---

**Any questions?**

# Agenda

---

14:00 – 14:45

Introduction to multiple testing  
*Dong Xi*

14:45 – 16:15

**Graphical approaches to multiple testing**  
*Frank Bretz*

Break

16:30 – 17:30

Extensions to group sequential designs  
*Ekkehard Glimm*

17:30 – 18:00

Extensions to pooled analyses from two studies  
*Dong Xi*

# Outline

---

- Graphical approaches to multiple testing
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example revisited

# Outline

---

- Graphical approaches to multiple testing
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example revisited

# Graphical approach

## Heuristics

---

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

- **$\alpha$ -propagation**

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

# Graphical approach

## Conventions

---

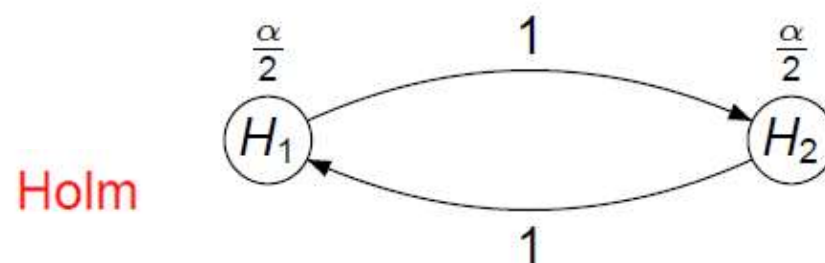
1 Hypotheses  $H_1, \dots, H_m$  represented as nodes



2 Split of significance level  $\alpha$  as weights  $\alpha_1, \dots, \alpha_m$



3 “ $\alpha$  propagation” through weighted, directed edges



# Outline

---

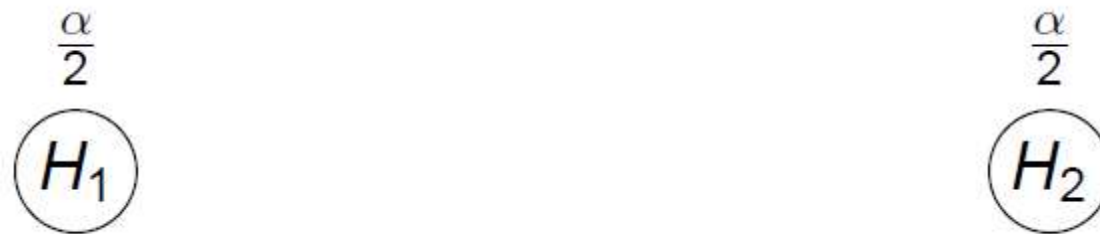
- Graphical approaches to multiple testing
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example revisited

# Graphical approach

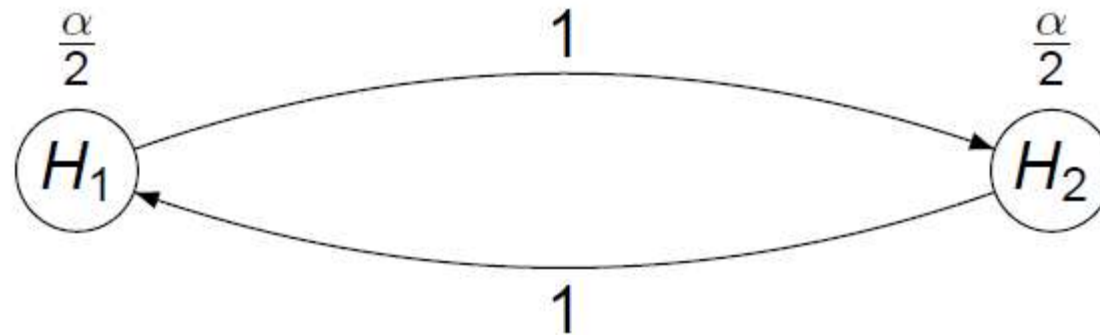
*Bonferroni test and Holm procedure:  $m=2$*

---

- **Bonferroni**: no  $\alpha$ -propagation, i.e. no edges between nodes



- **Holm**: includes  $\alpha$ -propagation and is thus more powerful





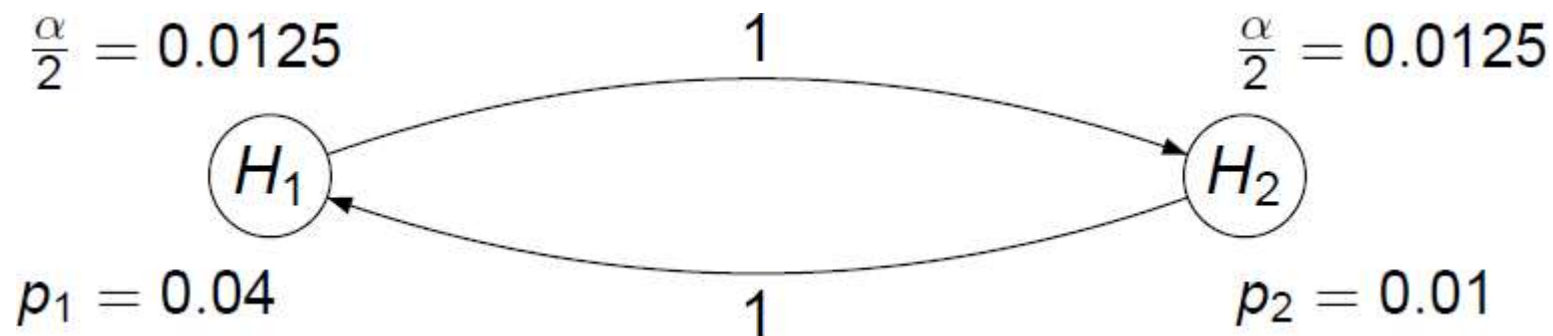
# Graphical approach

Holm procedure: Example with  $\alpha = 0.025$

---

Test  $H_1$  at level  $\alpha/2$

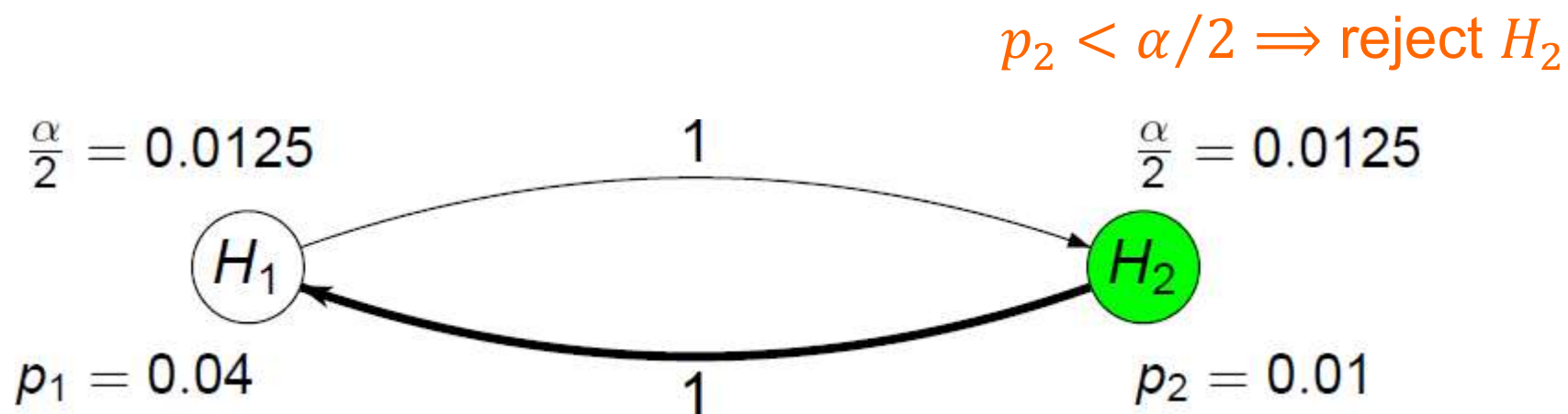
Test  $H_2$  at level  $\alpha/2$



# Graphical approach

Holm procedure: Example with  $\alpha = 0.025$

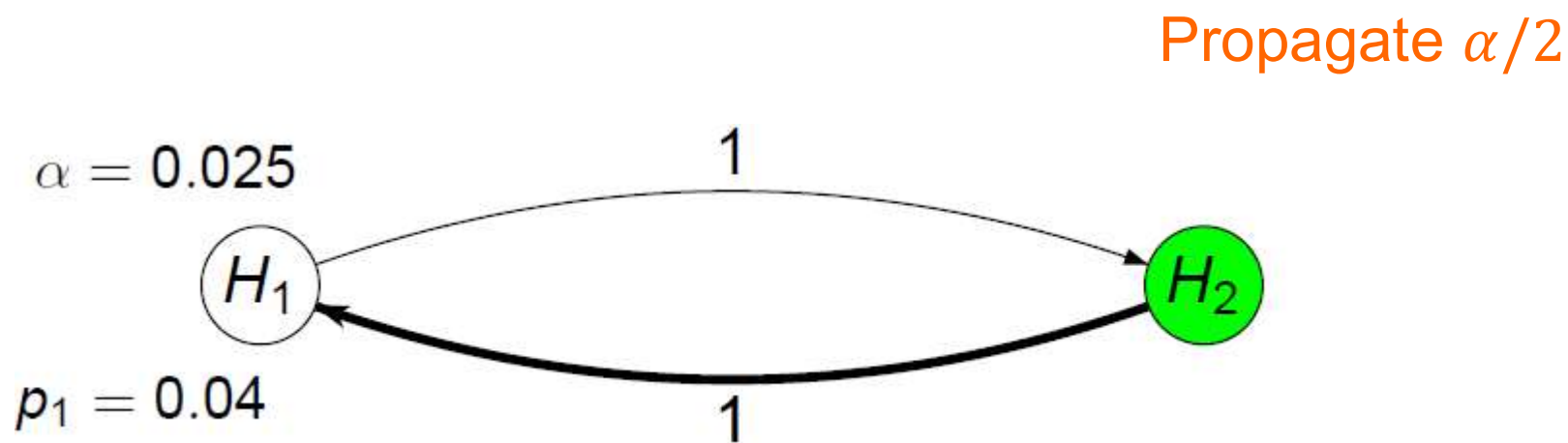
---



# Graphical approach

Holm procedure: Example with  $\alpha = 0.025$

---



# Graphical approach

*Holm procedure: Example with  $\alpha = 0.025$*

---

Remove node for  $H_2$

$$\alpha = 0.025$$

$H_1$

$$p_1 = 0.04$$

# Graphical approach

*Holm procedure: Example with  $\alpha = 0.025$*

---

Test  $H_1$  at level  $\alpha$

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

$$\alpha = 0.025$$



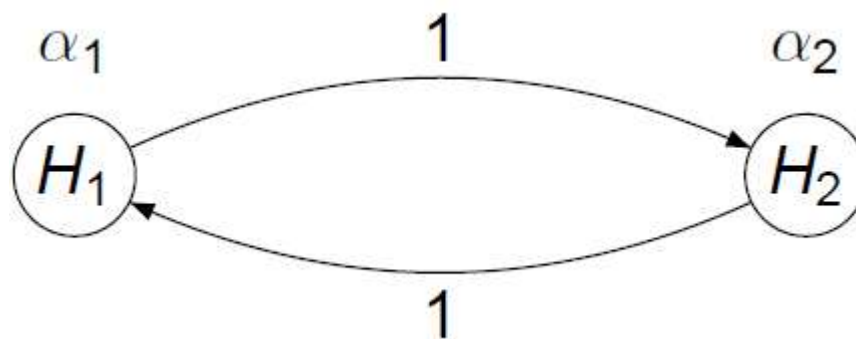
$$p_1 = 0.04$$

# Graphical approach

## *Weighted Holm procedure*

---

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

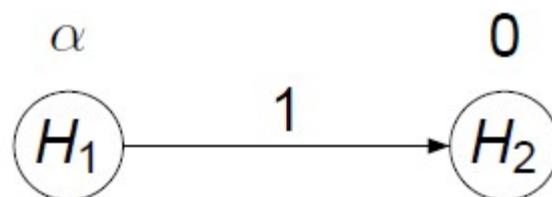


# Graphical Approach

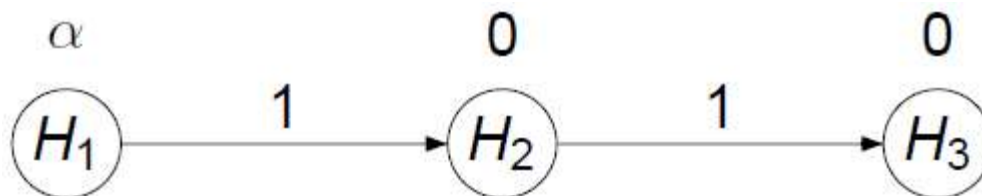
## Fixed sequence test

---

- Assume  $H_1 \rightarrow H_2$ 
  - That is,  $m = 2$  and  $H_1$  is more important than  $H_2$
  - Then the fixed sequence procedure is visualized as



- Similarly, assume for  $m = 3$  that  $H_1 \rightarrow H_2 \rightarrow H_3$ 
  - Then the fixed sequence procedure is visualized as



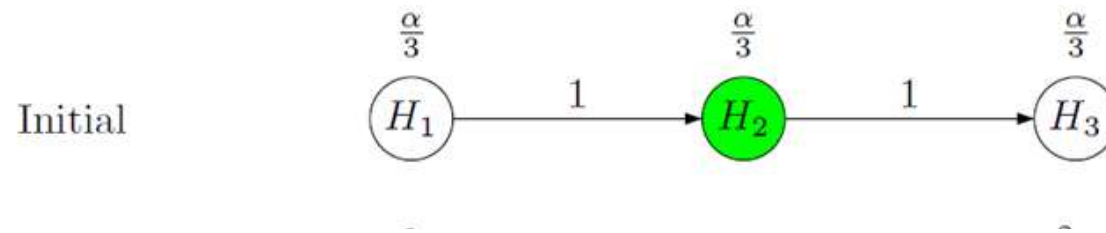
- **Caution:** If  $H_1$  cannot be rejected, we cannot test  $H_2, H_3$  regardless of their p-values

# Graphical Approach

## *Fallback procedure*

---

- 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:



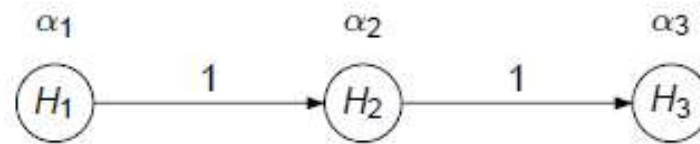


# Graphical Approach

## *Improved fallback procedures*

---

Original fallback



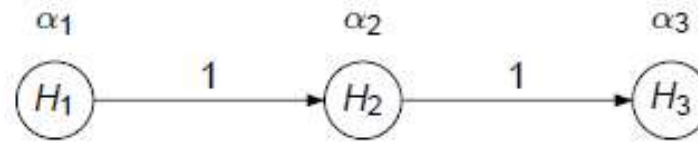
(Wiens, 2003)

# Graphical Approach

## Improved fallback procedures

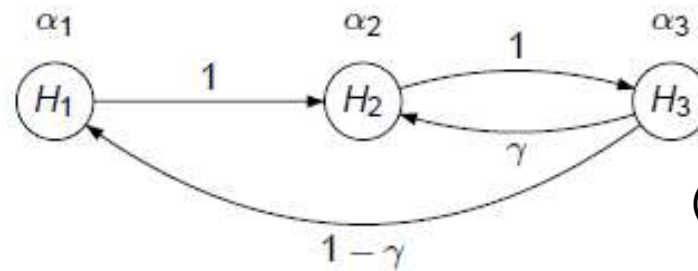
---

### Original fallback



(Wiens, 2003)

### Improved fallback I



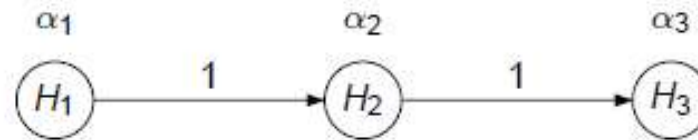
$$\gamma = \frac{\alpha_2}{\alpha_1 + \alpha_2}$$

(Wiens & Dmitrienko, 2005)

# Graphical Approach

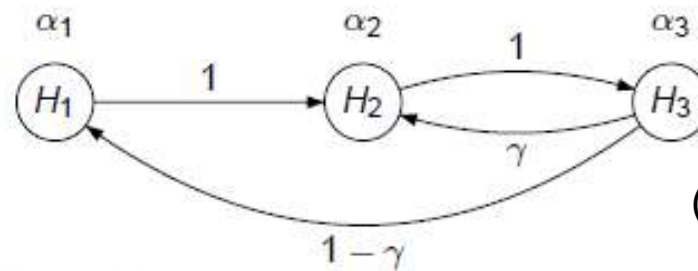
## Improved fallback procedures

### Original fallback



(Wiens, 2003)

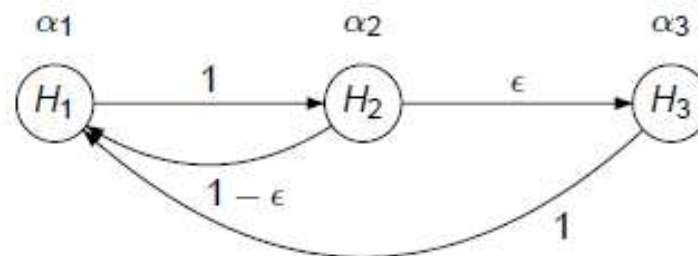
### Improved fallback I



$$\gamma = \frac{\alpha_2}{\alpha_1 + \alpha_2}$$

(Wiens & Dmitrienko, 2005)

### Improved fallback II



$$\epsilon \rightarrow 0$$

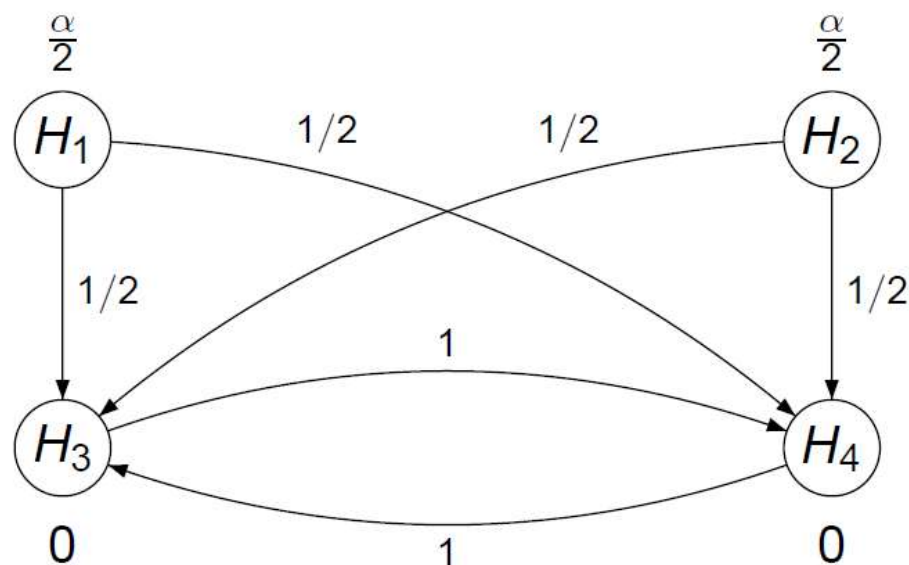
(Hommel & Bretz, 2008)

# Graphical Approach

*Parallel gatekeeping procedure (Dmitrienko et al., 2003)*

---

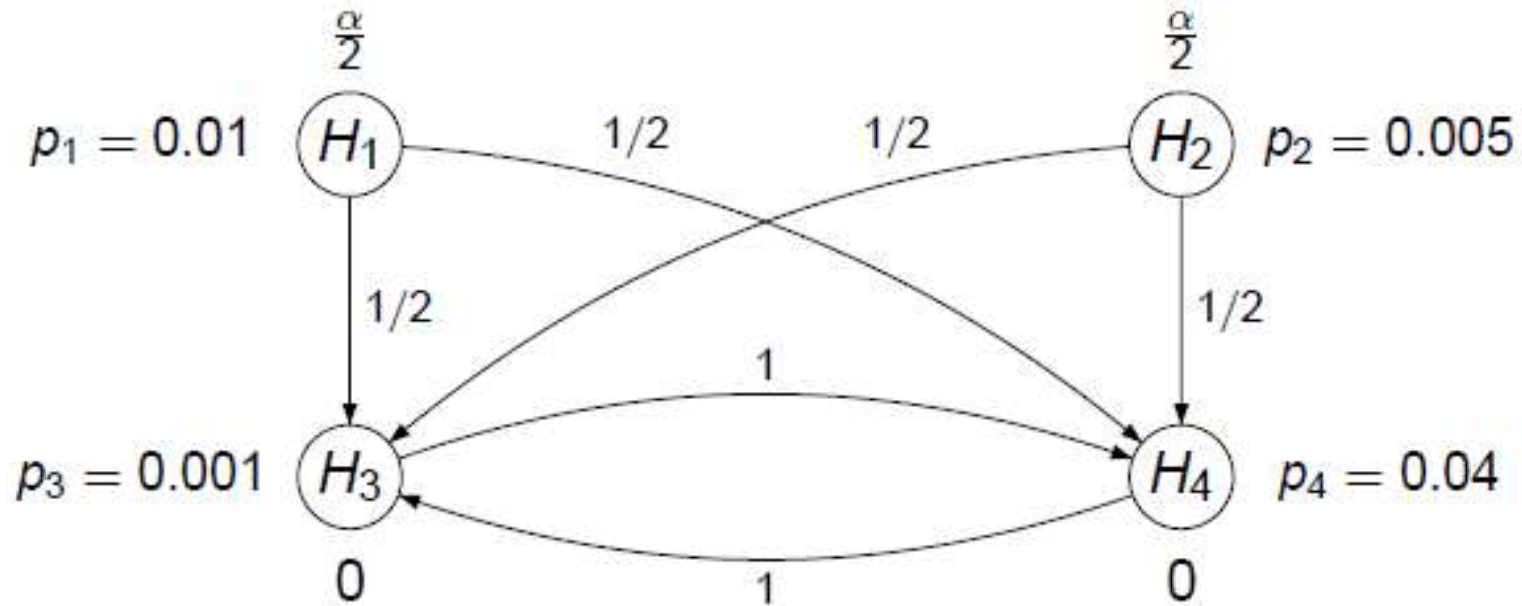
- $H_1, H_2$  are two primary hypotheses
  - For example, comparison of a new drug with placebo for two primary endpoints
- $H_3, H_4$  are two secondary hypotheses
  - For example, comparison of a new drug with placebo for two secondary endpoints
- Parallel gatekeeping: Testing of secondary hypotheses occurs if at least one of the primary hypotheses is rejected



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

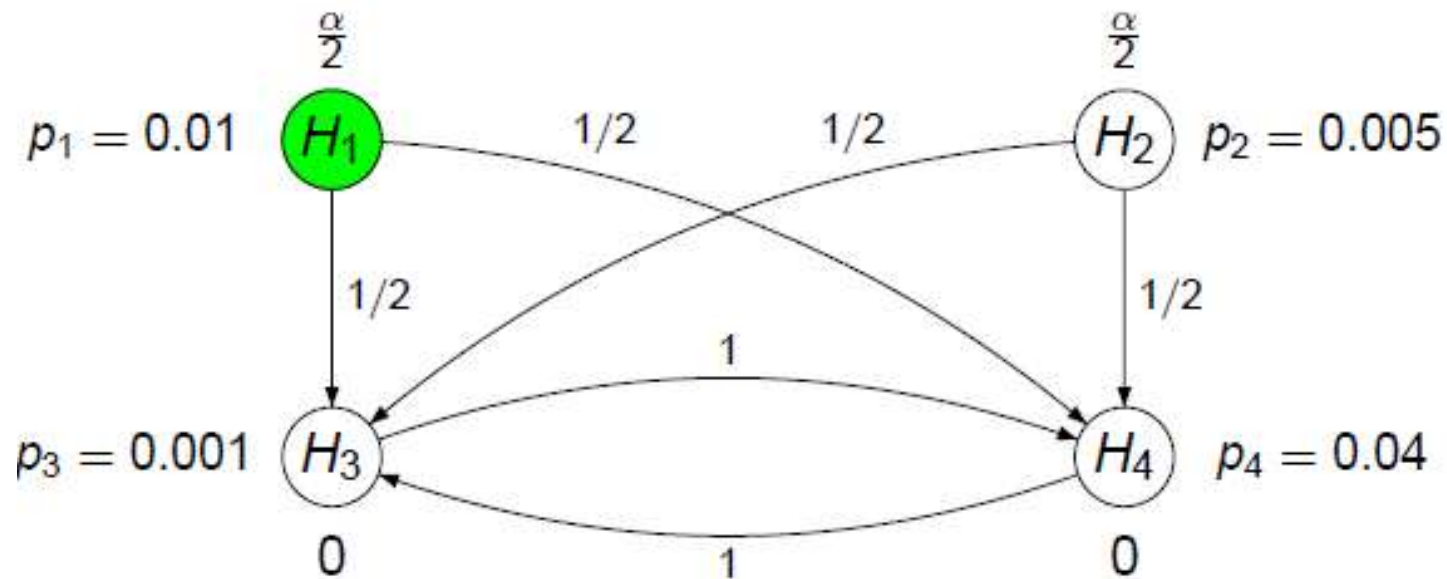
---



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

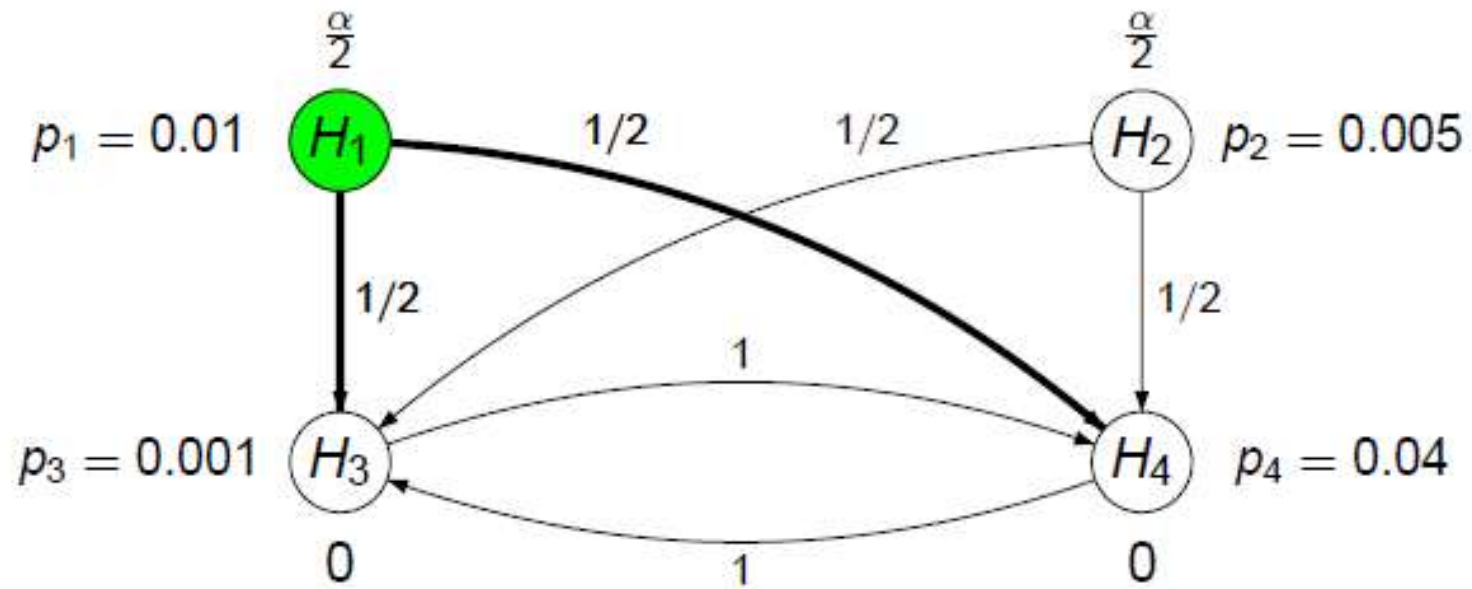
---



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

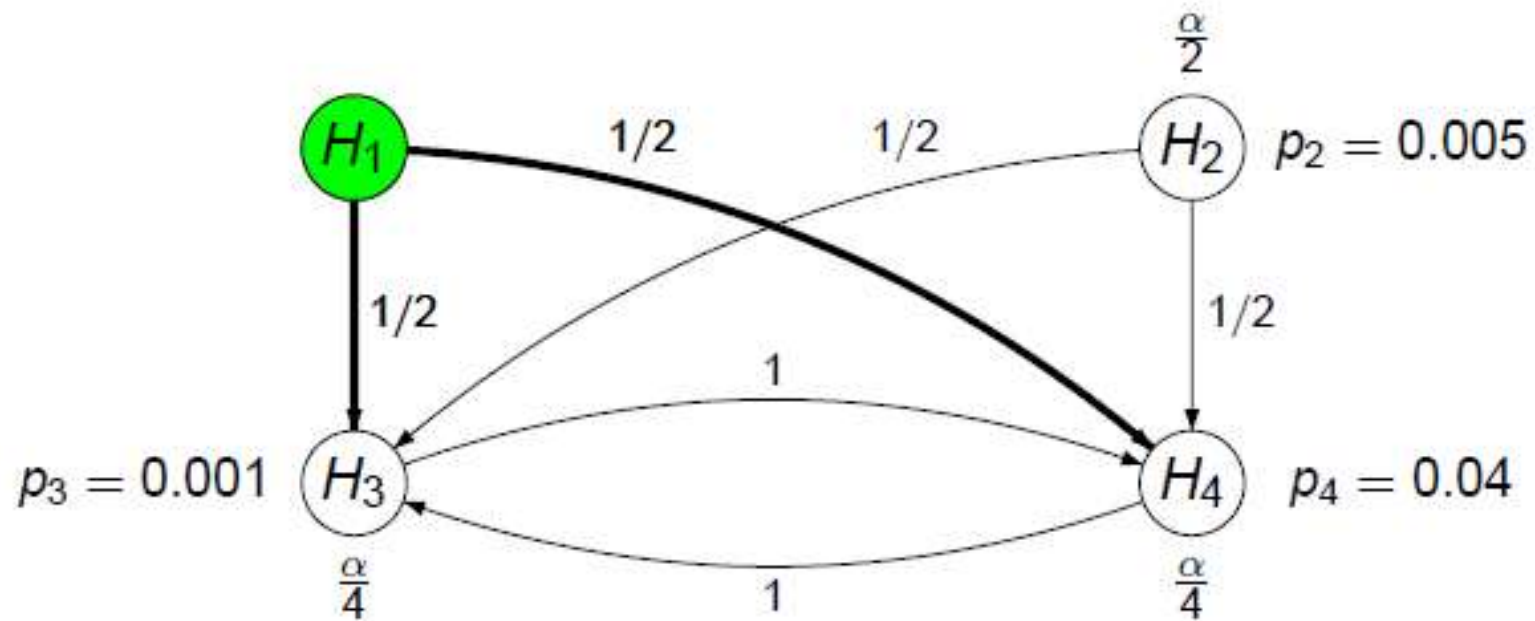
---



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

---

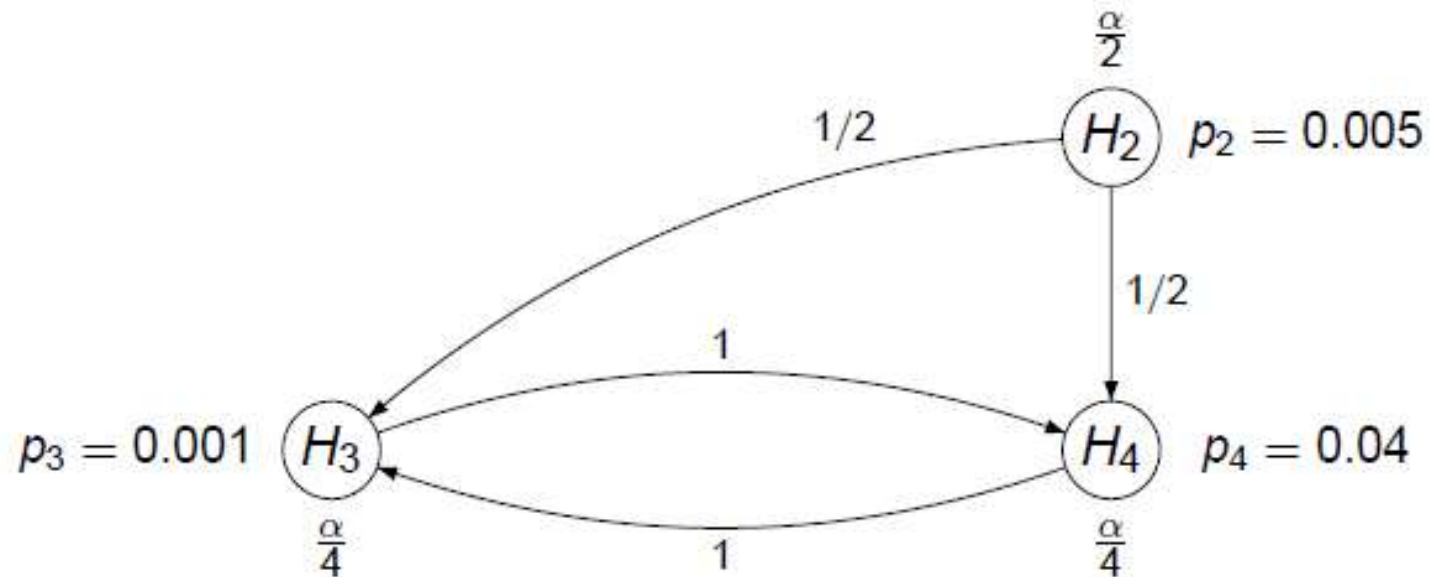




# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

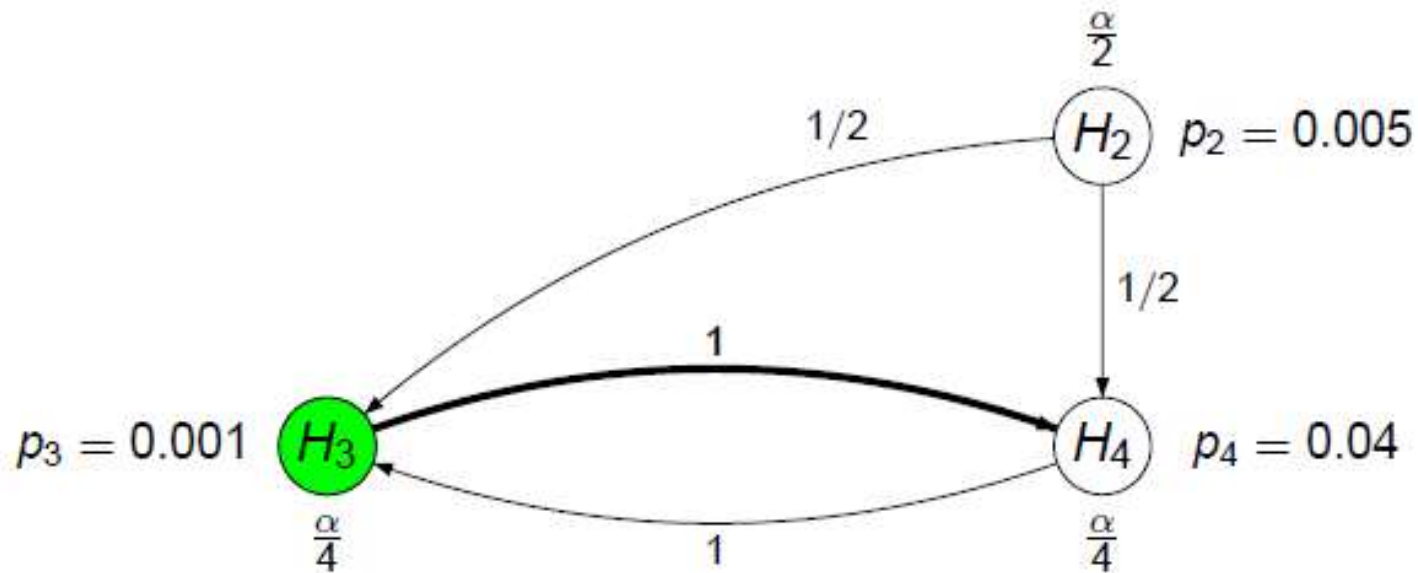
---



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

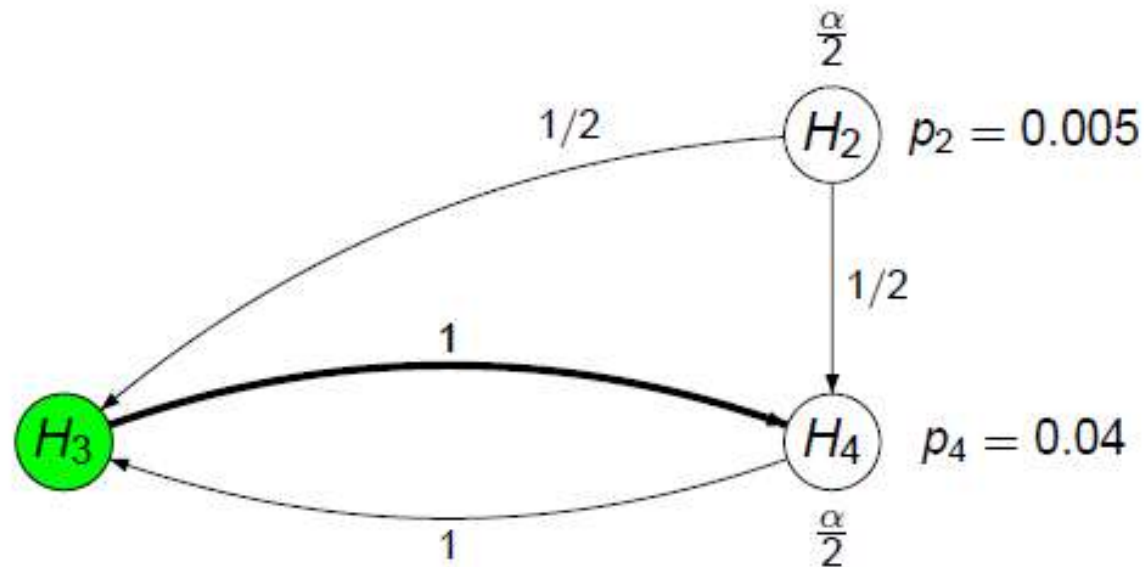
---



# Graphical Approach

Parallel gatekeeping – Example with  $\alpha = 0.025$

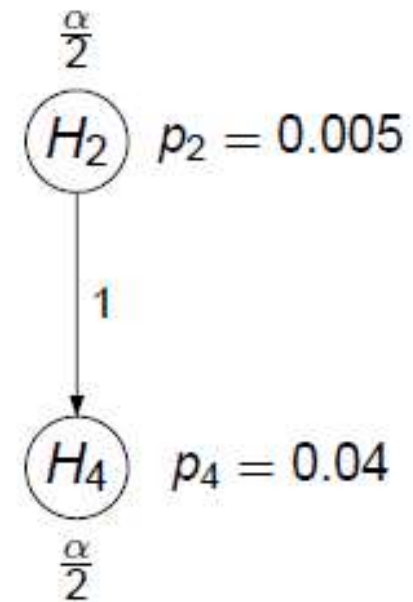
---



# Graphical Approach

*Parallel gatekeeping – Example with  $\alpha = 0.025$*

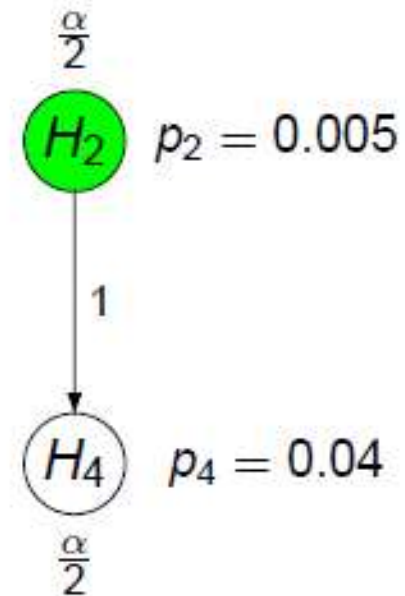
---



# Graphical Approach

*Parallel gatekeeping – Example with  $\alpha = 0.025$*


---



# Graphical Approach

*Parallel gatekeeping – Example with  $\alpha = 0.025$*

---



$p_4 = 0.04$   
 $\alpha$

# Outline

---

- Graphical approaches to multiple testing
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example revisited

# Graphical approach

## Formal definition

---

- Define

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

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  $\sum_{j=1}^m g_{ij} \leq 1$ ,  $\forall i = 1, \dots, m$

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



# Graphical approach

## Update algorithm

---

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

- 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:

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

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

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

- 3 Go to Step 1

# Graphical approach

## *Main result*

---

- The initial levels  $\alpha$ , the transition matrix  $G$ , and the algorithm define a unique sequentially rejective test procedure that controls the FWER at level  $\alpha$
  
- Remarks:
  - Any multiple test procedure derived and visualized by a graph  $(G, \alpha)$  is based on the closed test principle
  - The graph  $(G, \alpha)$  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

# Outline

---

- Graphical approaches to multiple testing
  - Conventions
  - Common multiple test procedures
  - Formal description
  - COPD example revisited

# COPD example revisited

## *Background*

---

- Objective: 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

Building a multiple test procedure: *Hypotheses*

---

primary

$H_1$

$H_2$

secondary

$H_3$

$H_4$

low dose

high dose

# COPD example revisited

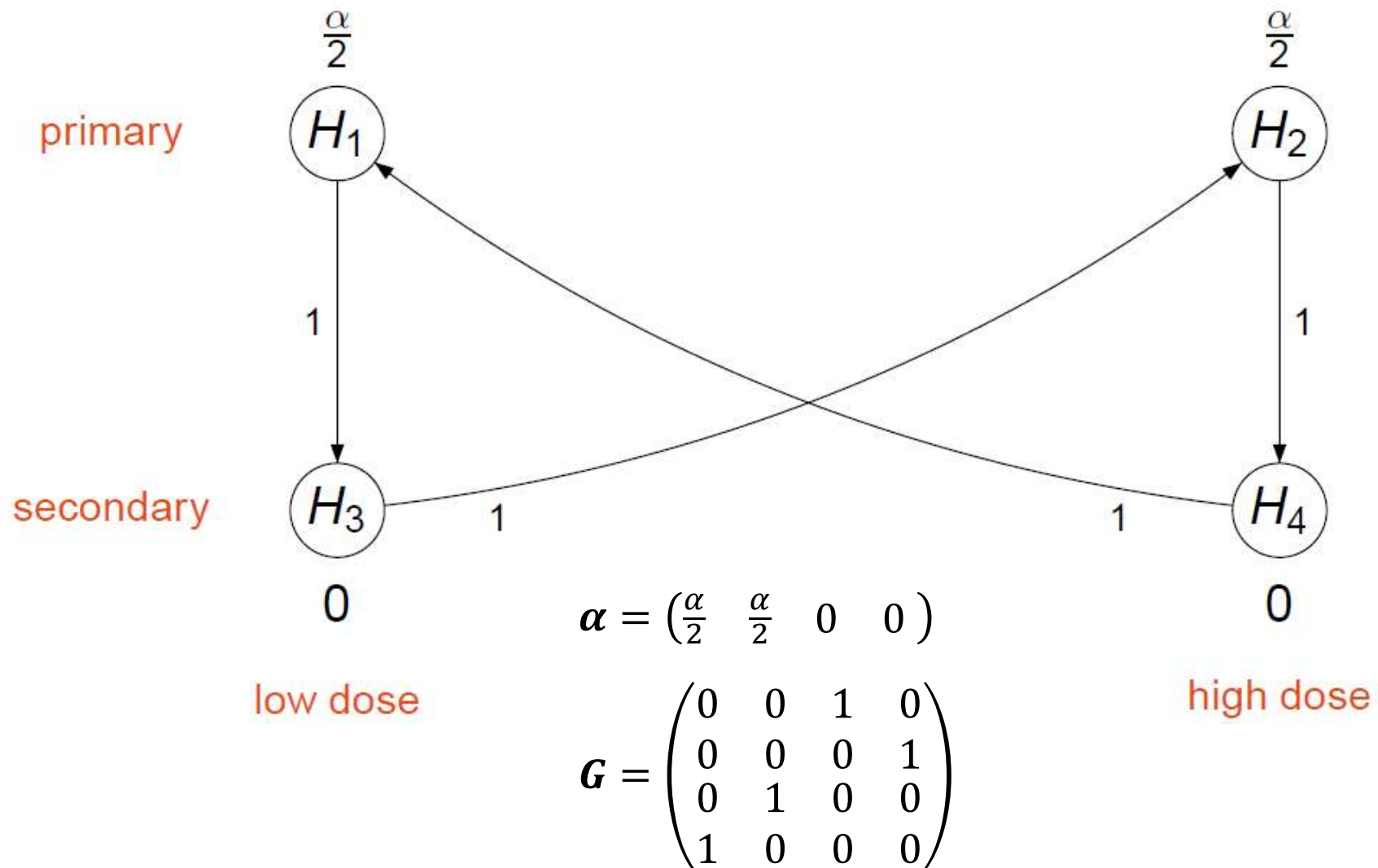
Building a multiple test procedure: *Initial levels  $\alpha$*

---



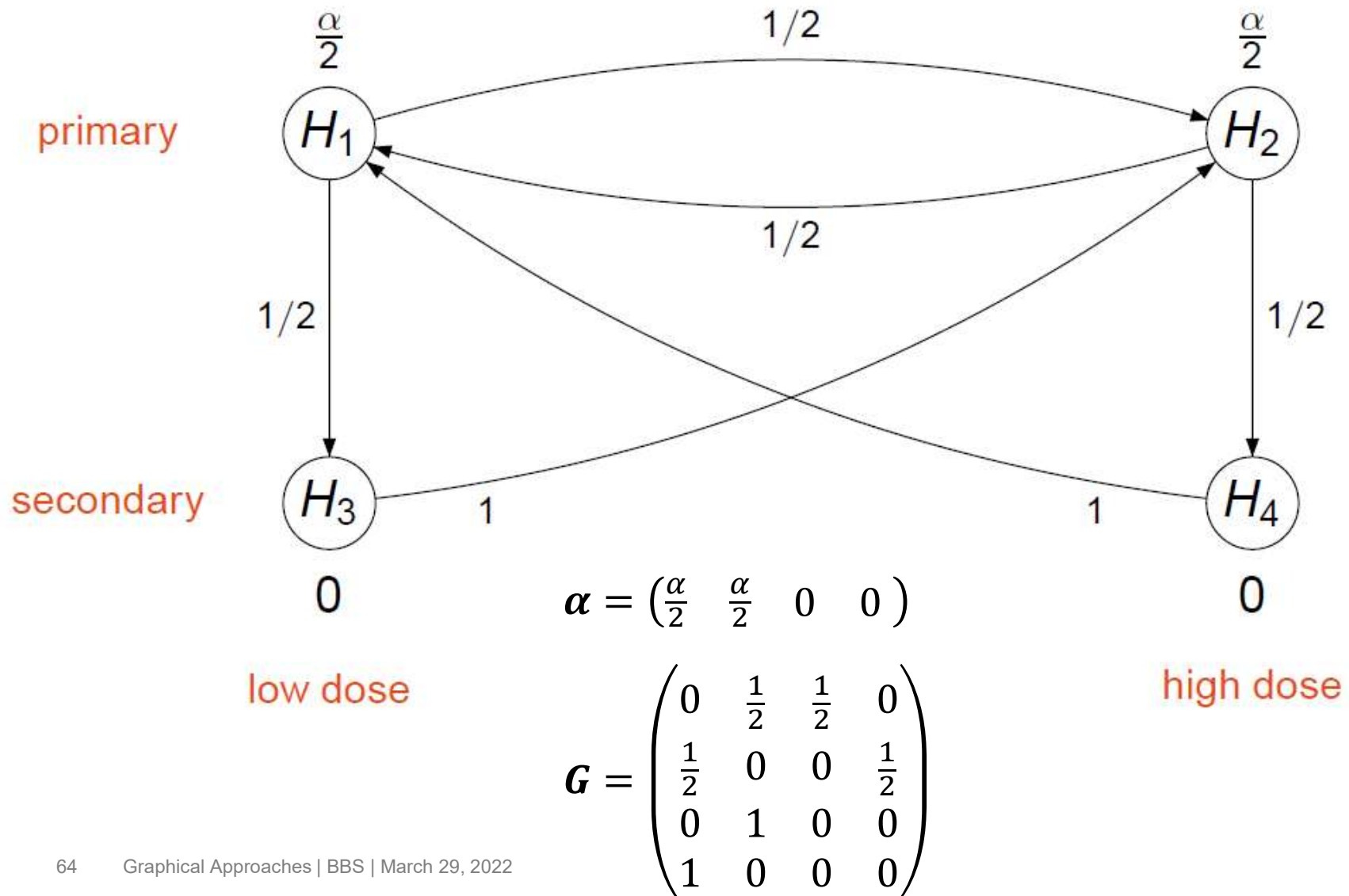
# COPD example revisited

Building a multiple test procedure:  $\alpha$ -propagation



# COPD example revisited

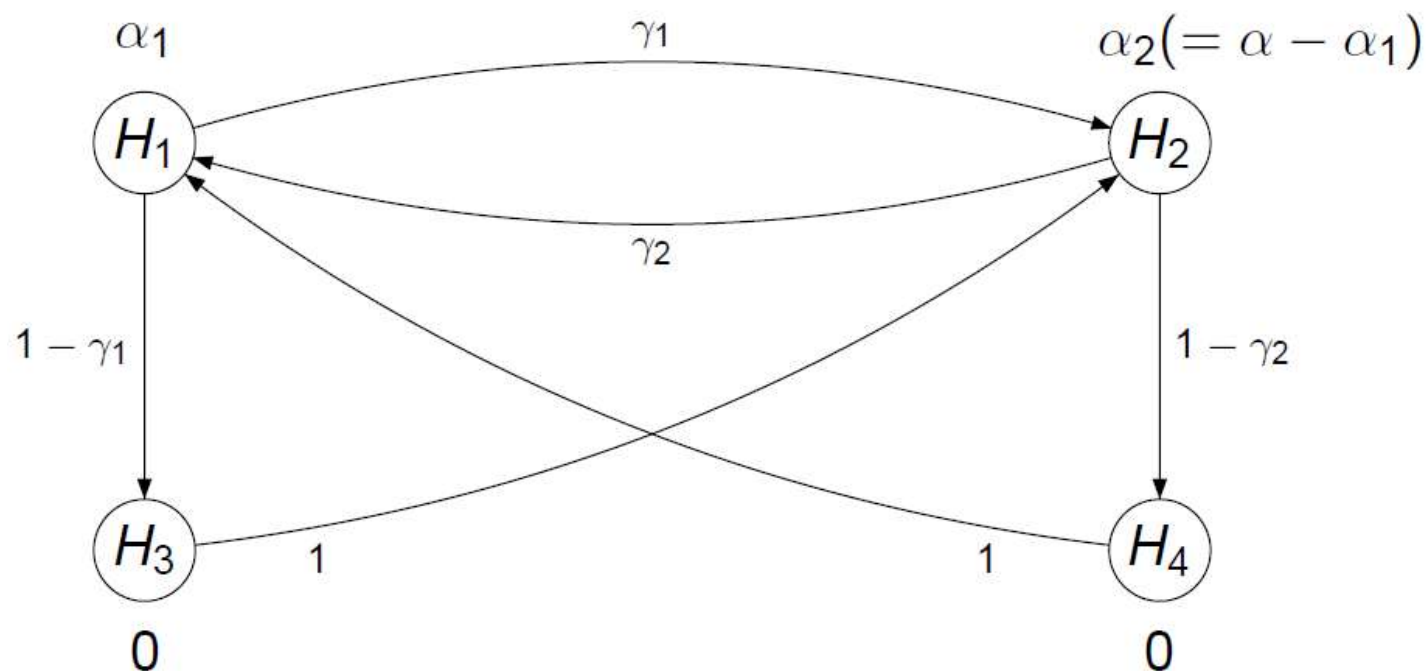
Building a multiple test procedure: *Alternative  $\alpha$ -propagation*





# COPD example revisited

Building a multiple test procedure: *General solution*



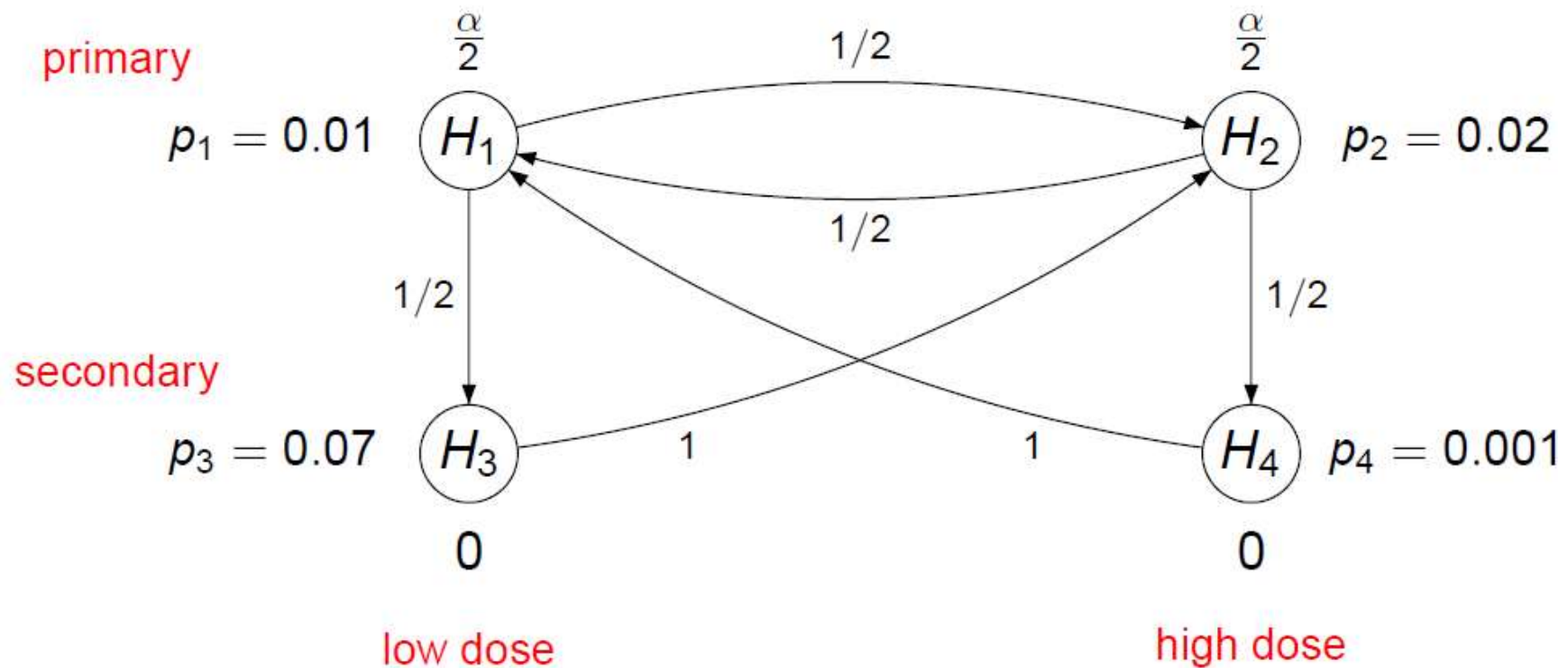
$$\alpha = (\alpha_1 \quad \alpha_2 \quad 0 \quad 0)$$

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

- Resulting graph depends on only three parameters  $\alpha_1$ ,  $\gamma_1$ , and  $\gamma_2$  that can be fine-tuned based on:
  - further clinical considerations, or
  - assumptions about effect sizes, correlations, ...

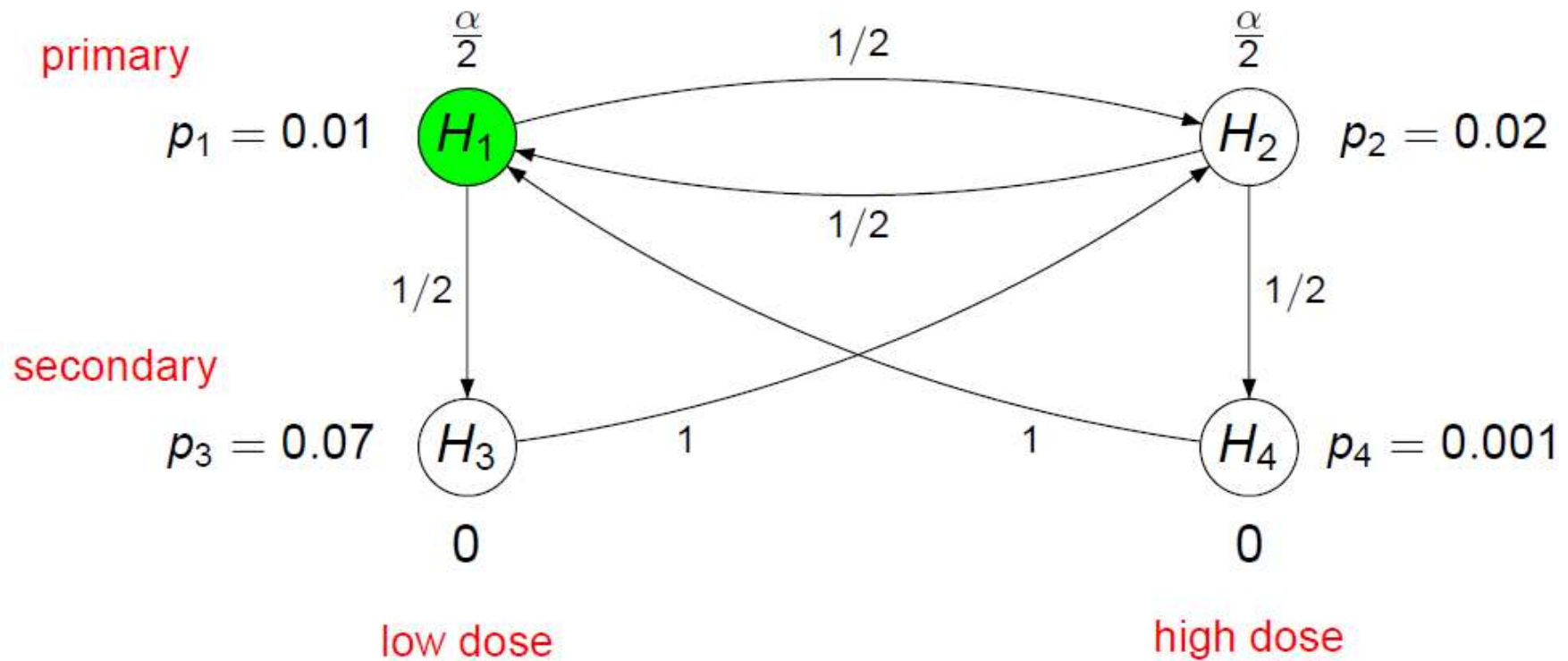
# COPD example revisited

Numerical example with  $\alpha = 0.025$



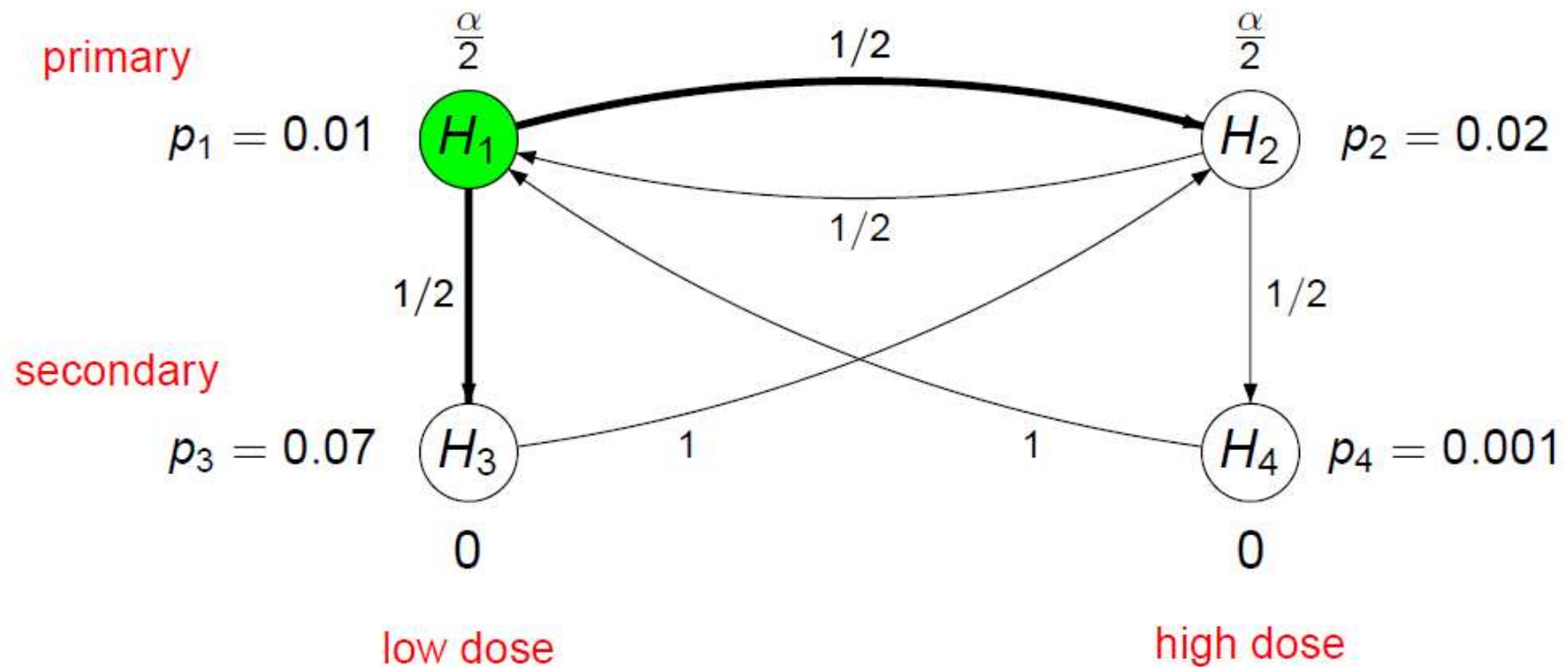
# COPD example revisited

Numerical example with  $\alpha = 0.025$



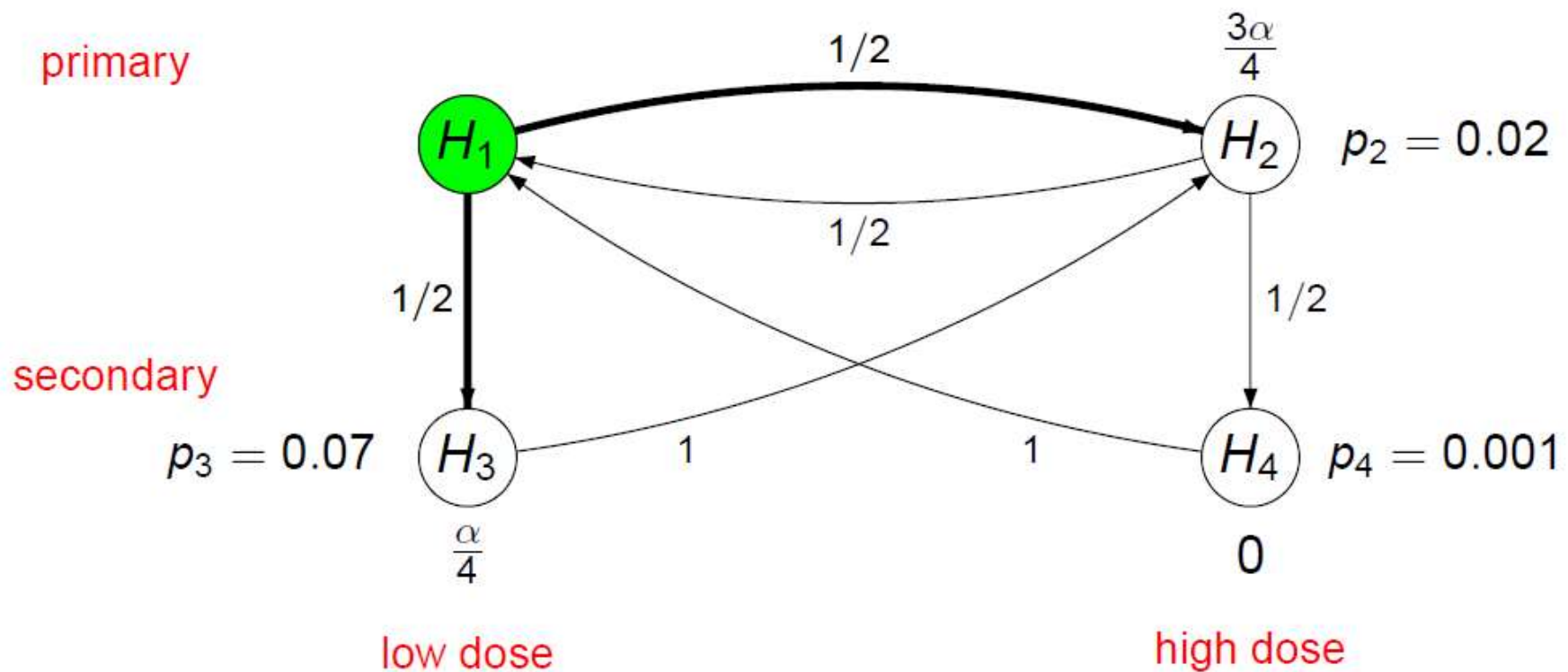
# COPD example revisited

Numerical example with  $\alpha = 0.025$



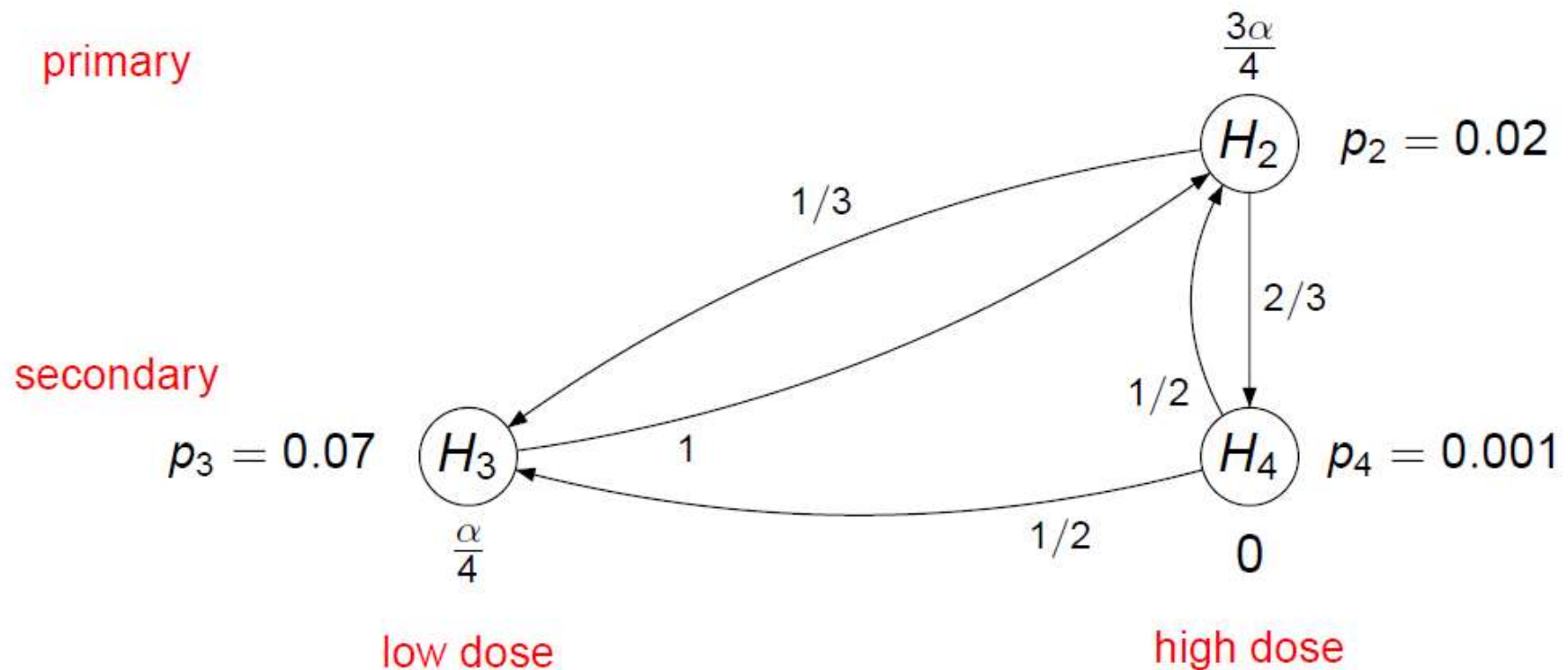
# COPD example revisited

Numerical example with  $\alpha = 0.025$



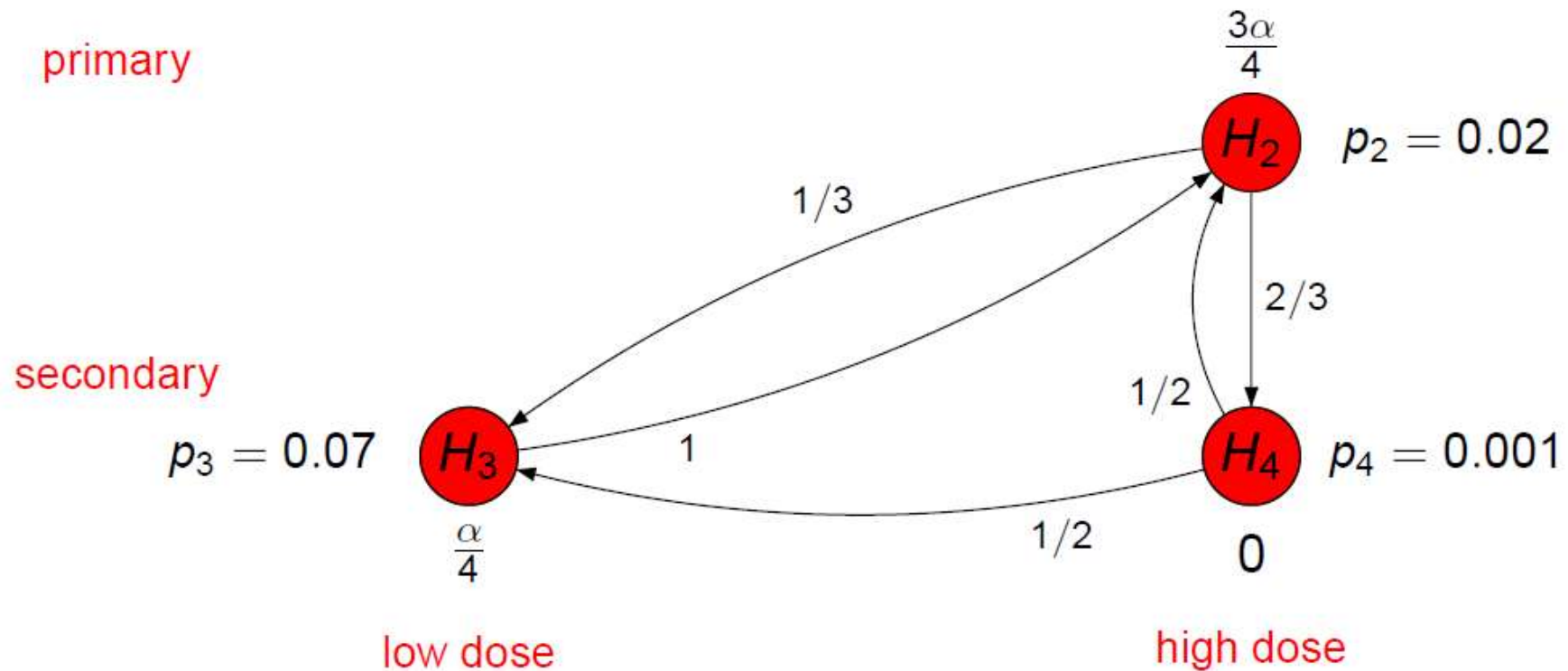
# COPD example revisited

Numerical example with  $\alpha = 0.025$



# COPD example revisited

Numerical example with  $\alpha = 0.025$



# 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)
   g: weights for the edges (n x n matrix)
   p: observed p-values (1 x n vector) */
START mcp(h, a, g, 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;
      g1 = J(n, n, 0);
      DO i = 1 TO n;
        a[i] = a[i] + a[rej]*g[rej,i];
        IF (g[i,rej]*g[rej,i]<1) THEN DO j = 1 TO n;
          g1[i,j] = (g[i,j] + g[i,rej]*g[rej,j])/(1 - g[i,rej]*g[rej,i]);
        END;
        g1[i,i] = 0;
      END;
      g = g1; g[rej,] = 0; g[,rej] = 0;
      a[rej] = 0;
      mata = mata // a;
    END;
    ELSE crit = 1;
  END;
  PRINT h; PRINT (ROUND(mata, 0.0001)); PRINT (ROUND(g,0.01));
FINISH;
```



# COPD example revisited

## SAS: Example call

```
PROC IML;
START mcp(h, a, g, p);
.....
FINISH;

/** Numerical example **/
h = {0      0      0      0      };
a = {0.0125 0.0125 0      0      };
g = {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, g, p);
QUIT;
```

The SAS System

h			
1	0	0	0

0.0125	0.0125	0	0
0	0.0188	0.0063	0

0	0	0	0
0	0	0.33	0.67
0	1	0	0
0	0.5	0.5	0

# 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

The screenshot displays the gMCP GUI 0.8.3 interface. The main window is titled "gMCP GUI 0.8.3" and contains a menu bar with "File", "Example graphs", "Analysis", "Extras", and "Help". Below the menu bar is a toolbar with icons for adding nodes, edges, and starting the test procedure. The central area shows a network diagram with four nodes: H1 (green circle, weight 1/2), H2 (white circle, weight 1/2), H3 (white circle, weight 0), and H4 (white circle, weight 0). Edges connect H1 to H2 (weight 0.5), H2 to H1 (weight 0.5), H1 to H3 (weight 0.5), H2 to H4 (weight 0.5), H3 to H4 (weight 1), and H4 to H3 (weight 1). The "Adjacency Matrix" table is shown on the right, and the "Hypothesis Weight P-Value" table is shown below it. The "Total  $\alpha$ " is set to 0.025. The "No Information about correlations" option is selected.

	H1	H2	H3	H4
H1	0	0.5	0.5	0
H2	0.5	0	0	0.5
H3	0	1	0	0
H4	1	0	0	0

Hypothesis	Weight	P-Value	
H1	1/2	0.01	Reject and pass $\alpha$
H2	1/2	0.02	Reject and pass $\alpha$
H3	0	0.07	Reject and pass $\alpha$
H4	0	0.001	Reject and pass $\alpha$

Sum of weights: 1; Load p-values from R

Total  $\alpha$ : 0.025

No Information about correlations

Select an R correlation matrix (No 4x4-matrices found.)

Correlation applicable for Simes test (new feature that still needs testing)

# Summary

---

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

# Graphical Approach

## *Summary*

---

- Extensions available to address other problems
  - Adjusted p-values and simultaneous confidence intervals available
  - Power considerations
  - Weighted and trimmed Simes tests
  - Weighted parametric test procedures to exploit correlation
  - Families of hypotheses
  - Convex combination of graphs to introduce “memory” (including truncated procedures)
  - Group-sequential and adaptive designs
  - Symmetric graphs (including Hochberg procedure)
  - Graphical test procedures controlling generalized error rates

Q & A

---

**Any questions?**

# Agenda

---

14:00 – 14:45

**Introduction to multiple testing**  
*Dong Xi*

14:45 – 16:15

**Graphical approaches to multiple testing**  
*Frank Bretz*

Break

16:30 – 17:30

**Extensions to group sequential designs**  
*Ekkehard Glimm*

17:30 – 18:00

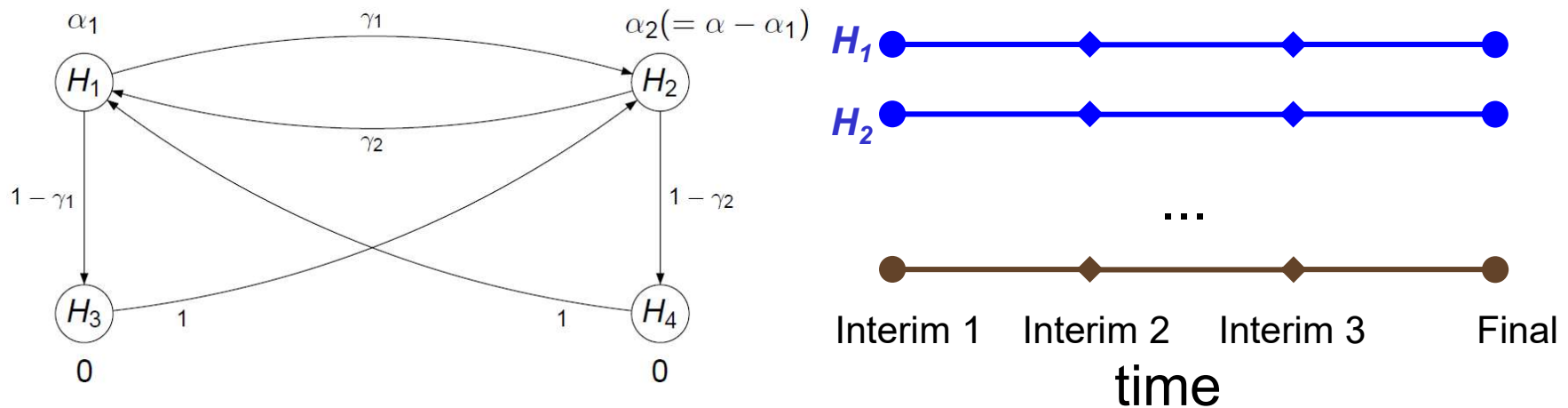
**Extensions to pooled analyses from two studies**  
*Dong Xi*

# Problem Statement

Combine multiplicity adjustment for multiple endpoints, multiple treatment arms, multiple subpopulations, ...

with

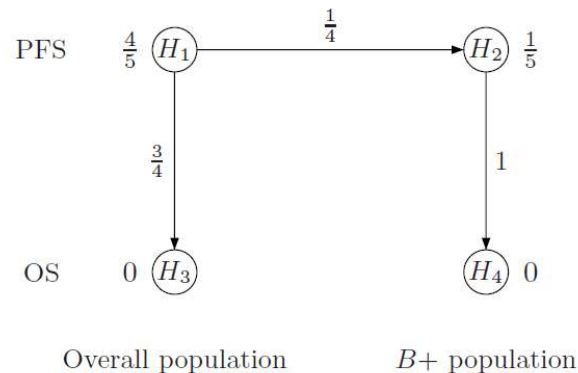
repeated testing in the framework of a group-sequential design.



# General principle we will follow

- Top layer:  
Design the multiplicity-adjustment method ignoring repeated testing for the moment

- E.g. the graphical procedure



- Bottom layer:  
Devise an alpha-spending approach for the hypotheses and all  $\alpha$ -levels which occur in the closed test procedure.

(Xi, Glimm, Bretz, 2016)



## Short recap of group-sequential testing

---

- Hypothesis  $H_0$  is tested repeatedly in time at times  $I_1, \dots, I_F$ .
- $H_0$  will be rejected if  $t_i \geq c_i$  (or alternatively if  $t_i \leq c_i$ ) at at least one time  $I_i$ 
  - $t_i$  is observation of test statistic  $T_i$  (e.g. a t-test statistic or a p-value) calculated from the data available up to time  $I_i$
  - $c_i$  are critical values fulfilling  $P_{H_0}(T_1 \geq c_1 \text{ or } T_1 \leq c_1 \text{ or } \dots T_F \geq c_F) \leq \alpha$
- Repeated testing poses a multiplicity problem, but there is just one hypothesis, so decision space is much simpler ( $H_0$  is either true or false).

## Short recap of group-sequential testing

---

- "Time" in this context refers to information time
  - In "conventional" trials: number of patients recruited
  - In time-to-event trials: number of events accrued
  - In general: information fraction, ratio of variance of parameter estimate at interim and final
- Very common assumption ("canonical distribution"):
  - i.  $T_1, \dots, T_F$  are multivariate normal
  - ii.  $T_i \sim N(\sqrt{I_i}\theta, 1)$
  - iii.  $\text{corr}(T_i, T_j) = \sqrt{I_i/I_j}$  for  $i \leq j$

holds asymptotically under relatively mild assumptions (Scharfstein et al., 1997; Jennison and Turnbull, 1997), e.g. for ML-estimates.

## Short recap of group-sequential testing

---

- Typically, we know  $I_1, \dots, I_F$  in advance
  - e.g. we planned IAs after 50, 100 and 200 patients
- Or we can condition on their observations
  - e.g. we plan IAs after 6, 12 and 24 months and condition on the number of events observed up to then
- This knowledge can be used to find the critical values  $c_i$  such that  $P_{H_0}(T_1 \geq c_1 \text{ or } T_1 \geq c_1 \text{ or } \dots T_F \geq c_F) = \alpha$ .
- As there are infinitely many solutions, additional restrictions are needed ("alpha-spending rules")
  - Most common are the Pocock and the O'Brien-Fleming Lan-deMets-alpha-spending approaches, but there are many more.

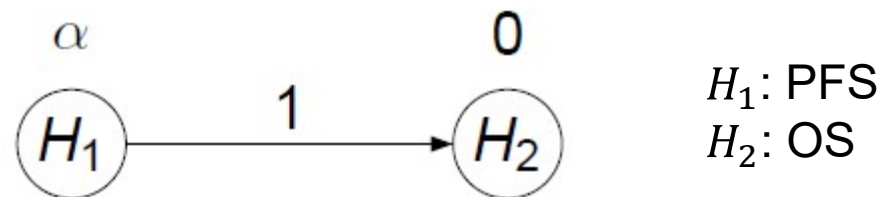
# Multiplicity + group-sequential

---

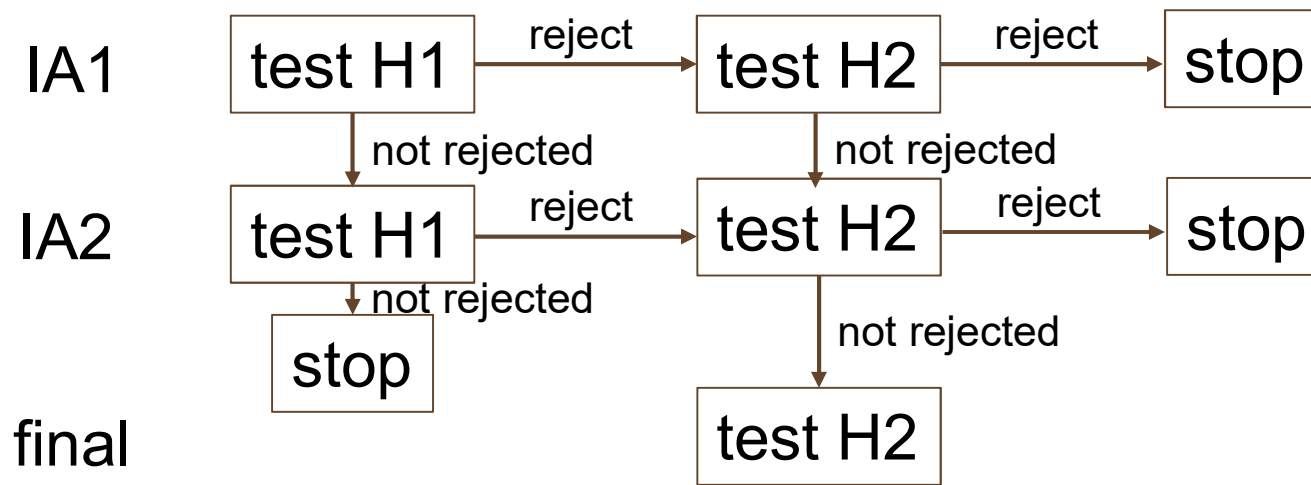
- Top layer:  
Design the multiplicity-adjustment method with a graph
- Bottom layer:  
Devise an alpha-spending approach for the hypotheses and all the  $\alpha$ -levels which occur in the closed test procedure defined by the graph.
- Whenever a hypothesis  $H_i$  is rejected (no matter when), it gives its  $\alpha$  to other hypotheses (according to the graphical procedure)
  - For  $\alpha$ -propagation, we ignore the group-sequential aspect

# Multiplicity + group-sequential

- A simple example: hierarchical testing of PFS and OS



Within this setup: Two interim analyses, 1 final



# Hierarchical testing of PFS and OS

---

- IA1 after 150 PFS events, IA2 after 300 PFS events
- Final after 200 OS events
- Information fractions
  - PFS: 0.5, 1
  - OS: 75, 150, 200 / 200 = 0.375, 0.75, 1 (estimated #OS events at IAs)
- Alpha-spending: OBF for PFS, Pocock for OS  $\Rightarrow$  critical values for the p-value:

	IA 1	IA 2	F
<b>PFS</b>	0.0015	0.0245	0
<b>OS (PFS not sign.)</b>	0	0	0
<b>OS (PFS signifkant)</b>	0.0124	0.0117	0.0100

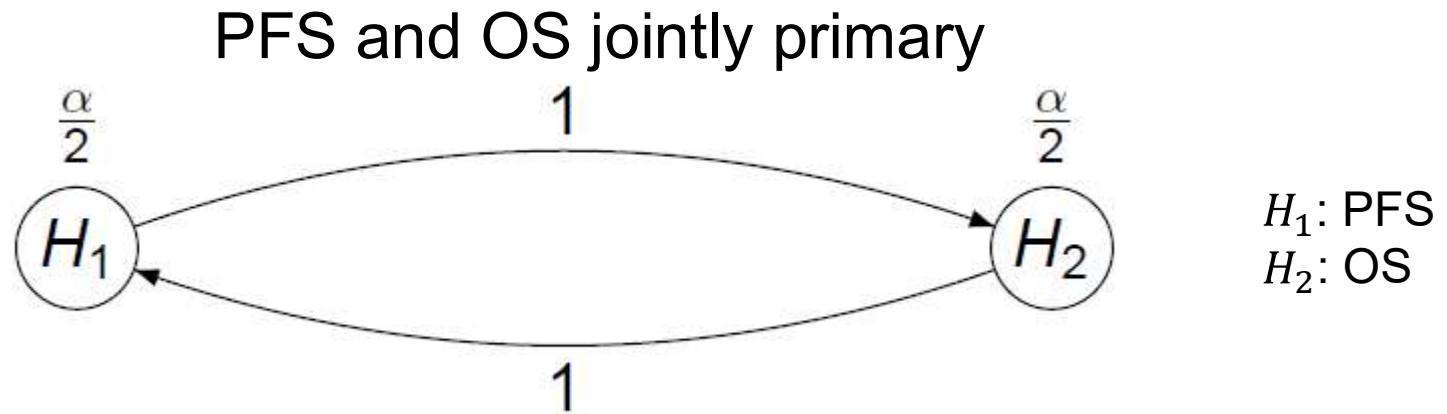
# Hierarchical testing of PFS and OS

---

## Remarks:

- The approach uses the known correlation between stage-wise test statistics, but *not* between PFS and OS
  - In the hierarchical procedure,  $\text{corr}(\text{PFS}, \text{OS})$  does not play a role.
- "Looking back" is allowed
  - If  $H_1$  is not rejected at IA1, rejected at IA2,  $H_2$  not rejected at IA2, we are allowed to "retest"  $H_2$  at IA1 at the level 0.0124. This preserves the FWER.  
... but does it make sense? (Some debate, see e.g. Tamhane et al., 2021)
- If in practice, observed OS events diverge, we recalculate
  - e.g. if 65, 160, 200 OS events observed, use 0.0111; 0.0133; 0.0097

# Modification of the example



## Critical values

(OBF-Lan/deMets for PFS, PK-Lan/deMets for OS)

	IA 1	IA 2	F
<b>PFS at <math>\alpha/2</math></b>	0.0004	0.0124	0
<b>PFS at <math>\alpha</math></b>	0.0015	0.0245	0
<b>OS at <math>\alpha/2</math></b>	0.0062	0.0056	0.0046
<b>OS at <math>\alpha</math></b>	0.0124	0.0117	0.0100



# Joint testing of PFS and OS

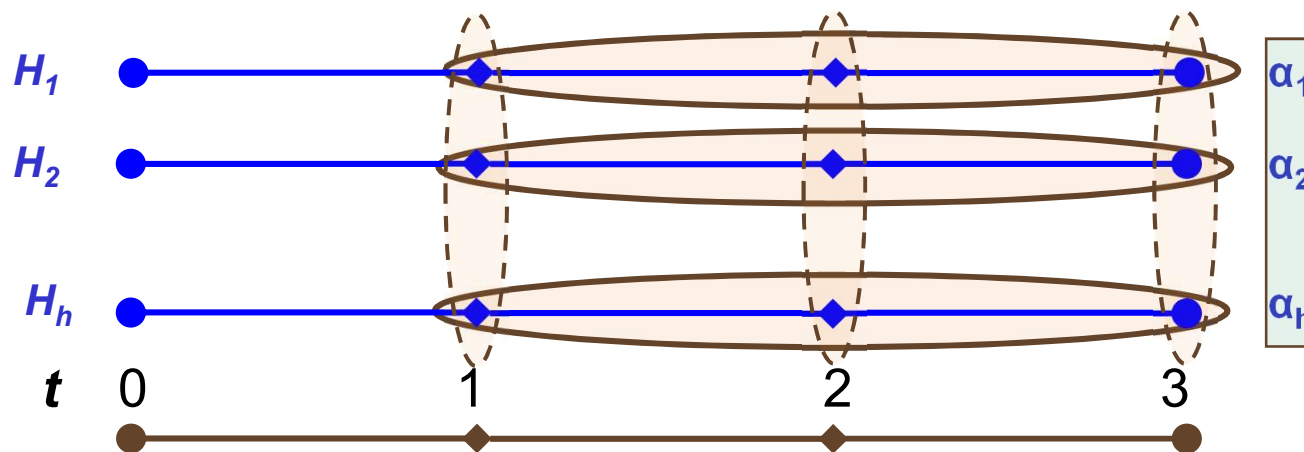
---

## Remarks:

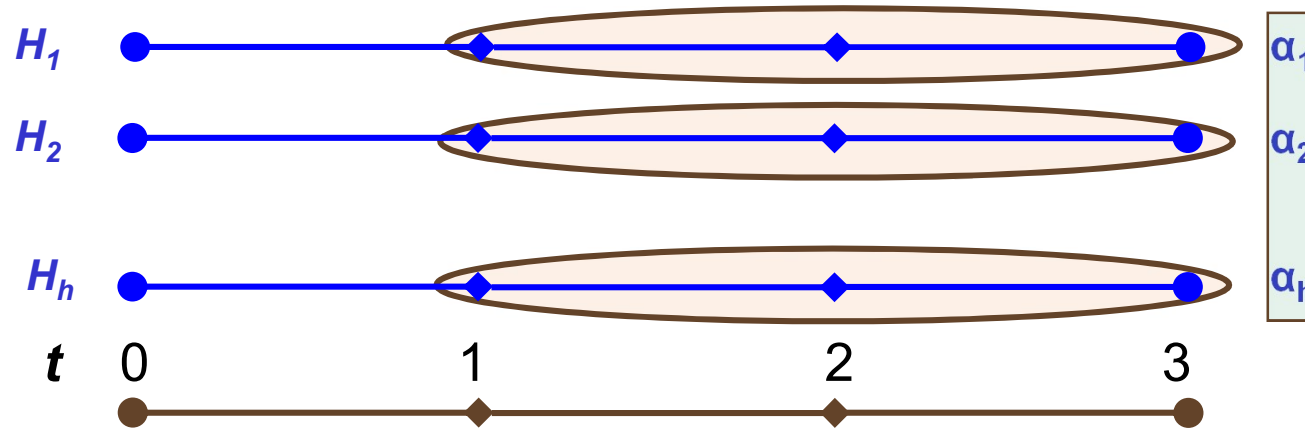
- Alpha-spending for  $H_1, H_2, \dots$  does not have to be the same
- Alpha-spending  $c_{i,k}(\alpha_{ij})$  for the different levels  $\alpha_{ij}$  arising in the top-layer multiple test procedure of  $H_i$  also does not have to be the same  
( $k$  is stage,  $\{\alpha_{i1}, \alpha_{i2}, \dots\}$  is set of all levels which can arise for  $H_i$  in the graph)
  - e.g., we could have used Pocock for OS at  $\alpha/2$  and then levels (0.0062, 0.0056, 0.0198) for OS at  $\alpha$   
$$c_{2,1}\left(\frac{\alpha}{2}\right) = 0.0062; c_{2,2}\left(\frac{\alpha}{2}\right) = 0.0056$$
- But we must obey the condition  $c_{i,k}(\alpha_{ij}) \leq c_{i,k}(\alpha_{ij'})$  for all  $\alpha_{ij} \leq \alpha_{ij'}$  and  $k$  (Maurer and Bretz, 2013)
  - e.g. mustn't use Pocock for OS at  $\alpha/2$  and then switch to OBF at  $\alpha$

## Several primary endpoints

- In group-sequential trials, correlation between stages is known:
  - $\sqrt{n_i/n_j}$  between stages  $i, j$  with non-TTE-endpoints and equal group sizes
  - $\sqrt{i_j/i_j}$  between stages  $i, j$  with TTE-endpoints ( $i_j$  information fraction of stage  $i$ )
- Occasionally correlation between endpoints is also known:
  - In practice usually only if primary endpoints pertain to several doses or regimens compared with a common control.



# Several primary endpoints: endpoint correlation unknown or not exploited



Strategy: "Bonferroni on hypotheses", then GS.

- Some improvements with partial knowledge of correlations between hypotheses are possible (e.g. Maurer et al., 2011)
  - A personal caveat: Don't try GS-splitting on full  $\alpha$ , then "bonferronize" GS-alphas (i.e. reverse top and bottom layer).
    - Becomes very complicated very quickly.
    - No power gains.
- ☞ Not "wrong", but also not worth the trouble.

## Several primary endpoints: correlation known

---

- Endpoints  $P$ ,  $S$  in 2 stages, normally distributed test statistics:  
Any set of critical values  $(c_{1P}, c_{1S}, c_{2P}, c_{2S})$  with  
 $1 - \Phi(c_{1P}, c_{1S}, c_{2P}, c_{2S}) = \alpha$ ,  
gives a valid test controlling the multiple level at  $\alpha$ .  
  
 $\Phi(c_{1P}, c_{1S}, c_{2P}, c_{2S})$  cdf of multivariate normal distribution with means 0, variances 1 and the known correlations
  - Equally important endpoints:  $c_{iP} = c_{iS}$
  - „Pocock-like“:  $c_{1P} = c_{2P}$
  - „O’Brien-Fleming-like“:  $c_{1P} = c_{2P} / \sqrt{\text{stage-1-info-fraction}}$
- Can be done sequentially:  
If one of  $P$  stage 1,  $S$  stage 1,  $P$  stage 2,  $S$  stage 2 is significant, cross out the corresponding endpoint  $P$  or  $S$  and apply the resulting univariate GS test to the remaining endpoint at full  $\alpha$  (as described on previous slides).

## Several primary endpoints: correlation known

---

- In theory, we could walk through the closure defined by the graph, calculate critical values for each intersection arising in it.
- Condition  $c_{i,k}(\alpha_{ij}) \leq c_{i,k}(\alpha_{ij'})$  for all  $\alpha_{ij} \leq \alpha_{ij'}$ ,  $k$  must be kept (if small values of the test statistic lead to rejection, otherwise reverse).
- In practice, this is complicated, the advantages of the graphical procedure are partly lost (see Bretz et al., 2011, Xi et al., 2017).
- For really complex cases with multiple sources of multiplicity (e.g. several doses, several endpoints and group-sequential testing), we usually do not know all correlations.
- Further literature on GS + (partly) known correlations: Tamhane et al., 2021; Anderson et al., 2021.

## Some applications

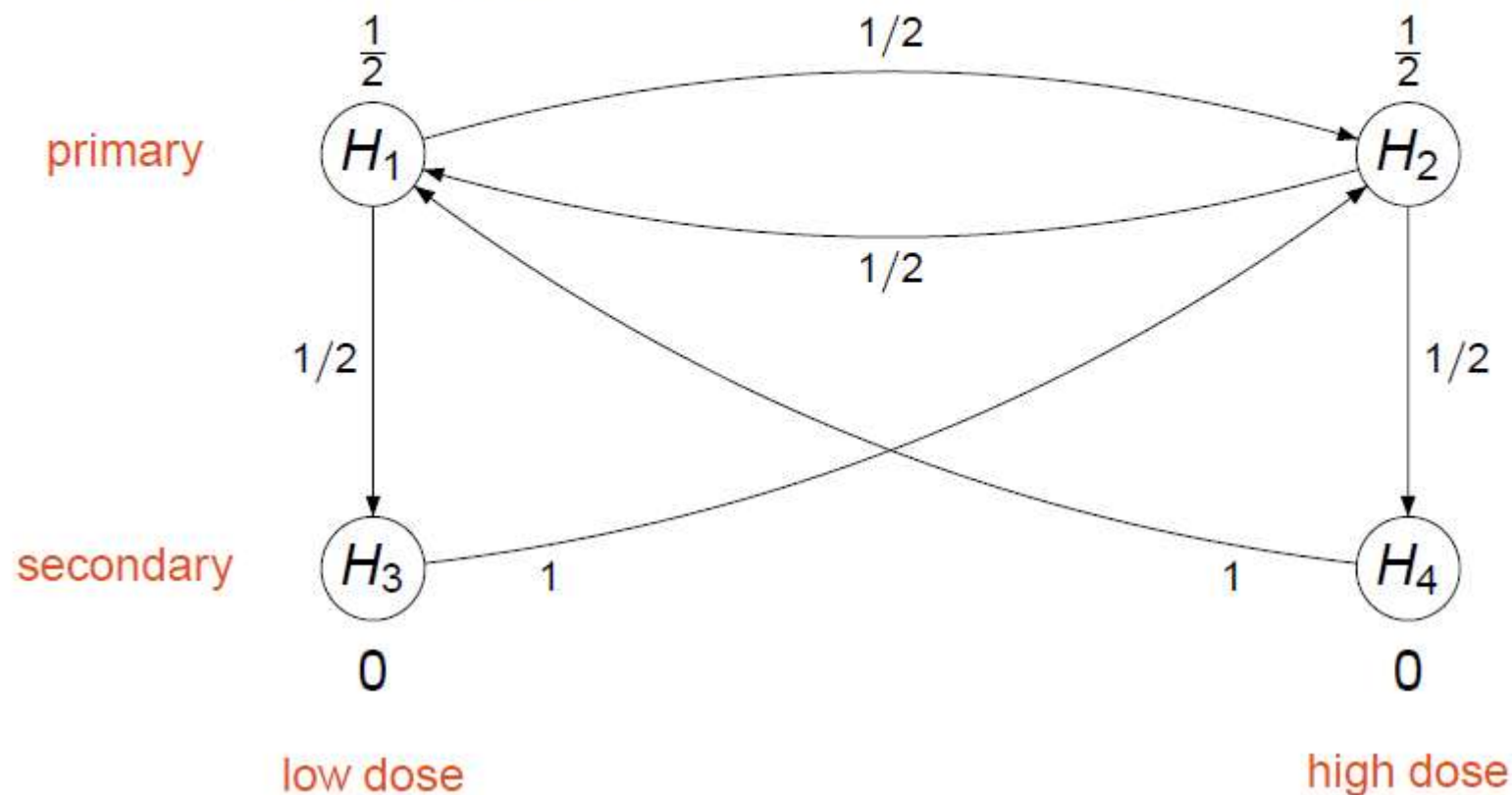
---

- A more complex example
- Matching interim alphas with desired decisions I
- Matching interim alphas with desired decisions II
- Some traps to avoid

# Example 1: two endpoints, two doses

*Study begin*

- Two interim and one final analysis, equally spaced in time
- O'Brien-Fleming-type spending function with  $\alpha = 0.025$
- Test procedure:



# Example 1

First interim analysis ( $t = 1$ )

observed p-values and critical values (p-value scale),  
information fraction (IF)=1/3

$i$	$p_{i,1}$	$\alpha_{i,1}^*(\alpha W_i(\{1, 2, 3, 4\}))$
1	0.0062	0.00002
2	0.017	0.00002
3	0.009	0
4	0.13	0

Critical value  $H_1$  at  $\frac{\alpha}{2}$ ;  
OBF split at  $IF=\frac{1}{3}$

No rejection



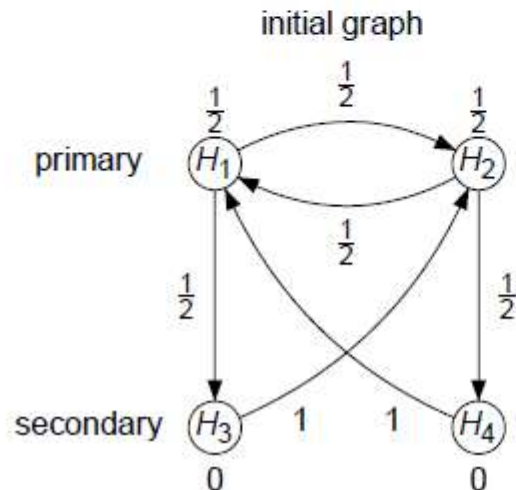
# Example 1

Second interim analysis ( $t = 2$ )

observed p-values and critical values (p-value scale),  
information fraction (IF)=2/3

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha W_i(\{1, 2, 3, 4\}))$
1	0.0002	0.0022
2	0.0035	0.0022
3	0.002	0
4	0.06	0

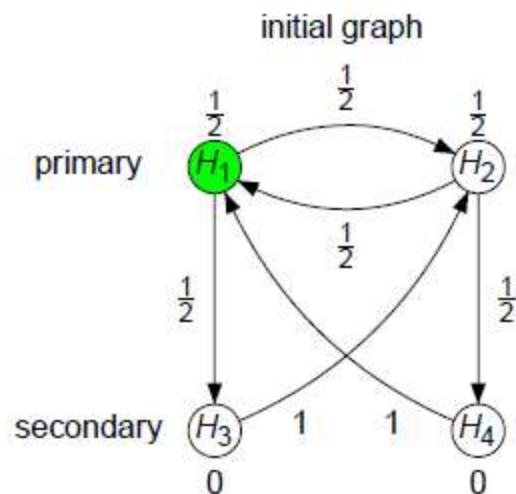
Critical value  $H_1$  at  $\frac{\alpha}{2}$ ;  
OBF split at  $IF = \frac{2}{3}$



# Example 1

Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$
1	0.0002	0.0022
2	0.0035	0.0022
3	0.002	0
4	0.06	0

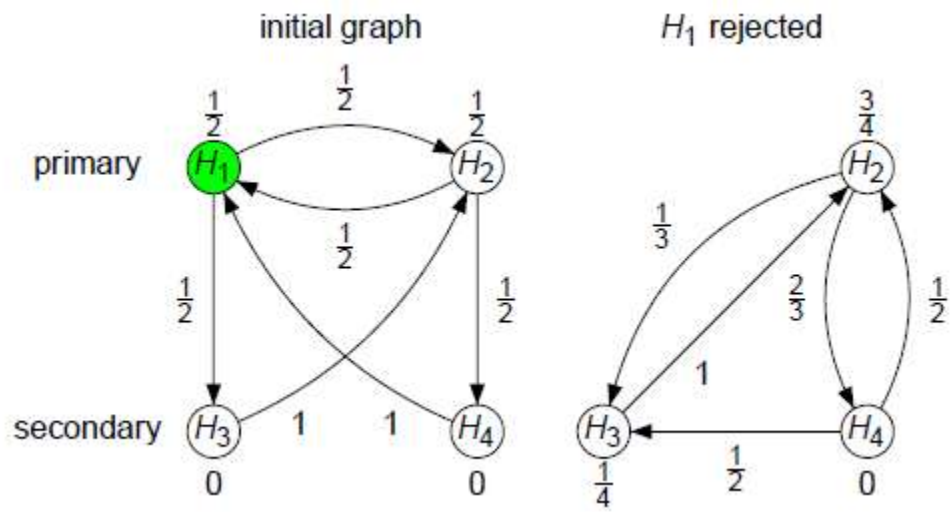


# Example 1

## Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$	$\alpha_{i,2}^*(w_i(\{2, 3, 4\})\alpha)$
1	0.0002	0.0022	0
2	0.0035	0.0022	0.004
3	0.002	0	0.0008
4	0.06	0	0

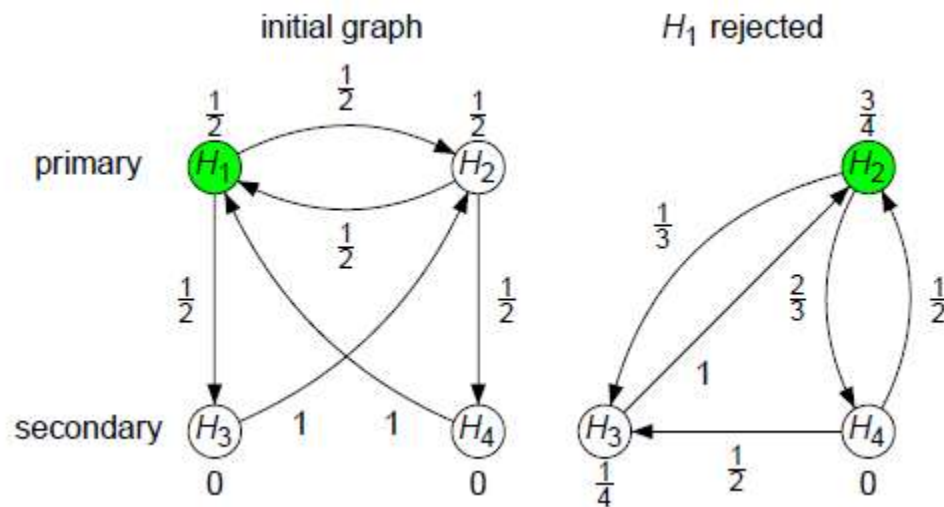
Critical value  $H_2$  at  $\frac{3\alpha}{4}$ ;  
 OBF split at  $IF = \frac{2}{3}$



# Example 1

Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$	$\alpha_{i,2}^*(w_i(\{2, 3, 4\})\alpha)$
1	0.0002	0.0022	—
2	0.0035	0.0022	0.004
3	0.002	0	0.0008
4	0.06	0	0

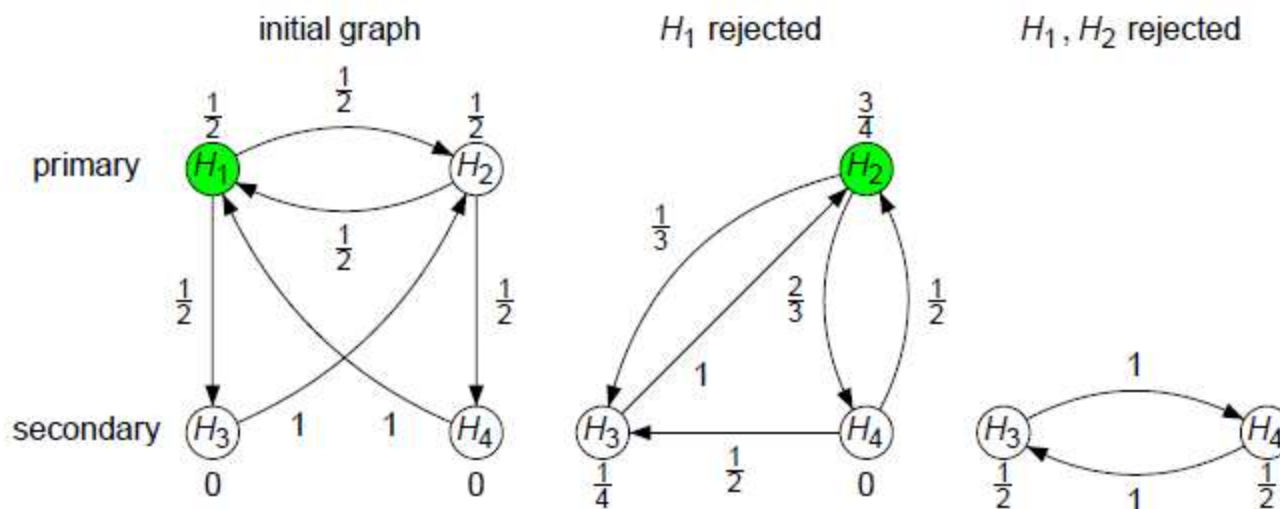


# Example 1

Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$	$\alpha_{i,2}^*(w_i(\{2, 3, 4\})\alpha)$	$\alpha_{i,2}^*(w_i(\{3, 4\})\alpha)$
1	0.0002	0.0022	-	-
2	0.0035	0.0022	0.004	-
3	0.002	0	0.0008	0.0022
4	0.06	0	0	0.0022

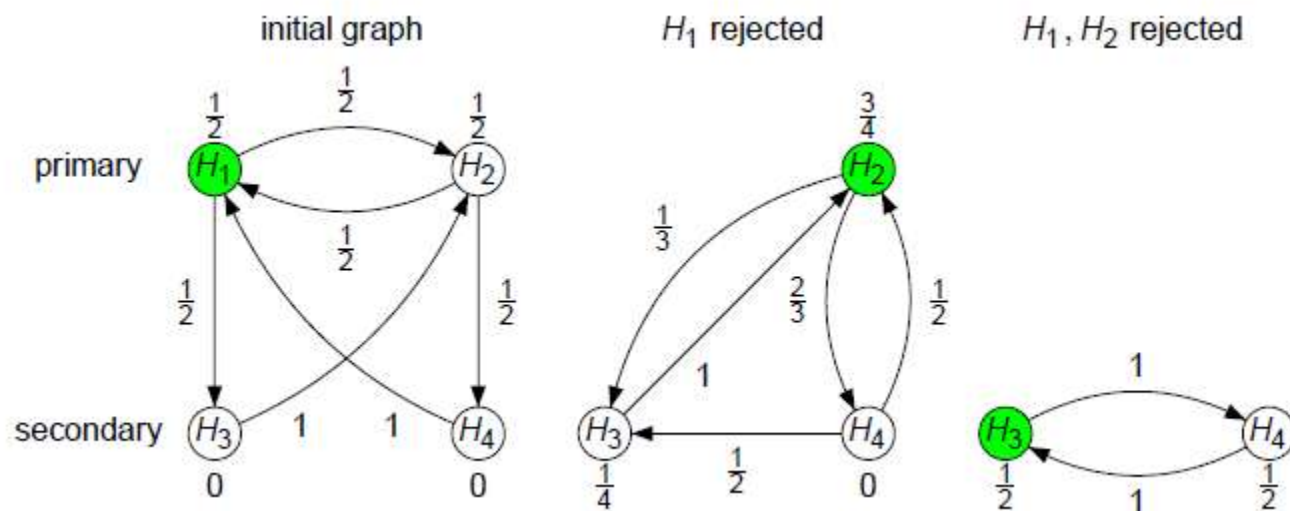
Critical value  $H_3$  at  $\frac{\alpha}{2}$ ;  
 OBF split at  $IF = \frac{2}{3}$



# Example 1

Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$	$\alpha_{i,2}^*(w_i(\{2, 3, 4\})\alpha)$	$\alpha_{i,2}^*(w_i(\{3, 4\})\alpha)$
1	0.0002	0.0022	–	–
2	0.0035	0.0022	0.004	–
3	0.002	0	0.0008	0.0022
4	0.06	0	0	0.0022

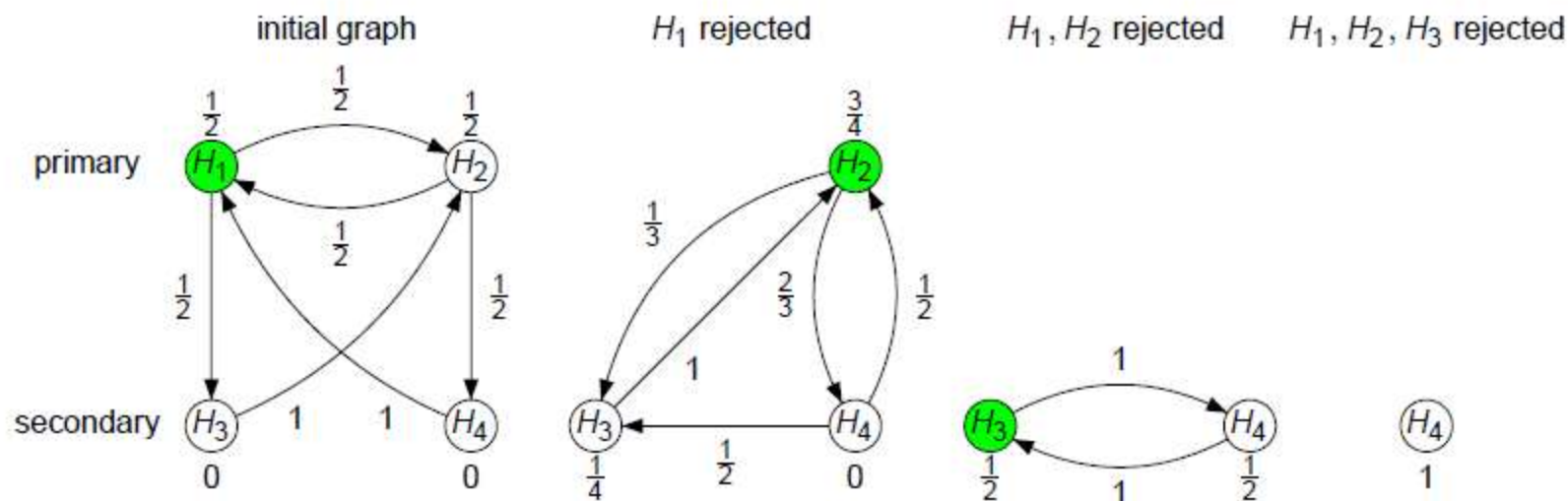


# Example 1: Decision

## Second interim analysis ( $t = 2$ )

$i$	$p_{i,2}$	$\alpha_{i,2}^*(\alpha w_i(\{1, 2, 3, 4\}))$	$\alpha_{i,2}^*(w_i(\{2, 3, 4\})\alpha)$	$\alpha_{i,2}^*(w_i(\{3, 4\})\alpha)$	$\alpha_{i,2}^*(w_i(\{4\})\alpha)$
1	0.0002	0.0022	—	—	—
2	0.0035	0.0022	0.004	—	—
3	0.002	0	0.0008	0.0022	—
4	0.06	0	0	0.0022	0.006

Critical value  $H_4$  at  $\alpha$ ;  
 OBF split at  $IF = \frac{2}{3}$



Decision to stop the trial

## Example 2: matching interim alphas

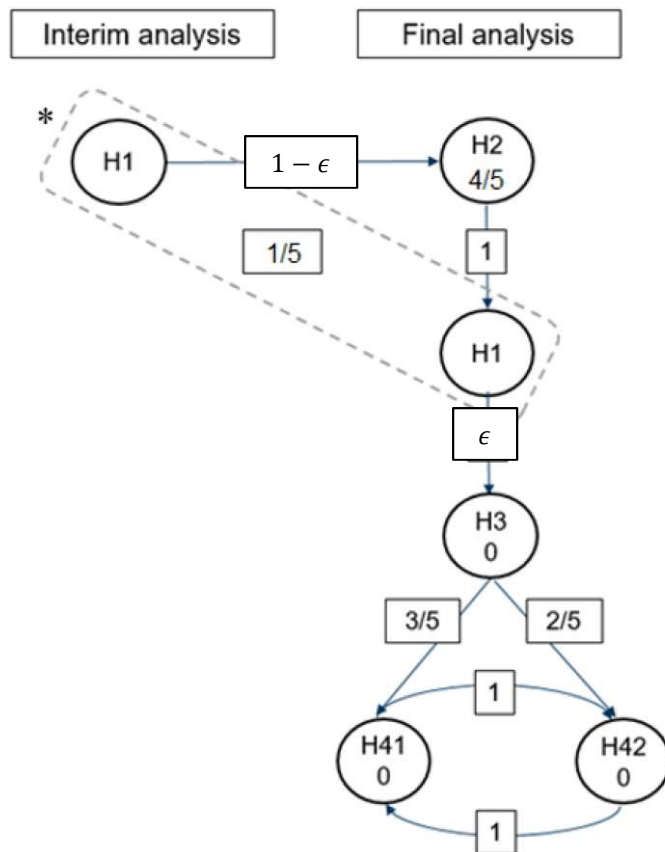
---

### Background:

- 2 jointly primary endpoints with hypotheses  $H_1$  and  $H_2$ ;  $H_1$  tested twice (IA and F),  $H_2$  only at F.
- Subsequent secondary endpoints, all just tested once at F
- Study continues irrespective of result of IA (due to long-term safety data collection and immature data for key secondary objectives)
- Conditional approval might be granted based on very convincing IA results.



## Example 2: graph from the protocol



$H_1$  initially tested at level  $\frac{\alpha}{5}$   
 $H_2$  initially tested at level  $\frac{4 \cdot \alpha}{5}$

If  $H_1$  rejected,  $H_2$  tested at  $\alpha$   
 If  $H_2$  rejected,  $H_1$  tested at  $\alpha$

**How do we best split  $\frac{\alpha}{5}$  and  $\alpha$  onto IA and F ?**

\*Note that  $H_1$  is **only** tested at the final analysis (at  $\alpha_F$  adjusted to account for  $H_1$  being tested and not rejected at the IA) in case of non-rejection at the IA and rejection of  $H_2$ .

## Example 2: graph from the protocol

---

How do we best split  $\frac{\alpha}{5}$  and  $\alpha$  onto IA and F ?

IA after 250 patients, F after 430 patients.

$\frac{\alpha}{5}$  will be completely used up at IA.

⇒ No re-testing of  $H_1$  if  $H_2$  cannot be rejected.

If  $H_2$  rejected at F,  $H_1$  will get  $4 \cdot \frac{\alpha}{5}$ . This will go entirely to the final analysis of  $H_1$ .

GS-levels for  $H_1$ ,  $\alpha = 2.5\%$ :  $(0.005, 0)$  for  $\frac{\alpha}{5}$  (if  $H_2$  not rejected)  
 $(0.005, 0.023)$  for  $\alpha$  (if  $H_2$  rejected)

## Example 2: R code

---

```
library(mvtnorm)
alpha<-0.025 # Set overall 1-sided alpha for hypothesis testing
nIA<-250 # Planned sample size at the Interim Analysis (IA: stage 1)
nFin<-430 # Planned sample size at the Final analysis (FA: stage 2)
IF<-nIA/nFin # Information Fraction at the IA
corr<-((1-sqrt(IF))*diag(2)+sqrt(IF)) #correlation matrix

# To adjust for multiplicity in the group-sequential test of H1 alone, alpha for the test
# of H1 is split to (alpha/5, alphaF) for the IA and Final Analysis respectively.
# calculate critical value for alpha spent at stage 1
c1_a<-qnorm(1-alpha/5) # alpha/5=1-sided alpha allocated at the IA

# Spending function to calculate the adjusted critical value x for stage 2,
# given alpha=overall alpha and c0 is the critical value of stage 1.
adjCrit<-function(alpha1, alpha, c0, corr){
  x<-qnorm(1-alpha1)
  check<-pmvnorm(upper=c(c0,x),corr=corr, algorithm=Miwa)
  return(1-check-alpha)
}
c2starp_a<-uniroot(adjCrit, lower = alpha/5, upper = alpha, alpha=alpha, c0=c1_a, corr=corr, tol=1E-12)

# adjusted alpha (on the % scale) for group sequential test of H3: UPCR at Final Analysis
# with alpha/5=0.5% spent at IA.
adjustedAlphaF<-round(c2starp_a$root*100,1)
adjustedAlphaF
```

## Example 2: Why split like this?

---

GS-levels for  $H_1$ ,  $\alpha = 2.5\%$ :

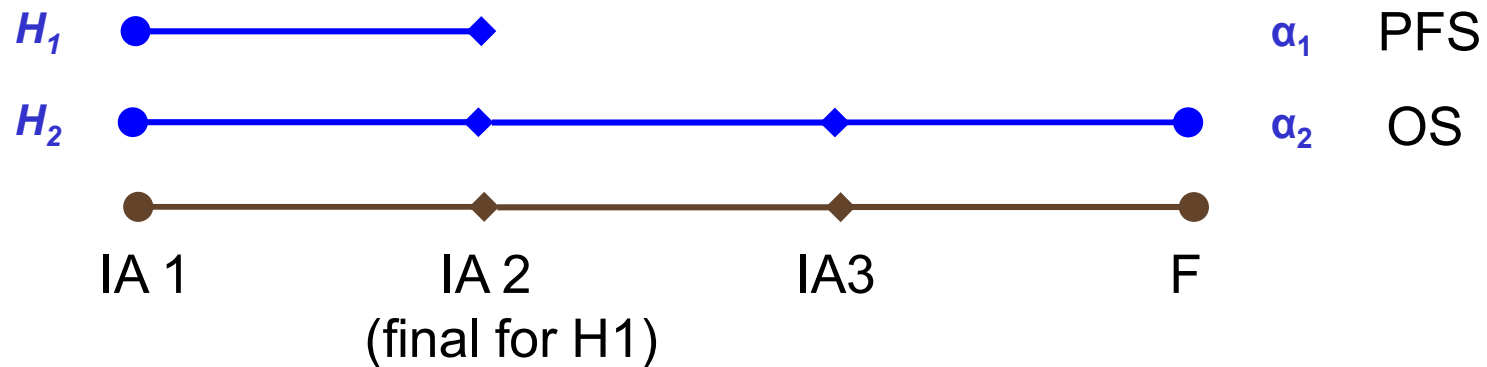
$(0.005, 0)$  for  $\frac{\alpha}{5}$  (if  $H_2$  not rejected)

$(0.005, 0.023)$  for  $\alpha$  (if  $H_2$  rejected)

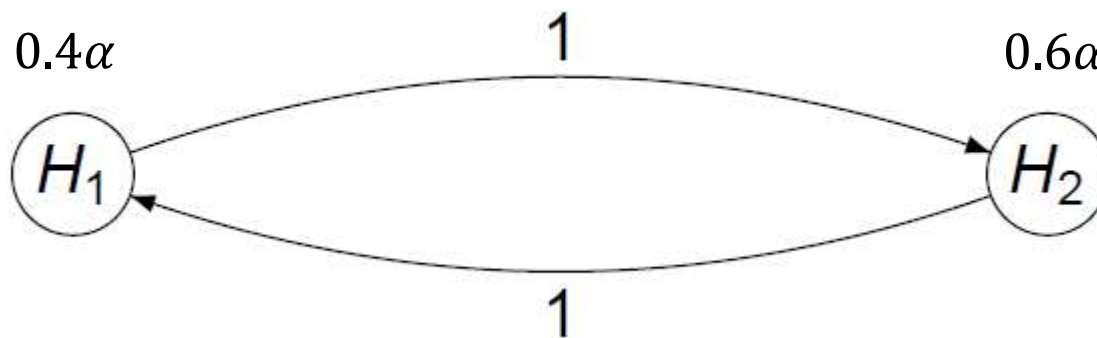
- This way, we avoid "having to look back": GS-spending at IA uses exactly the same value for all  $\alpha_{1j}$  of  $H_1$  at  $j = I, F$ .

## Example 3: matching interim alphas

- Jointly primary endpoints: PFS and OS



Graph:



## Example 3: matching interim alphas

---

- PFS: O'Brien-Fleming
- OS:
  - spending at 1.5%: O'Brien-Fleming
  - spending at 2.5%: For 1st interim, same as 1.5% OBF, for 2nd and 3rd interim same as 2.5% OBF, for the final all that's left.

Numerical example:

Information fractions of OS: 0.35, 0.5, 0.77, 1

Critical values for OS (Z-scale):

	IA1	IA2	IA3	F
OBF at 1.5%	3.949	3.254	2.550	2.218
at 2.5%	3.949	2.973	2.321	2.019
OBF at 2.5%	3.613	2.973	2.321	2.020

## Example 3: matching interim alphas

---

- Same motivation as in example 2: Matching the critical values avoids having to "look back" at previous interim analyses for alpha-adjustment.
- We could distribute the saving from IA1 in other ways.
- In this example, hardly any difference between conventional OBF at 2.5% and the modification.
  - Reason: OBF spends "next to nothing" at an interim analysis with 0.35 information fraction:  $\text{pnorm}(-3.613)=0.00015$
  - Hence, there is next to nothing to redistribute.

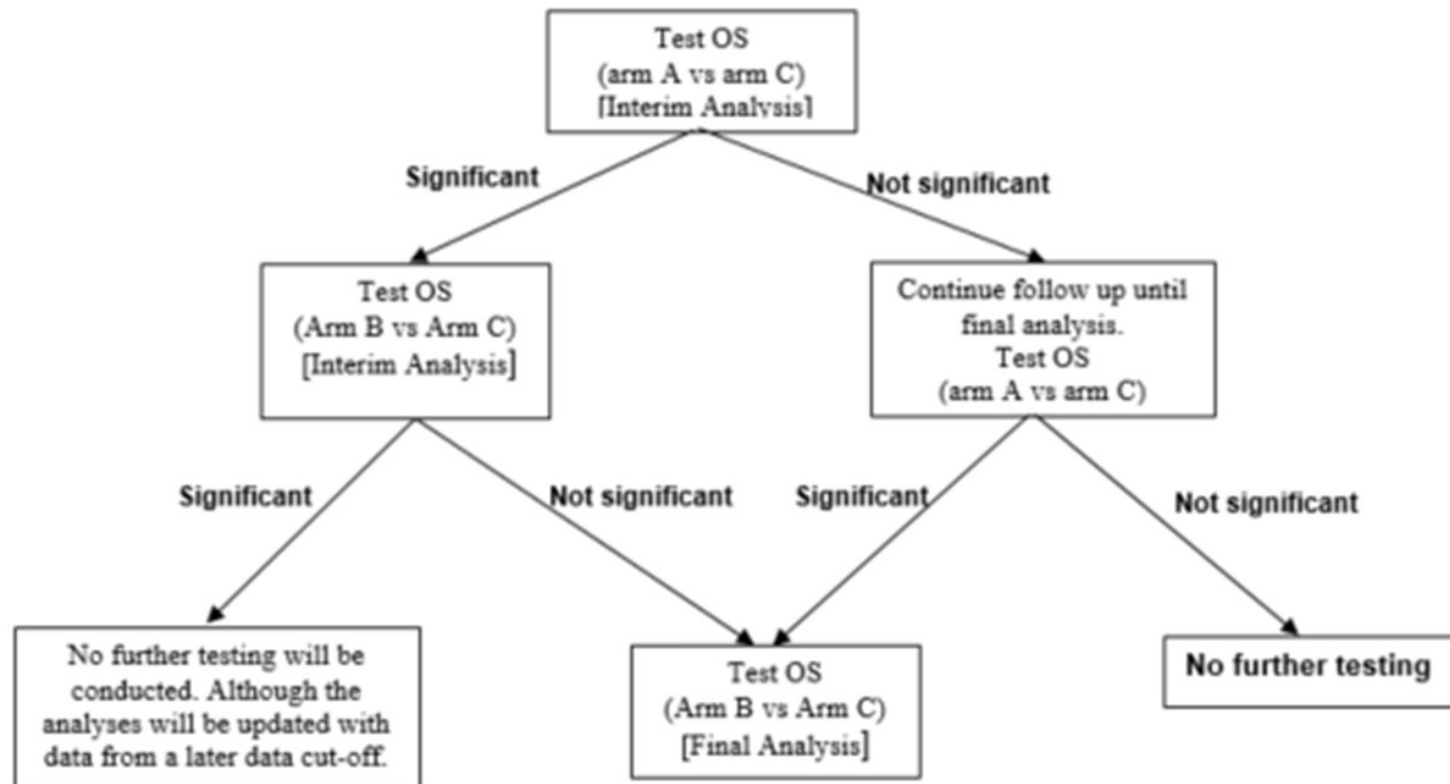
- R code:



modOBF\_last.R

## Example 4: traps to avoid

- 3-arm study with an interim and a final analysis
- Hierarchical testing (A vs C, then B vs C)

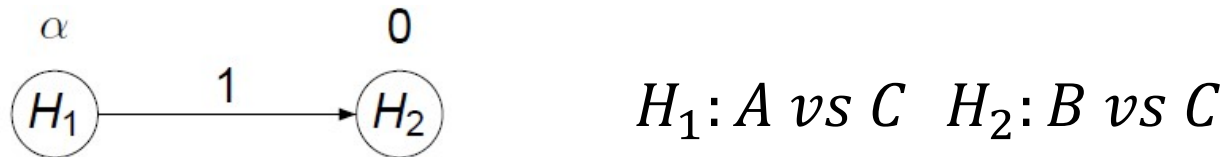




## Example 4: traps to avoid

---

- Looks very straightforward.

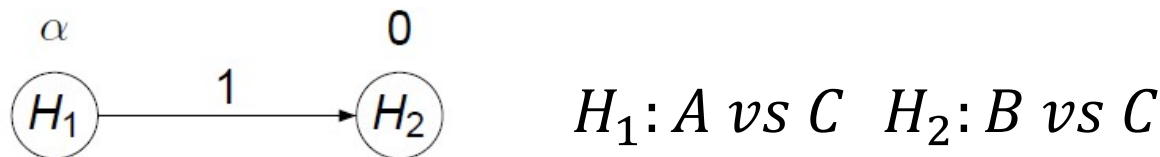


OBF for interim and final, both  $H_1$  and  $H_2$

- When protocol is almost finished, clinical team decides to bring in a futility stop for  $H_1$ .
- Idea: If  $H_1$  is stopped for futility (and hence "not tested"), we can test  $H_2$  at level  $\alpha$ .
- That's obviously (?) not true.
  - The trial statistician caught this, but was uncertain, so reached out to confirm.

## Example 4: traps to avoid

---



- Assume we spend  $(0, \alpha)$  at I and F for both  $H_1$  and  $H_2$ .
- Then the futility stop is just like deciding at the interim whether to test  $H_1$  and  $H_2$ .
- That's an adaptive design with endpoint selection.
- It is easy to calculate the inflation in this case analytically.
  - Inflation decreases with increasing correlation between between test statistics.
  - R code for calculation of the inflation:



alphainflationprimswitch.R

# Agenda

---

14:00 – 14:45

**Introduction to multiple testing**  
*Dong Xi*

14:45 – 16:15

**Graphical approaches to multiple testing**  
*Frank Bretz*

Break

16:30 – 17:30

**Extensions to group sequential designs**  
*Ekkehard Glimm*

**17:30 – 18:00**

**Extensions to pooled analyses from two studies**  
*Dong Xi*

## Background: FWER and two-study paradigm

---

- Regulatory guidance mandates strong FWER control at a pre-specified significance level  $\alpha$  for **a single study**
  - FDA (2017), EMA (2017)
- “Requirement” for two positive confirmatory studies
  - FDA (1998) guidance
  - many examples of diseases under the two-study paradigm
  - *“replication”, “independent substantiation”*
- Single study approvals generally limited to “mortality or irreversible morbidity” settings
  - *“statistically very persuasive”, “very low p-value”*

## Background: Pooled analysis to address resource imbalance

---

- Different sample sizes are needed to achieve a certain power (e.g., 80%) for different endpoints
  - A short-term symptom endpoint ( $E_1$ , e.g., FEV1)
  - A long-term outcome endpoint with low frequency/prevalence ( $E_2$ , e.g., COPD exacerbation)
  - $E_2$  may require a sample size twice as large as  $E_1$
- These unbalanced requirements of resources in a single study are amplified under the two-study paradigm
- One solution is to *pool data from the two studies* for  $E_2$  without doubling the sample size of each study

## Problem statement

---

- Pooling data from two studies increases statistical efficiency
- Different ways to pool
  - Naive pooling
  - Meta-analytic approach using 'study' as a stratification factor
- Poolability needs careful examination to avoid systematic bias/difference

*What approaches could be considered for managing multiplicity when data on an endpoint from two or more trials were planned to be pooled, and each trial had multiple endpoints managed?*

Lavange (2019)

# Principles

Bretz and Xi (2019)

---

- **Strong FWER control** at (one-sided) level  $\alpha = 0.025$  within each of the two confirmatory studies
- **Confirmation of independent substantiation** from at least one other endpoint prior to the pooled analysis
- Control of the **submissionwise error rate (SWER)** across both studies at an appropriate level
  - Probability to make a false claim of success for an endpoint while taking into account that a significant result on the same endpoint has to be obtained in both studies

# Three roles of the pooled analysis

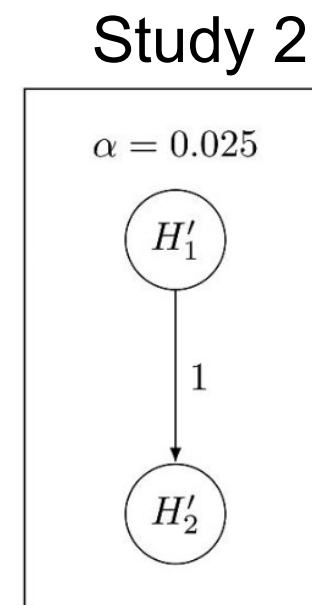
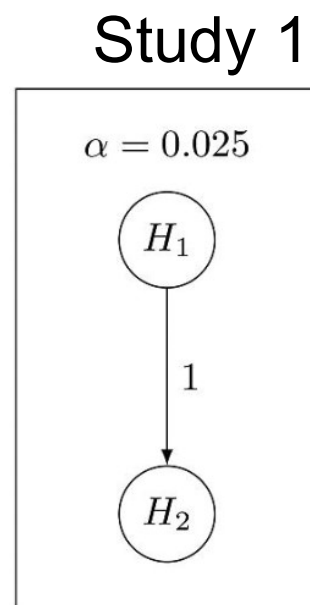
---

- Two endpoints
  - $E_2$  requires twice the sample size of  $E_1$  to satisfy a reasonable power
- Two-study paradigm
  - Independent and identically designed
  - $H_1$  and  $H'_1$  for  $E_1$  are tested independently in Study 1 and 2, respectively
  - $H_2$  and  $H'_2$  for  $E_2$  are tested independently in Study 1 and 2, respectively
- Pooled analysis for  $E_2$ 
  - $\tilde{H}_2$  is tested using data from both studies
- Role of the pooled analysis
  - Secondary
  - Primary
  - Co-primary



## Pooled analysis as a secondary analysis

- Two endpoints ( $E_1$ : primary and  $E_2$ : secondary)
  - Without the pooled analysis
    - Hierarchical test within each study
    - Study 1: test  $H_1$  at level  $\alpha = 0.025$ 
      - If rejected, test  $H_2$  at level  $\alpha = 0.025$
    - Study 2: test  $H'_1$  at level  $\alpha = 0.025$ 
      - If rejected, test  $H'_2$  at level  $\alpha = 0.025$
  - Summary
    - FWER for  $E_1$  and  $E_2 \leq 0.025$
    - SWER for  $E_1$  and  $E_2 \leq 0.025^2$



# Pooled analysis as a secondary analysis

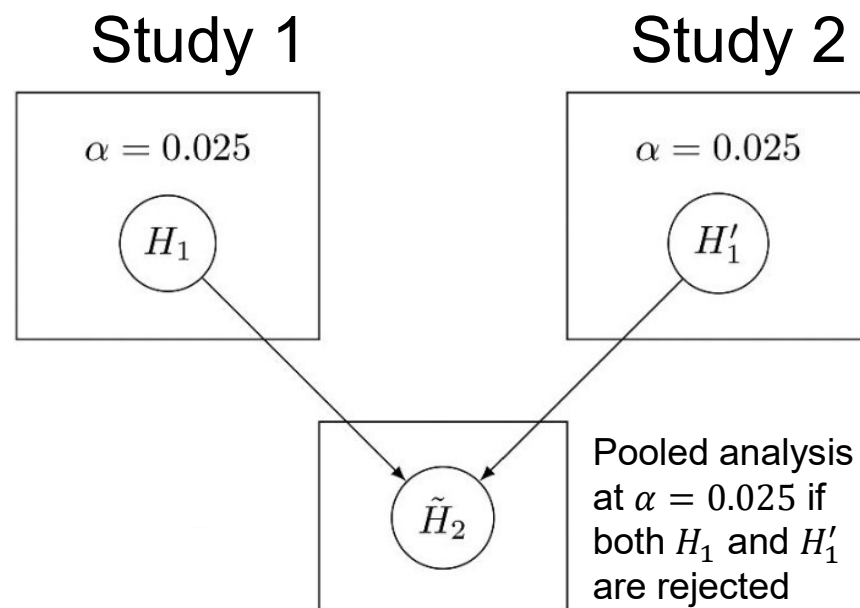
- Two endpoints ( $E_1$ : primary and  $E_2$ : secondary)

- With the pooled analysis

- Study 1: test  $H_1$  at level  $\alpha = 0.025$
- Study 2: test  $H'_1$  at level  $\alpha = 0.025$
- If *both  $H_1$  and  $H'_1$  are rejected*,  $\tilde{H}_2$  is tested using data from both studies at level  $\alpha = 0.025$

- Summary

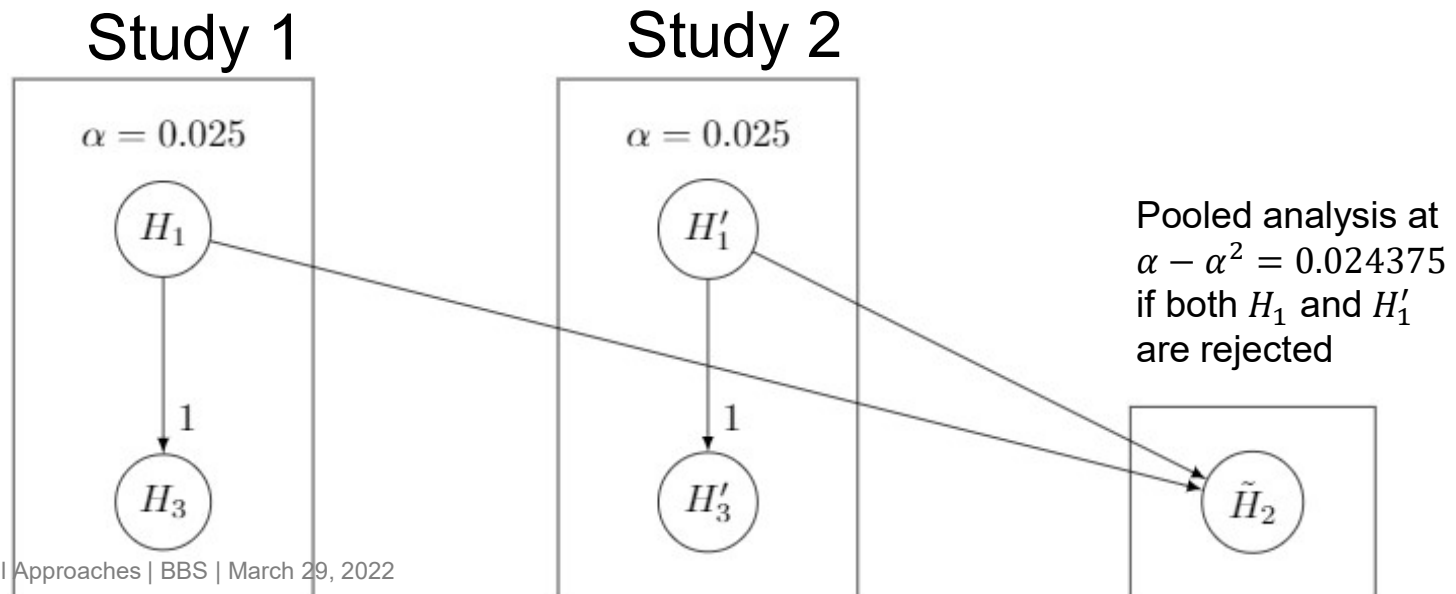
- FWER for  $E_1 \leq 0.025$
- SWER for  $E_1 \leq 0.025^2$
- Independent substantiation via  $E_1$
- Level  $\alpha = 0.025$  for  $\tilde{H}_2$  is determined by the conventional level of proof for a single hypothesis



# Pooled analysis as an additional secondary analysis

Bretz, Maurer, and Xi (2019)

- Three endpoints ( $E_1$ : primary and  $E_2, E_3$ : secondary)
- With the pooled analysis
  - Hierarchical test within each study for  $E_1$  and  $E_3$
  - If **both  $H_1$  and  $H'_1$  are rejected**,  $\tilde{H}_2$  is tested using data from both studies at level  $\alpha - \alpha^2$ 
    - Bonferroni split between  $H_3, H'_3$  and  $\tilde{H}_2$
- Summary
  - FWER within each study, i.e., for  $E_1$  and  $E_3$ , is controlled at level  $\alpha = 0.025$
  - SWER for  $E_1 \leq 0.025^2$
  - Type I error rate for secondary endpoints, i.e., for  $E_2$  and  $E_3$ , is controlled at level  $\alpha = 0.025$

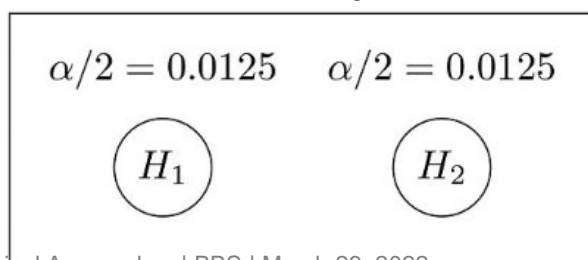


## Pooled analysis as a primary analysis

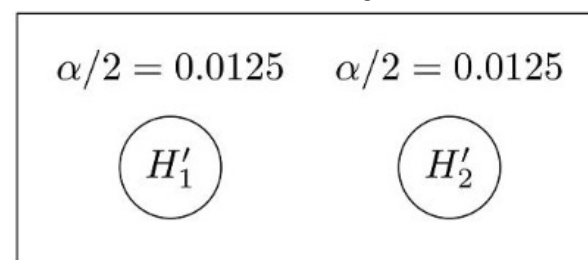
---

- Two endpoints ( $E_1, E_2$ : primary)
- Without the pooled analysis
  - For example, Bonferroni test within each study
  - Study 1: test  $H_1$  and  $H_2$  at level  $\alpha/2 = 0.0125$
  - Study 2: test  $H'_1$  and  $H'_2$  at level  $\alpha/2 = 0.0125$
- Summary
  - FWER for  $E_1$  and  $E_2 \leq 0.025$
  - SWER for  $E_1$  and  $E_2 \leq 2 \cdot 0.0125^2 < 0.025^2$

### Study 1

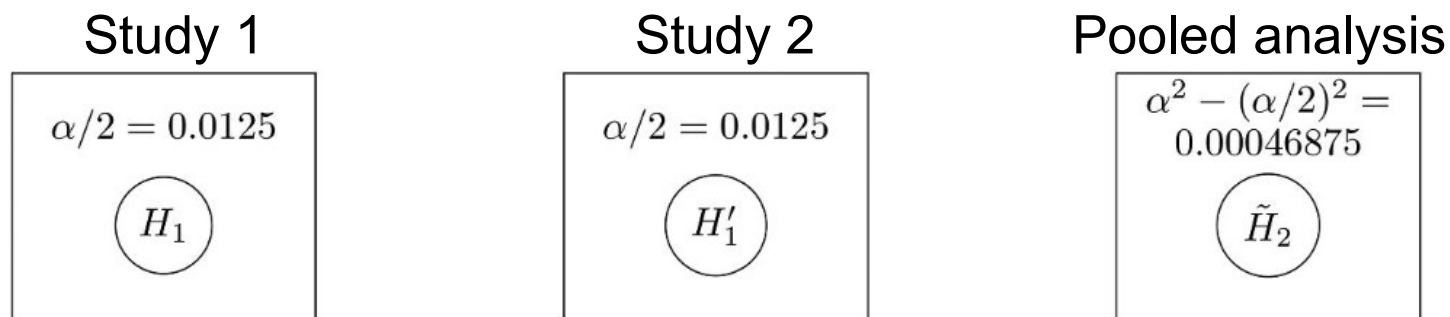


### Study 2



## Pooled analysis as a primary analysis

- Two endpoints ( $E_1, E_2$  primary)
- With the pooled analysis
  - Study 1: test  $H_1$  at level  $\alpha/2 = 0.0125$
  - Study 2: test  $H'_1$  at level  $\alpha/2 = 0.0125$
  - Test  $\tilde{H}_2$  at level  $\alpha^2 - (\alpha/2)^2 = 0.00046875$  (Bonferroni split between  $H_1, H'_1$  and  $\tilde{H}_2$ )
- Summary
  - FWER for  $E_1 \leq 0.0125$
  - SWER for  $E_1$  and  $E_2 \leq 0.025^2$
  - *If only  $\tilde{H}_2$  is significant, independent substantiation may be questioned since either  $H_1$  or  $H'_1$  is not significant*



# Pooled analysis as a co-primary analysis

---

- Two endpoints ( $E_1, E_2$ : co-primary)
- Without the pooled analysis
  - Study 1: test  $H_1$  and  $H_2$  each at level  $\alpha = 0.025$
  - Study 2: test  $H'_1$  and  $H'_2$  each at level  $\alpha = 0.025$
  - Claim study success only if both hypotheses are rejected
- Summary
  - FWER for  $E_1$  and  $E_2 \leq 0.025$
  - SWER for  $E_1$  and  $E_2 \leq 0.025^2$

# Pooled analysis as a co-primary analysis

---

- Two endpoints ( $E_1, E_2$ : co-primary)
- With the pooled analysis
  - Study 1: test  $H_1$  at level  $\alpha = 0.025$
  - Study 2: test  $H_1'$  at level  $\alpha = 0.025$
  - Test  $\tilde{H}_2$  at level  $\alpha = 0.025$ 
    - Determined by the conventional level of proof for a single hypothesis
- Summary
  - FWER for  $E_1 \leq 0.025$
  - SWER for  $E_1 \leq 0.025^2$  and for  $E_2 \leq 0.025$
  - Independent substantiation via  $E_1$
  - Significance level for  $\tilde{H}_2$  could be determined to be  $[\alpha^2, \alpha]$ , in order to balance the level of replication standard and the feasibility of the trials

# ASCLEPIOS I and II – Design

Hauser et al. (2020)

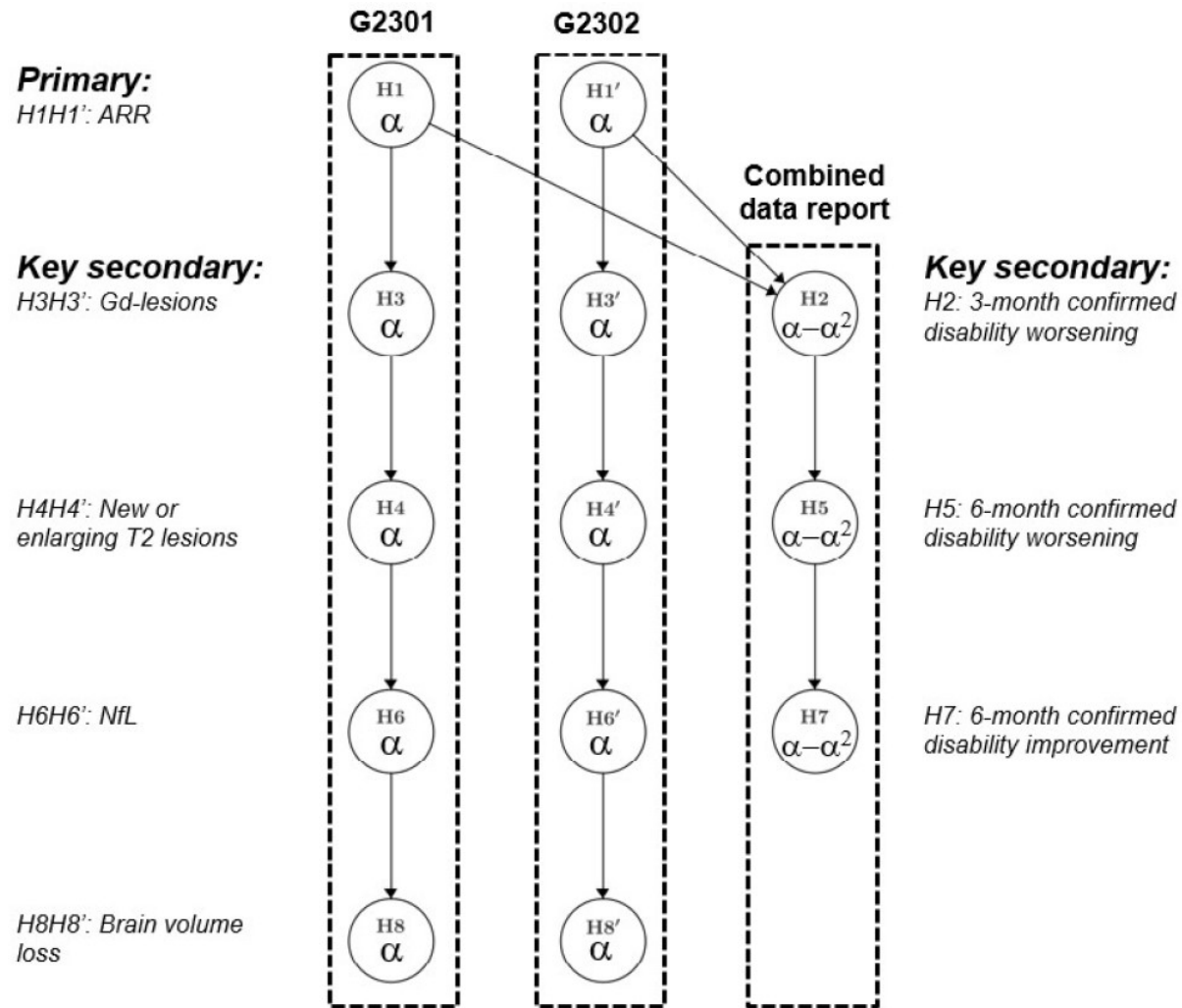
---

- Two confirmatory studies of identical design in patients with multiple sclerosis to compare ofatumumab versus teriflunomide
- Primary endpoint was the annualized relapse rate (ARR)
- Key secondary endpoints:
  - disability worsening after 3 months, disability worsening after 6 months, disability improvement after 6 months
  - number of Gd lesions, number of new or enlarging T2 lesions, neurofilament light (NfL) chain, brain volume loss
- Randomizing 900 patients per study would provide  $> 90\%$  power in each study to detect a 40% lower ARR
- Combining the data from both studies, a total of 1800 patients would provide 90% power and 80% power to detect a 38.6% lower risk of disability worsening at 3 months and at 6 months, respectively



# ASCLEPIOS I and II – Testing scheme

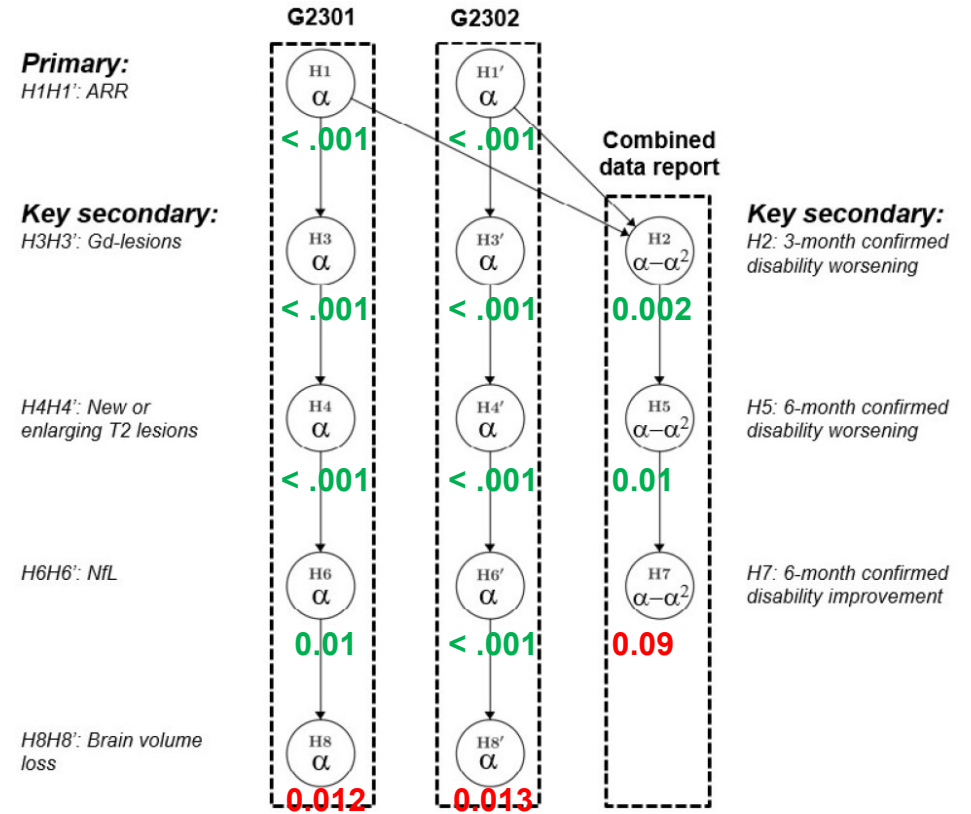
Hauser et al. (2020)



# ASCLEPIOS I and II – Results

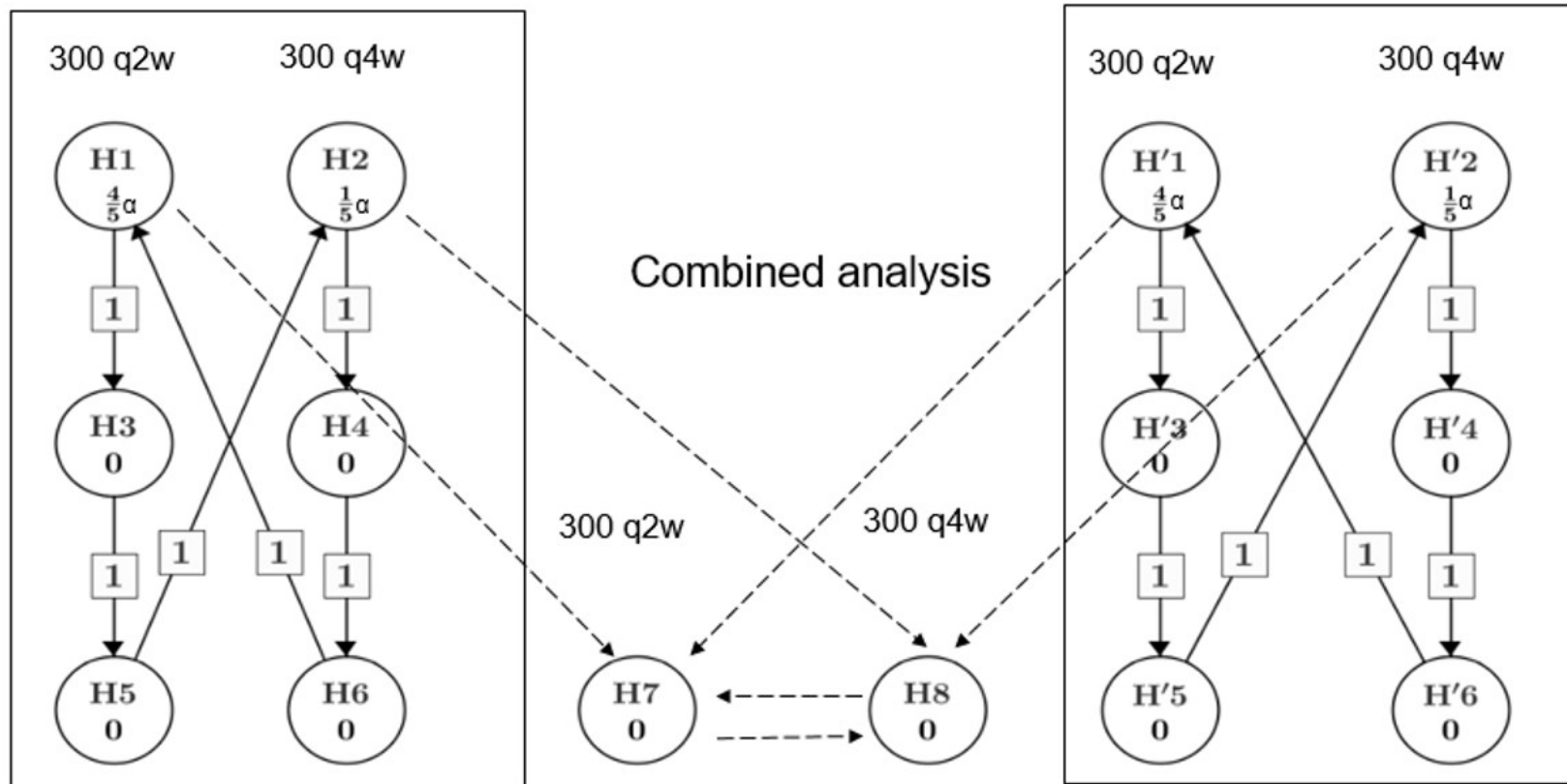
Hauser et al. (2020)

- Overall, 946 patients were assigned to receive ofatumumab and 936 to receive teriflunomide
- All confidence intervals and p-values in the study report were presented without adjustments



# More examples

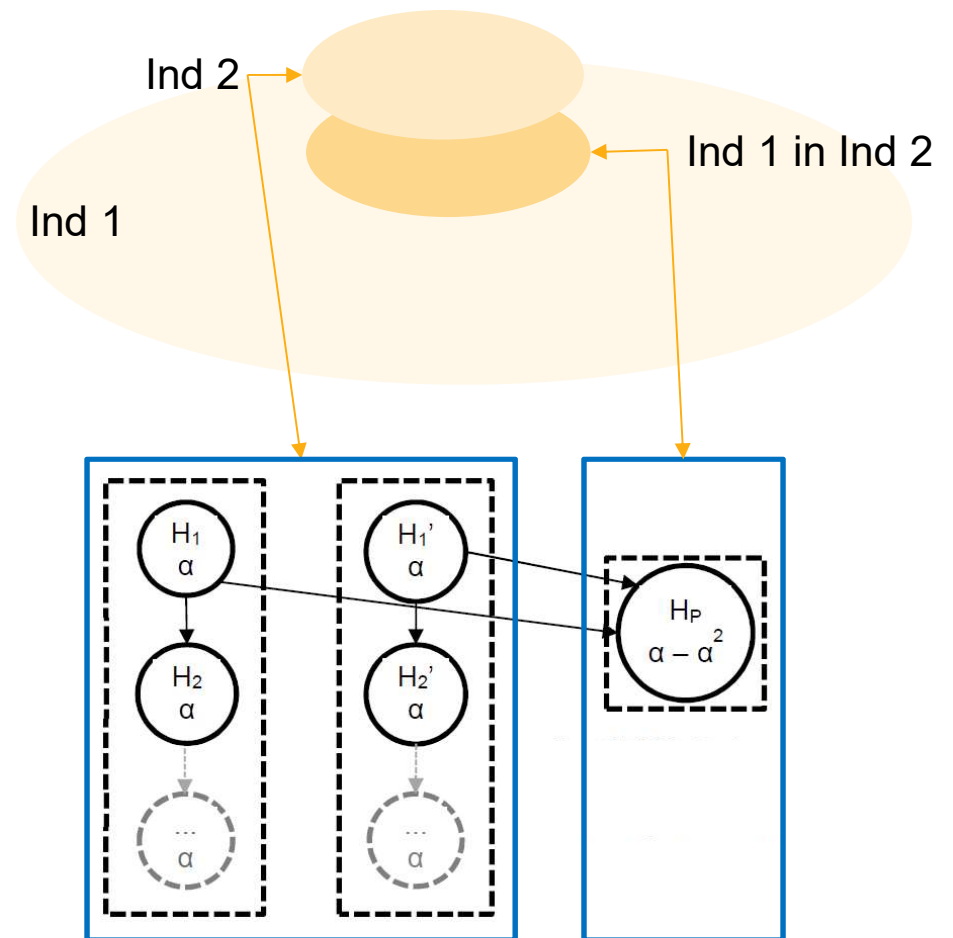
## Multiple doses and multiple endpoints



# More examples

## Two birds, one stone

- File a single dossier for two related indications ('Ind 2' and 'Ind 1 in Ind 2') based on two sets of endpoints from the two confirmatory studies
  - The two confirmatory studies used 'Ind 2' in their testing strategy
  - The project level testing strategy incorporated the endpoint relevant for 'Ind 1 in Ind 2'



## Conclusions and other considerations

---

- Several test strategies are proposed, based on a few key principles and depending on the role on the pooled analysis
- Pooled analysis would be done in a timely manner if both studies are finished simultaneously
- Reduce the dependency of individual trial reports on the pooled analysis for logistic efficiency
  - Not recommend to include the pooled analysis into the study testing strategy, see Bretz, Maurer, and Xi (2019) for a case study
- Pooled analysis relies on independent substantiation
  - Efficiency of the pooled analysis may be outweighed by the risk of inconsistency (e.g., two studies of different designs/populations)
  - Maca, Gallo, Branson, and Maurer (2002) discuss a consistency requirement for testing the pooled analysis

Q & A

---

**Any questions?**

# References

---

- Anderson, Guo, Zhao, Sun (2021): A unified framework for weighted parametric group sequential design (WPGSD). [arXiv:2103.10537](https://arxiv.org/abs/2103.10537)
- Bretz, Maurer, Brannath, Posch (2009) A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 28: 586–604.
- Bretz, Posch, Glimm, Klinglmueller, Maurer, Rohmeyer (2011) Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. *Biometrical Journal* 53: 894–913.
- Bretz, Xi (2019) Commentary on ‘Statistics at FDA’. *Statistics in Biopharmaceutical Research*, 11, 20-25
- Bretz, Maurer, Xi (2019) Replicability, Reproducibility, and Multiplicity in Drug Development. *Chance*, 32(4), 4-11
- Burman, Sonesson, Guilbaud (2009) A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 28: 739–761.
- Dmitrienko, Offen, Westfall (2003) Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 22: 2387–2400.

# References

---

- EMA (2017) Guideline on multiplicity issues in clinical trials
- FDA (1998) Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products
- FDA (2017) Multiple Endpoints in Clinical Trials
- Hauser et al. (2020) Ofatumumab versus teriflunomide in multiple sclerosis. *New England Journal of Medicine*, 383(6), 546-557
- Jennison, Turnbull (1997): Group-sequential analysis incorporating covariate information. *J Am Stat Assoc* 92, 1330–1341.
- LaVange (2019) Statistics at FDA: Reflections on the Past Six Years. *Statistics in Biopharmaceutical Research*, 11, 1-12
- Maca, Gallo, Branson, Maurer (2002) Reconsidering Some Aspects of the Two-Trials Paradigm. *Journal of Biopharmaceutical Statistics*, 12, 107-119
- Maurer, Bretz (2013): Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research* 5, 311–320.



# References

---

- Maurer, Glimm, Bretz (2011): Multiple and Repeated Testing of Primary, Coprimary, and Secondary Hypotheses. *Statistics in Biopharmaceutical Research* 3, 336-352
- Scharfstein, Tsiatis, Robins (1997): Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *J Am Stat Assoc* 92, 1342–1350.
- Tamhane, Xi, Gou (2021): Group sequential Holm and Hochberg procedures. *Statistics in Medicine* 40, 5333-5350.
- Xi, Glimm, Bretz (2016): Multiplicity. In: *Cancer clinical trials – current and controversial issues in design and analysis*, edited by Stephen L. George, Xiaofei Wang, and Herbert Pang, Boca Raton, FL: Chapman & Hall/CRC.
- Xi, Glimm, Maurer, Bretz (2017): A unified framework for weighted parametric multiple test procedures. *Biometrical Journal* 59, 918-931.
- Xi, Bretz (2021) Graphical Approaches for Multiple Comparison Procedures. In: *Handbook of Multiple Comparisons*. Chapman and Hall/CRC.