The Labyrinth of Multiple Testing: How to avoid the pitfall of false positives

FWER control

12th SISMEC National Congress 2023

Livio Finos and Angela Andreella

Outline

1. **FamilyWise Error Rate (FWER)**
2. Bonferroni (single-step)
3. Holm (step-wise)
4. Closed Testing
5. Gatekeeping strategies
6. Summary

FamilyWise Error Rate (FWER)

		Null hypothesis		
		False	True	Tot
Test	Rejected	S	V	R
	Not rejected	T	U	$m - R$
	Tot	m_1	m_0	m

$$\text{FWER} = \mathbb{P}(\text{at least one type I error}) = \mathbb{P}(V > 0)$$

A procedure **controls** it if FWER $\leq \alpha$.

Šidàk correction

If I want to check the **FWER** at the α level, at which individual $\widetilde{\alpha}$ level should I reject the individual tests?

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}\big(p_i \leq \widetilde{\alpha} \text{ for at least one } i \text{ true null hypothesis}\big) \\
&= \mathbb{P}\Big(\bigcup_{i \in \{\text{true null hypotheses}\}} \{p_i \leq \widetilde{\alpha}\} \Big) \\
&= 1 - \mathbb{P}\Big(\bigcap_{i \in \{\text{true null hypotheses}\}} \{p_i > \widetilde{\alpha}\} \Big) = \\
&\quad (\textit{deMorgan}) \\
&= 1 - (1 - \widetilde{\alpha})^{m_0} = (m_0 : \#\{\text{true null hypothesis}\}) \\
&\quad (\text{we don't know } m_0, \text{ we know that though } m_0 \leq m) \\
&\leq 1 - (1 - \widetilde{\alpha})^{m}
\end{aligned}
$$

Šidàk correction

From this, we get:

$$\begin{aligned} \alpha &= 1-(1-\widetilde{\alpha})^m \\ 1-\alpha &= (1-\widetilde{\alpha})^m \\ (1-\alpha)^{1/m} &= (1-\widetilde{\alpha}) \\ \widetilde{\alpha} &= 1-(1-\alpha)^{1/m} \end{aligned}$$

Then, it is enough to reject every single hypothesis at the level

$$\boxed{\widetilde{\alpha} = 1-(1-\alpha)^{1/m}}$$

(i.e., I reject the p-values for which $p \leq \widetilde{\alpha}$)

Šidàk correction

From this, we get:

$$\begin{aligned} \alpha &= 1-(1-\widetilde{\alpha})^m \\ 1-\alpha &= (1-\widetilde{\alpha})^m \\ (1-\alpha)^{1/m} &= (1-\widetilde{\alpha}) \\ \widetilde{\alpha} &= 1-(1-\alpha)^{1/m} \end{aligned}$$

Then, it is enough to reject every single hypothesis at the level

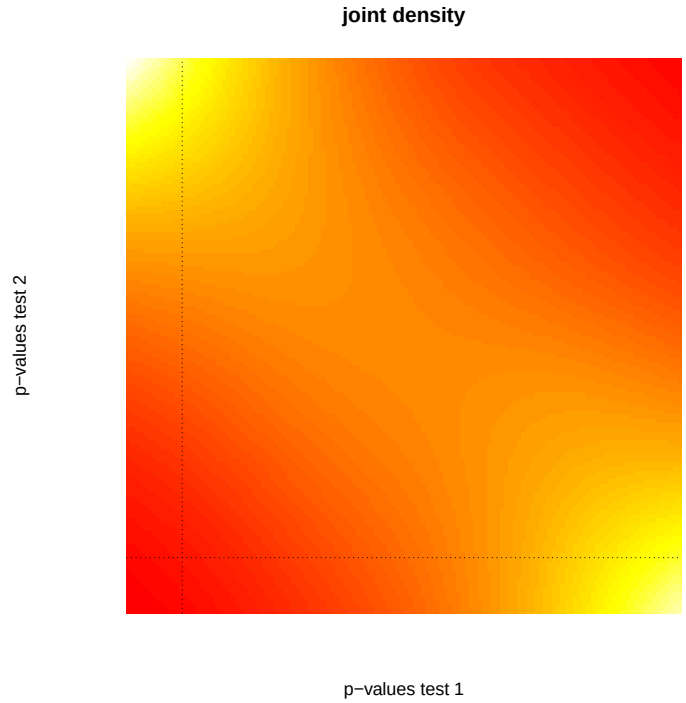
$$\boxed{\widetilde{\alpha} = 1-(1-\alpha)^{1/m}}$$

(i.e., I reject the p-values for which $p \leq \widetilde{\alpha}$)

PROBLEM: it is valid only when the p-values are **INDEPENDENT**.
In most cases, the tests have dependence-induced dependence between the original variables.

Dependent P-values

It may happen that $\mathbb{P}(\text{at least one false rejection of } H_0) > (!)1 - (1-\alpha)^2$



Remark: Remember, however, that the marginal distributions are uniform because the two tests are under H_0.

Outline

1. FamilyWise Error Rate (FWER)
2. **Bonferroni (single-step)**
3. Holm (step-wise)
4. Closed Testing
5. Gatekeeping strategies
6. Summary

Boole

Boolean inequality

Let two events A e B:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

so

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Generalizing $A_1, \ldots, A_m$:

$$\mathbb{P}(\bigcup_{i=1}^{m} A_i) \leq \sum_{i=1}^{m} \mathbb{P}(A_i)$$

Equality

Equality occurs when events are disjointed

FamilyWise Error Rate (FWER)

Bonferroni inequality

Reduce α

Reject H_i if $\boxed{p_i \leq \widetilde{\alpha} = \alpha/m}$ ($m =$ **number of hypotheses**)

FWER control

$$
\begin{aligned}
\text{FWER} &= \mathbb{P}\big(p_i \leq \alpha/m \text{ for } \textbf{at least one } i \text{ true null hypothesis}\big) \\
&= \mathbb{P}\Big(\bigcup_{i \in \{\text{true null hypothesis}\}} \{p_i \leq \alpha/m\} \Big) \\
&\leq \sum_{i \in \{\text{true null hypothesis}\}} \mathbb{P}(p_i \leq \alpha/m) \\
&\leq m_0 \frac{\alpha}{m} \leq m \frac{\alpha}{m} = \alpha
\end{aligned}
$$

Bonferroni Procedure

Multiplicity adjusted p-value

$\boxed{\tilde{p}_i = mp_i}$ for $i = 1, \ldots, m$ and reject if $\tilde{p}_i \leq \alpha$

Advantages

- Very **simple**
- Controls the FWER under **any** dependency

Disadvantages

- **Conservative** (adjusted p-values are often high, leading to few rejections)

Outline

1. FamilyWise Error Rate (FWER)
2. Bonferroni (single-step)
3. **Holm (step-wise)**
4. Closed Testing
5. Gatekeeping strategies
6. Summary

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

Adj. p-value:	$p_A 5$	$p_B 5$	$p_C 5$	$p_D 5$	$p_E 5$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:	A	B	C	D	E	
$\mathcal{R}$:						

Bonferroni

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

Suppose p_A and p_C are significant

Adj. p-value:	$p_A 5$	$p_B 5$	$p_C 5$	$p_D 5$	$p_E 5$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:	A	B	C	D	E	
$\mathcal{R}$:						

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

Adjusted p-value: $p \cdot 3$

Adj. p-value: - $p_B 3$ - $p_D 3$ $p_E 3$ $\leq ? \alpha$

$\mathcal{H} \setminus \mathcal{R}$: B D E

$\mathcal{R}$: A C

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

Suppose p_D is significant

Adj. p-value: - $p_B 3$ - $p_D 3$ $p_E 3$ $\leq ? \alpha$

$\mathcal{H} \setminus \mathcal{R}$: B D E

$\mathcal{R}$: A C

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

Adjusted p-value: $p \cdot 2$

Adj. p-value:	-	$p_B 2$	-	-	$p_E 2$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B			E	
$\mathcal{R}$:	A		C	D		

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Holm's Procedure[1]

1. First step: adjusted p-value: $p \cdot m$; reject if $\leq \alpha$
2. After r rejections, adjusted p-value: $p \cdot (m - r)$
3. Stop as soon as nothing is rejected

No rejections. Stop

Adj. p-value:	-	$p_B 2$	-	-	$p_E 2$	$\leq ? \alpha$
$\mathcal{H} \setminus \mathcal{R}$:		B			E	
$\mathcal{R}$:	A		C	D		

[1] Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*; 6(2):65–70.

Outline

1. FamilyWise Error Rate (FWER)
2. Bonferroni (single-step)
3. Holm (step-wise)
4. **Closed Testing**
5. Gatekeeping strategies
6. Summary

Closed Testing[2]

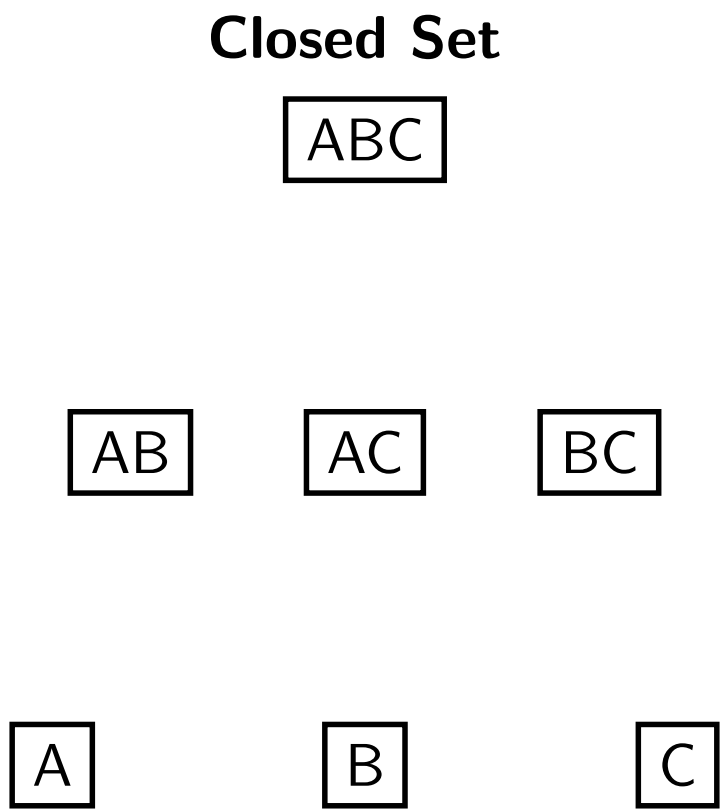
Closed Set of Hypotheses (all possible intersections)

Initial Hypotheses

[2] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Closed Testing[2]

Top Node Test (e.g., MANOVA)



[2]R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Closed Testing[4]

- Test each hypothesis at level α
- Reject an individual hypothesis A if you reject all hypotheses where A is component.
  Example: Reject H_A if $p_A \leq \alpha$, $p_{AB} \leq \alpha$, $p_{AC} \leq \alpha$ and $p_{ABC} \leq \alpha$

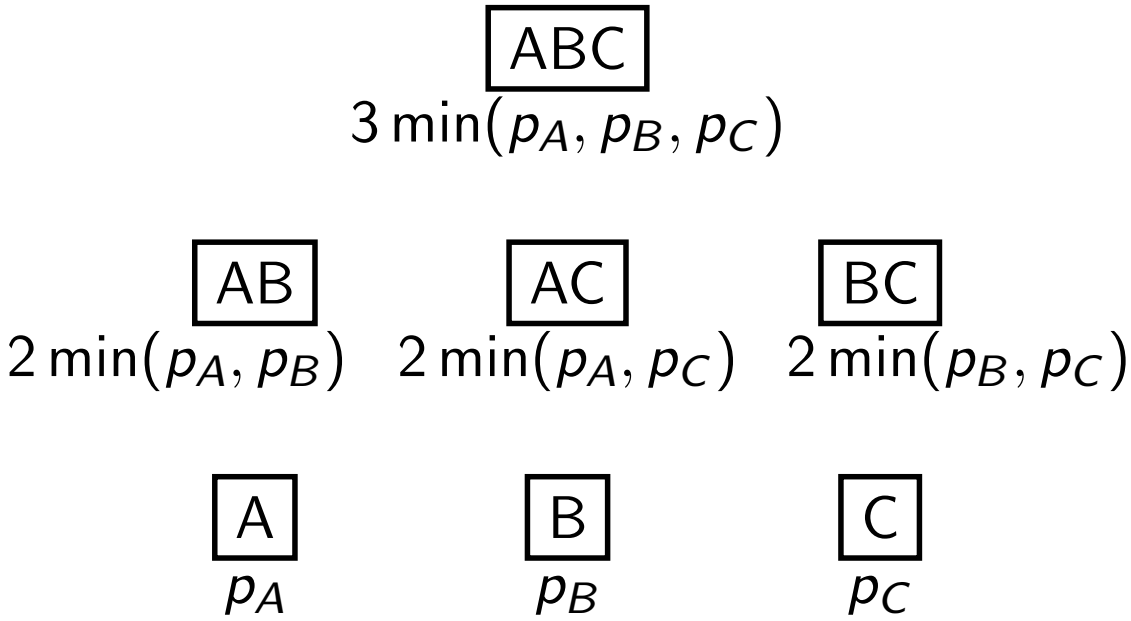
Relevance: **ANY** multiple testing procedure can be written as Closed Testing procedure. [3] This makes easy to prove the control of the Type I error for new procedures.

Disadvantage: The tested hypotheses often become too many: $= 2^{\text{hypotheses}} - 1$ (i.e., research focuses on shortcuts. see next...)

[3] Goeman, Hemerik, and Solari (2021) Only closed testing procedures are admissible for controlling false discovery proportions. Ann. Statist. 49(2): 1218-1238

[4] R Marcus, E Peritz, KR Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63: 655-660.

Holm as a Closed Testing procedure



Assume w.l.o.g. $p_A \leq p_B \leq p_C$:

- $3\min(p_A, p_B, p_C) \leq \alpha$ implies $2\min(p_A, p_B) \leq \alpha$, $2\min(p_A, p_C) \leq \alpha$ and $p_A \leq \alpha$, that is Reject H_A
- $2\min(p_B, p_C) \leq \alpha$ implies $p_B \leq \alpha$ (already $3\min(p_A, p_B, p_C) \leq \alpha$ and $2\min(p_A, p_B) \leq \alpha$), that is Reject H_B.

Outline

1. FamilyWise Error Rate (FWER)
2. Bonferroni (single-step)
3. Holm (step-wise)
4. Closed Testing
5. **Gatekeeping strategies**
6. Summary

Union-intersection hypothesis

Reject the global hypothesis of no effect if there is evidence of an effect with respect to **AT LEAST ONE** individual objective:

- the trial's outcome is declared positive if at least one analysis produces a significant result,
- Each analysis is independently clinically relevant
- Each endpoint, dose, or population analysis independently provides proof of efficacy

Let denote the hypotheses $H_1, \ldots, H_m$ corresponding to the multiple objectives tested against the alternative hypotheses $K_1, \ldots, K_m$:

$$\boxed{H_I : \cap_{i=1}^{m} H_i \quad \text{versus} \quad H_U : \cup_{i=1}^{m} K_i}$$

$\rightarrow$ **adjust for multiplicity!**

Gatekeeping strategies

Multiple objectives pursued in clinical trials typically exhibit a **hierarchical structure** $\rightarrow$ Primary and secondary objectives.

To construct a **gatekeeping procedure**, one first needs to define two or more families of analyses, for example:

- **Primary and secondary endpoints**: Primary endpoints determine the **trial's outcome**, and key secondary endpoints provide useful **supportive** information about efficacy and safety
- **Primary and secondary populations**: General population versus **subgroups** of patients who are more likely to benefit from treatment
- **Primary and secondary tests**: **Noninferiority** assessment as the primary analysis followed by a **superiority** assessment

Gatekeeping strategies

Family 1: $F_1 = \{H_1, \ldots, H_{k_1}\}$, null hypotheses

...

Family m: $F_m = \{H_{k_{m-1}+1}, \ldots, H_{k_m}\}$, null hypotheses

Each family (except for the last one) serves as a **gatekeeper** for the next one, in the sense that one must pass it to perform analyses in the next family

$\rightarrow$ **increase power** by accounting for **hierarchical** structure of multiple families.

Gatekeeping Procedures

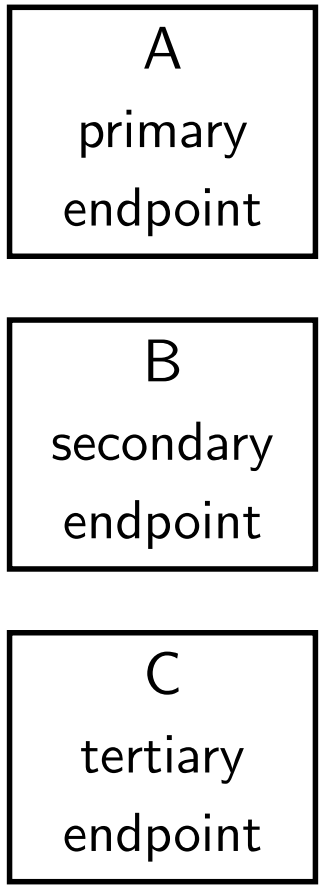
α **allocation**: Initial distribution of the error rate α across the hypotheses.

α **propagation**: *"Use it or lose it"* principle: After a hypothesis is rejected, the significance level used can be applied to remaining non-rejected hypotheses. $\rightarrow$ follows from the closure principle and does not result in FWER inflation.

- **Sequential Testing**: Families of hypotheses are tested sequentially starting with Family 1 $\rightarrow$ Error rate is transferred along the sequence
- **Sequential testing with re-testing**: Families of hypotheses are tested sequentially starting with Family 1 with a re-testing loop $\rightarrow$ Error rate is transferred along the sequence and then back to Family 1
- **Simultaneous testing**: Families of hypotheses are tested simultaneously $\rightarrow$ Error rate is transferred among families

Gatekeeping Procedures

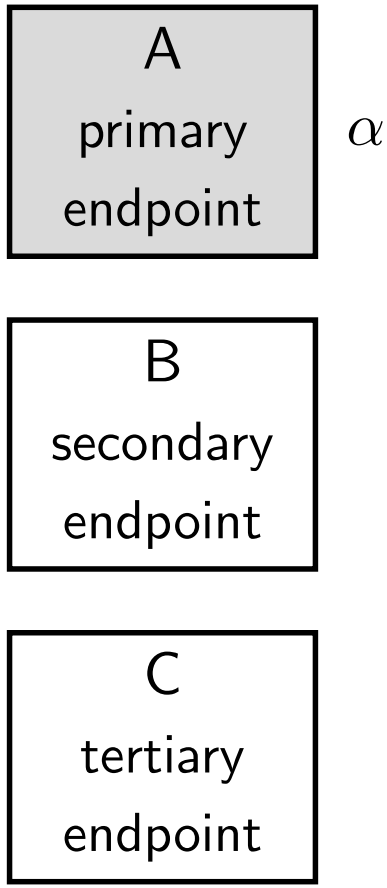
Sequential Testing
(three ordered endpoints: high, medium, low doses versus placebo)



Gatekeeping Procedures

Sequential Testing
(three ordered endpoints: high, medium, low doses versus placebo)

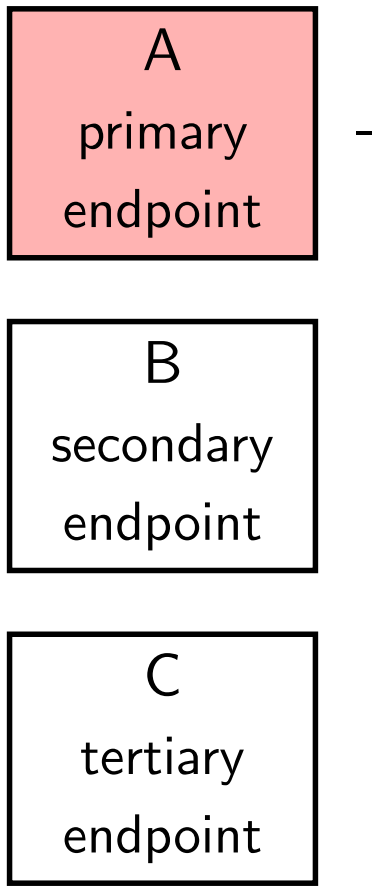
Start test A at α



Gatekeeping Procedures

Sequential Testing
(three ordered endpoints: high, medium, low doses versus placebo)

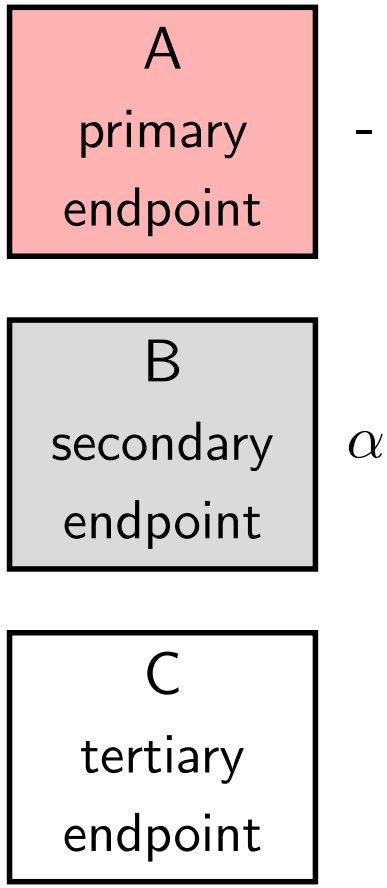
Suppose $p_A < \alpha$



Gatekeeping Procedures

Sequential Testing
(three ordered endpoints: high, medium, low doses versus placebo)

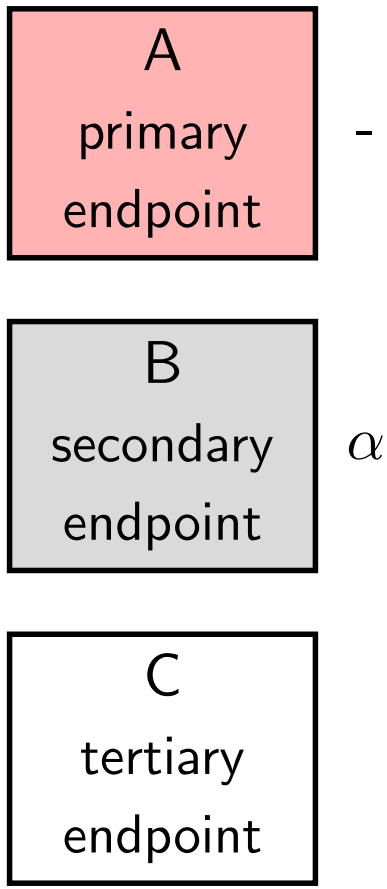
Go on to test B at α



Gatekeeping Procedures

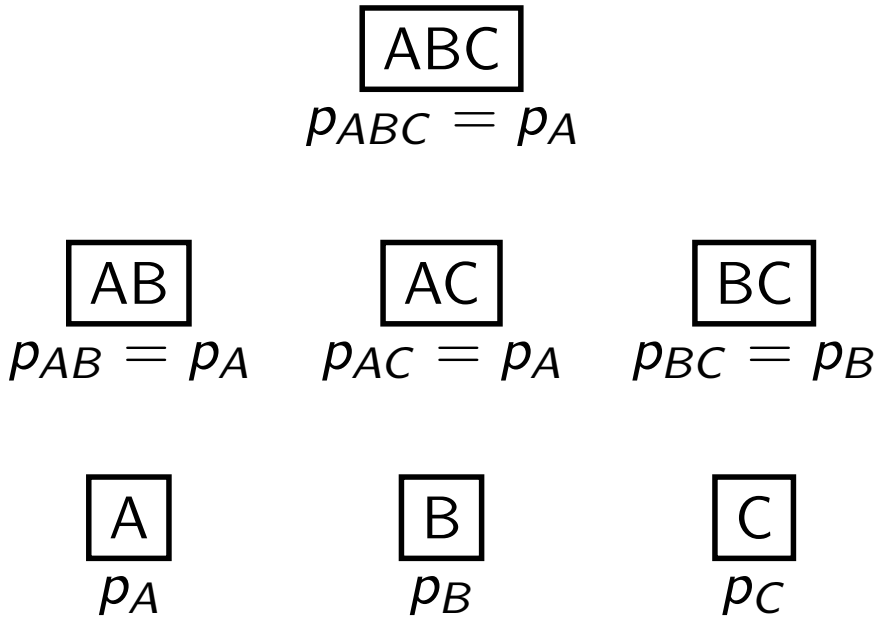
Sequential Testing
(three ordered endpoints: high, medium, low doses versus placebo)

Suppose $p_B > \alpha$. Stop



Sequential procedure as a Closed Testing procedure

Hypotheses are pre-ordered, and individual tests are also used for intersection hypotheses:



- $p_A \leq \alpha$ implies $p_{AB} = p_{AC} = p_{ABC} = p_A \leq \alpha$
- $p_B \leq \alpha$ implies $p_{BC} = p_B \leq \alpha$ (and $p_{AB} = p_{ABC} \leq \alpha$ already).

Parallel Gatekeeping [5]

Family 1 is a **parallel gatekeeper** for Family 2, i.e., **at least one** hypothesis must be rejected in Family 1 to proceed to Family 2.

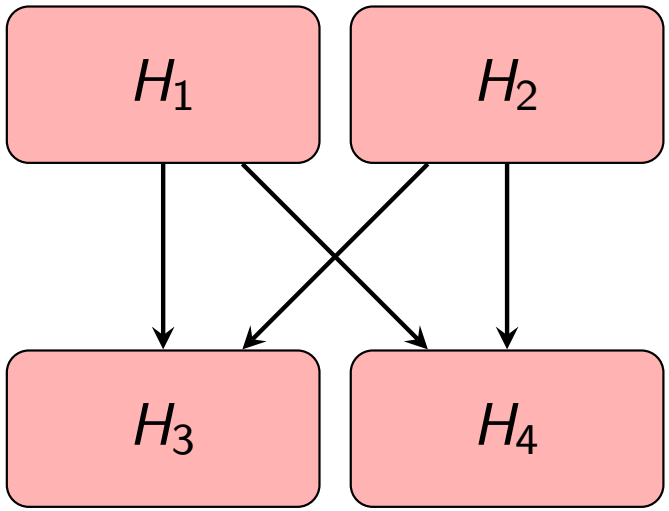
Example: Schizophrenia trial
Objective: Evaluate the efficacy of a treatment in patients diagnosed with schizophrenia
Design: Two doses of treatment (Doses L and H) versus placebo. Treatment effect on at least one dose must be significant.
Primary endpoint: Positive and Negative Symptoms Scale (PANSS) total score
Two patient populations: General population and subpopulation (based on a genotypic classifier)

[5] Dmitrienko, A., Offen, W. W., and Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics in medicine, 22(15), 2387-2400.

Parallel Gatekeeping

Family 1

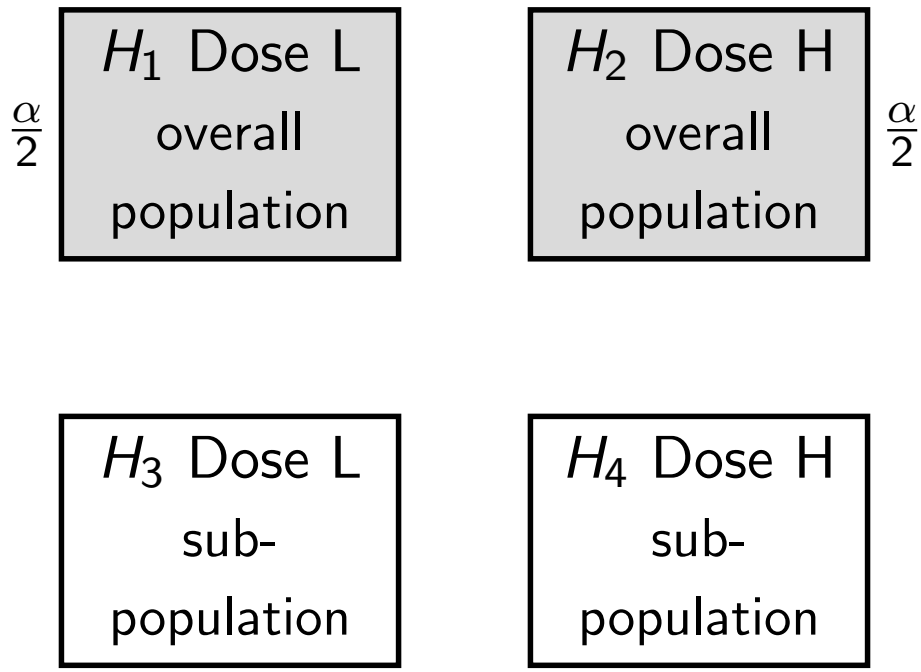
Family 2



Family 1: $\{H_1, H_2\}$ Doses L and H versus Placebo in overall population
Family 2: $\{H_3, H_4\}$ Doses L and H versus Placebo in subpopulation

Parallel Gatekeeping

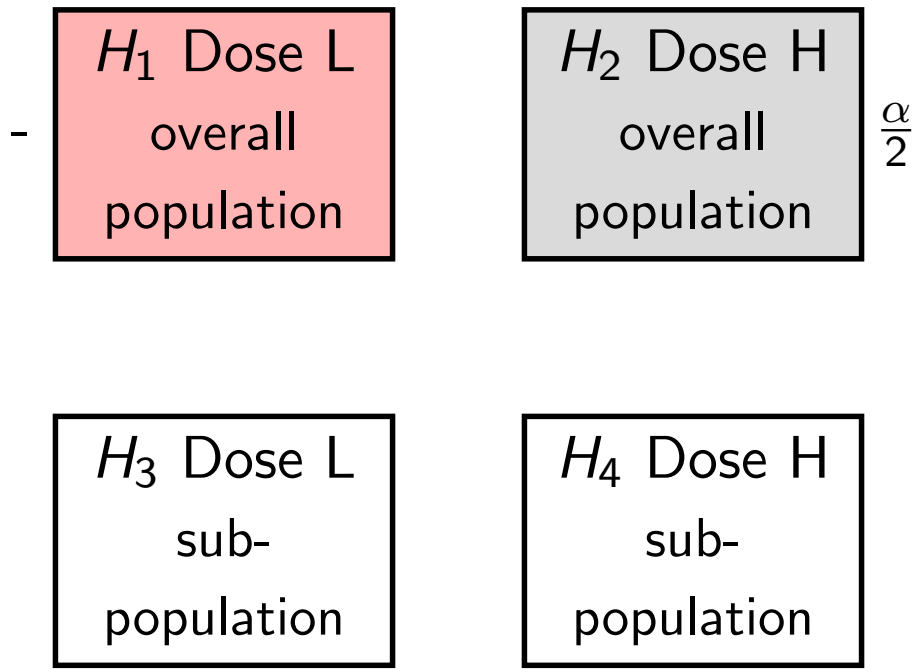
Start test H_1 and H_2 at $\alpha/2$



secondary endpoints are tested if at least one primary test is significant

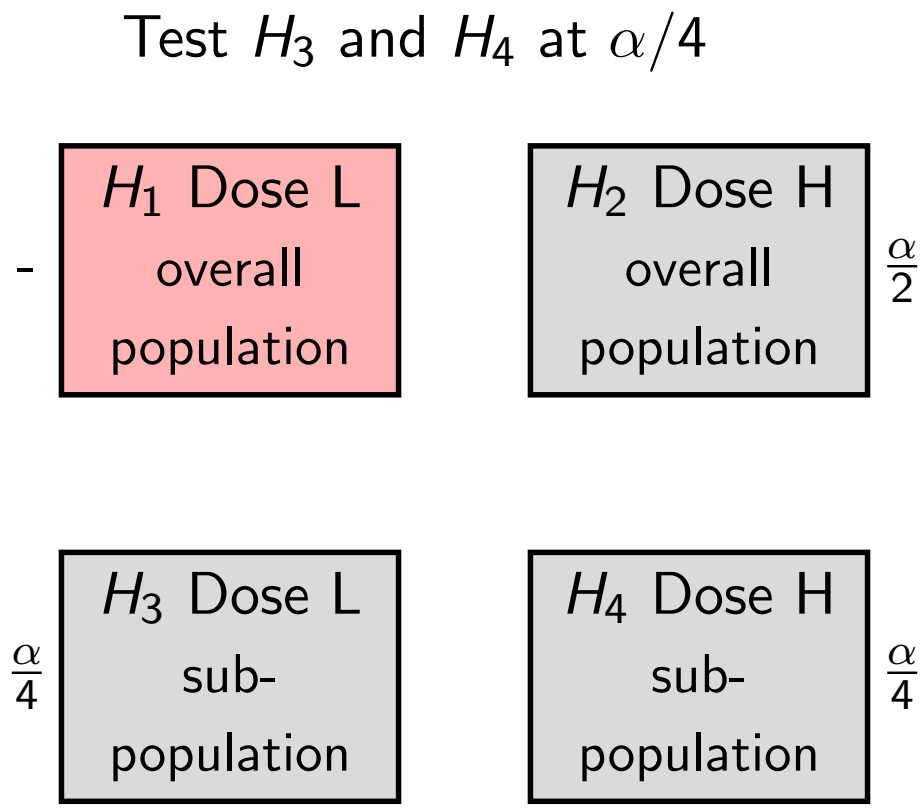
Parallel Gatekeeping

Suppose $p_1 < \alpha/2$



secondary endpoints are tested if at least one primary test is significant

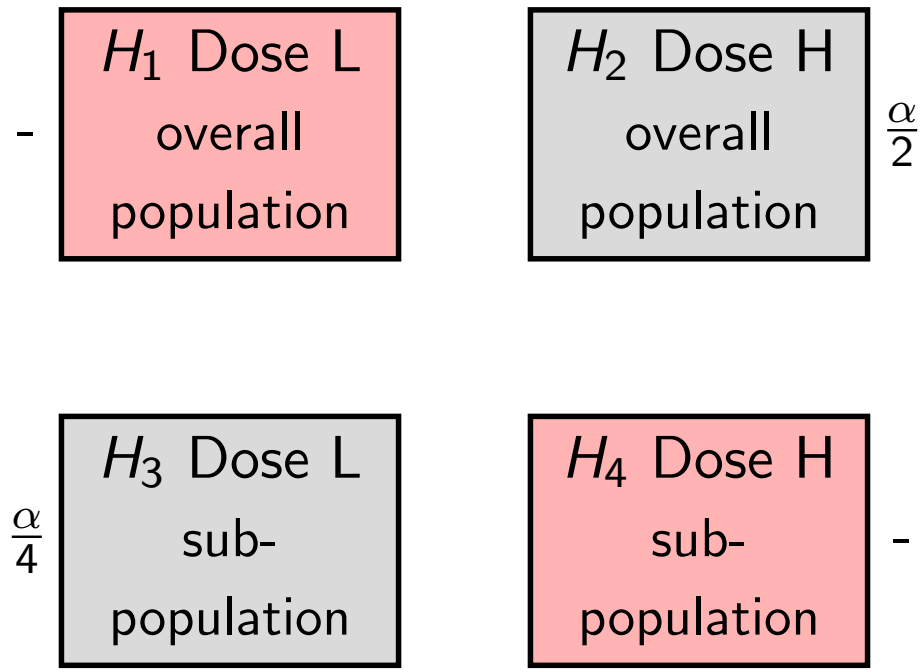
Parallel Gatekeeping



secondary endpoints are tested if at least one primary test is significant

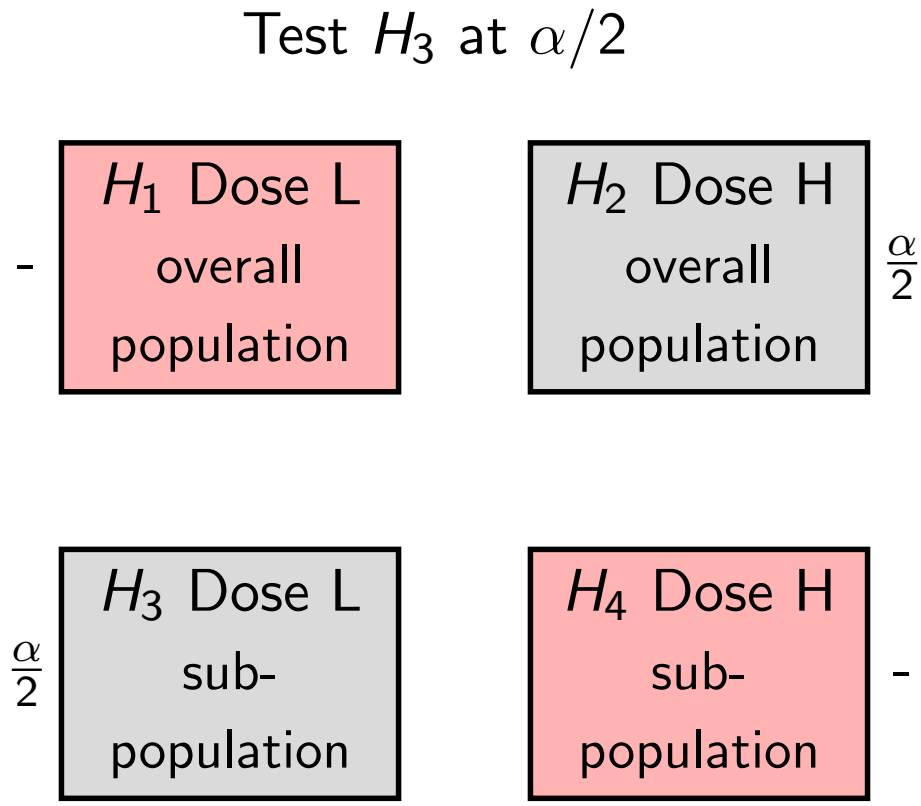
Parallel Gatekeeping

Suppose $p_4 < \alpha/4$



secondary endpoints are tested if at least one primary test is significant

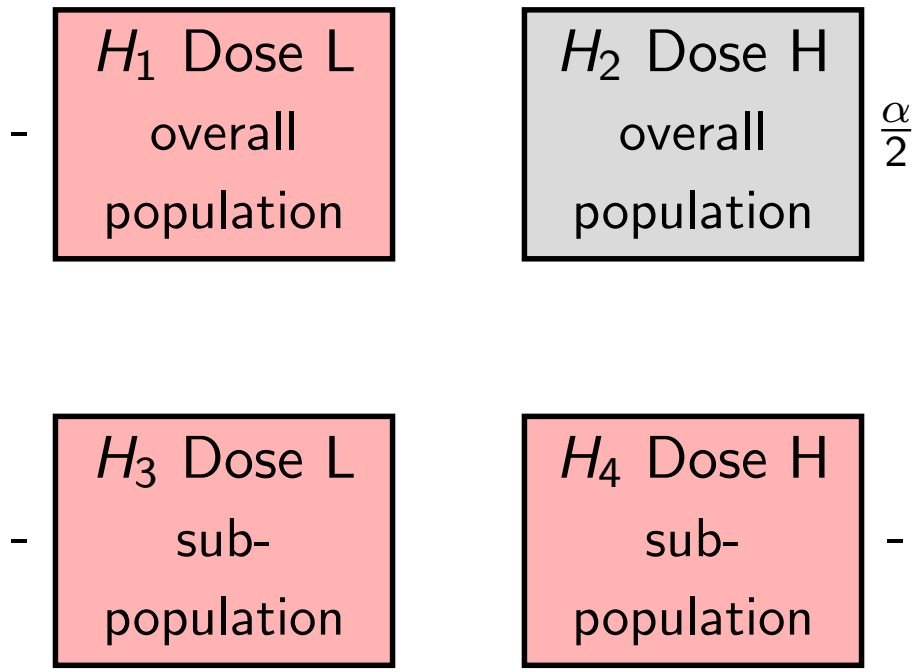
Parallel Gatekeeping



secondary endpoints are tested if at least one primary test is significant

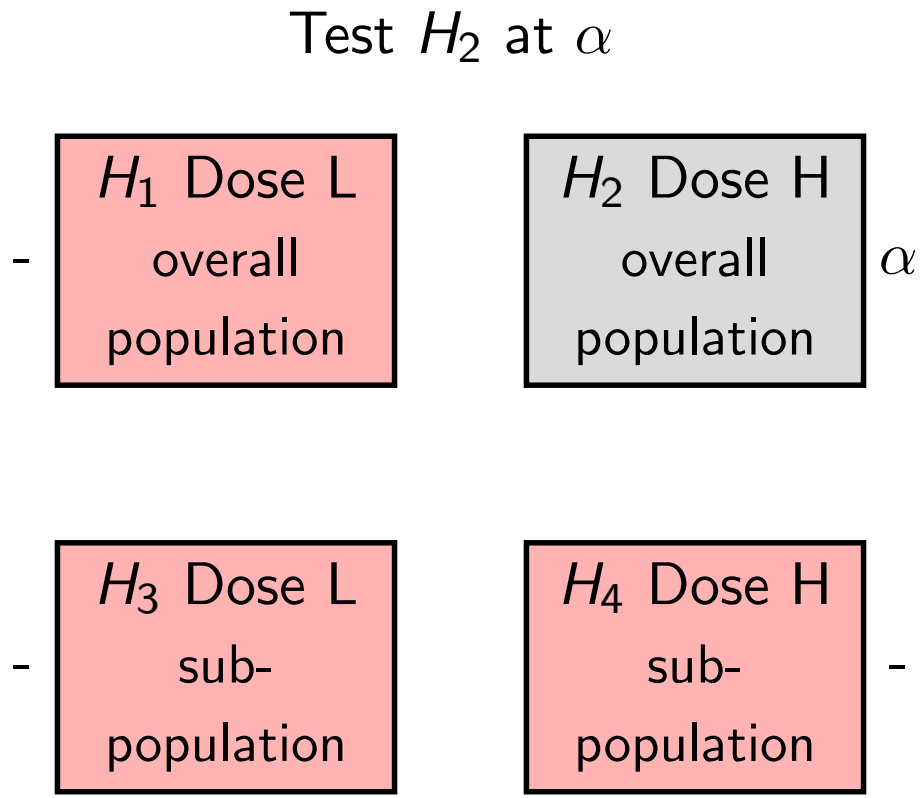
Parallel Gatekeeping

Suppose $p_3 < \alpha/2$



secondary endpoints are tested if at least one primary test is significant

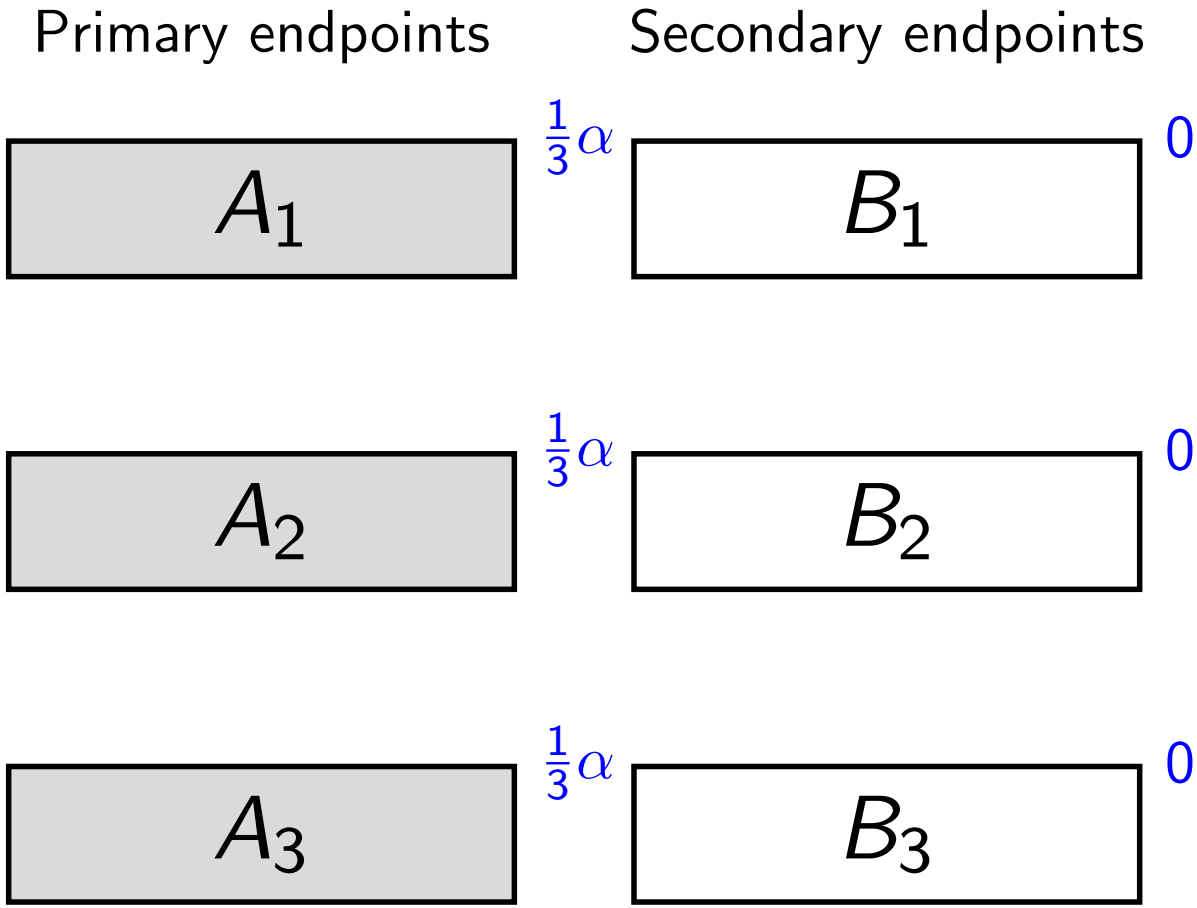
Parallel Gatekeeping



secondary endpoints are tested if at least one primary test is significant

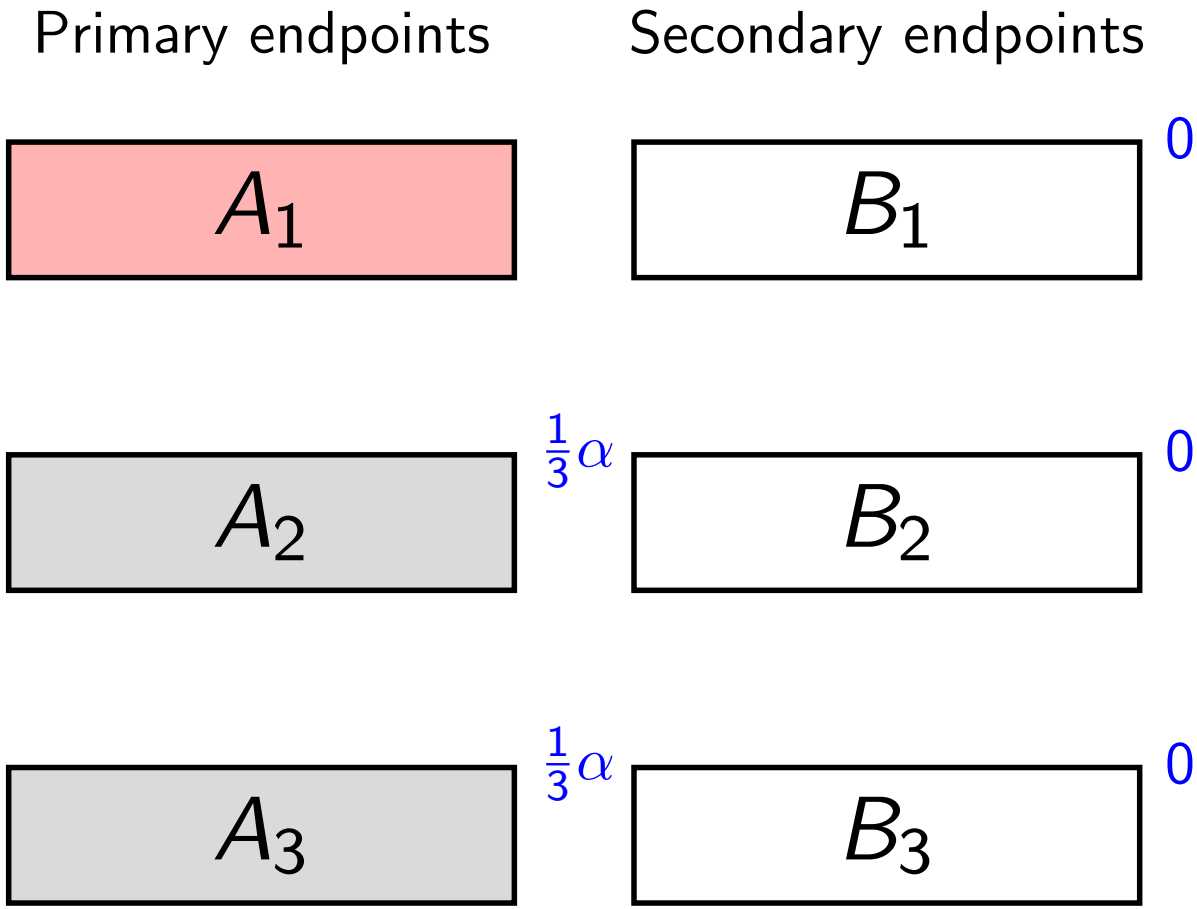
Parallel Gatekeeping

Test all primary endpoints at $\alpha/3$



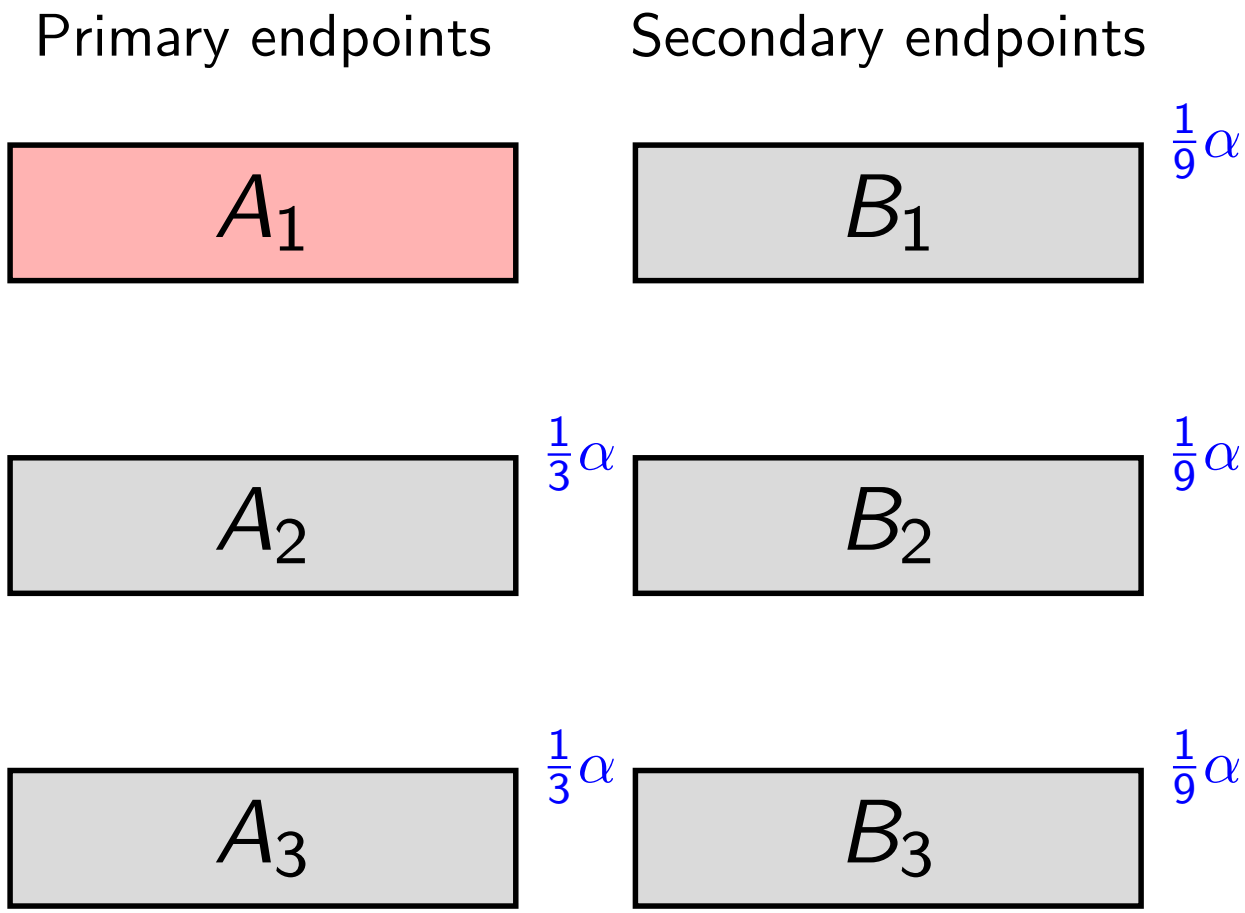
Parallel Gatekeeping

If we reject a few. . .



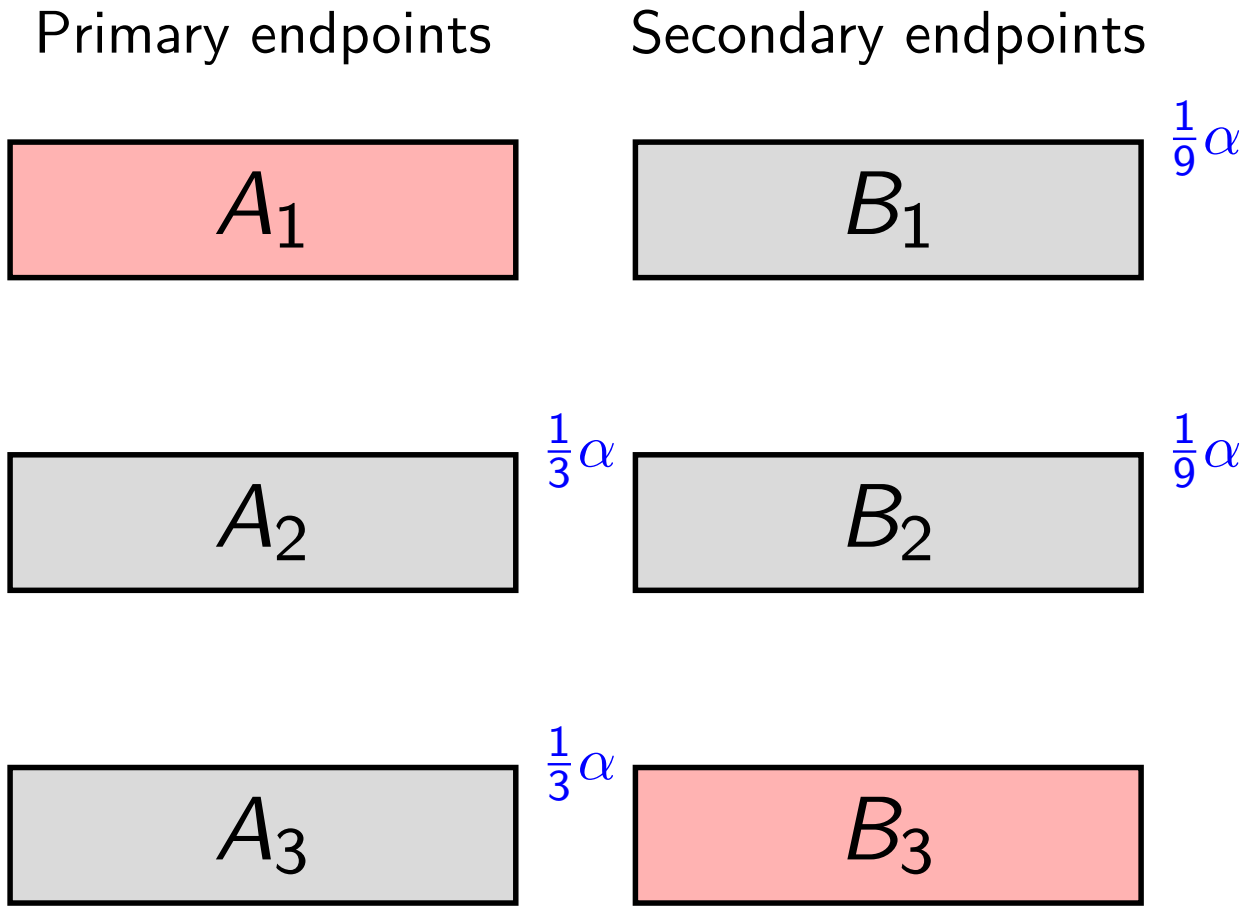
Parallel Gatekeeping

Go on with the secondary endpoints with the available α



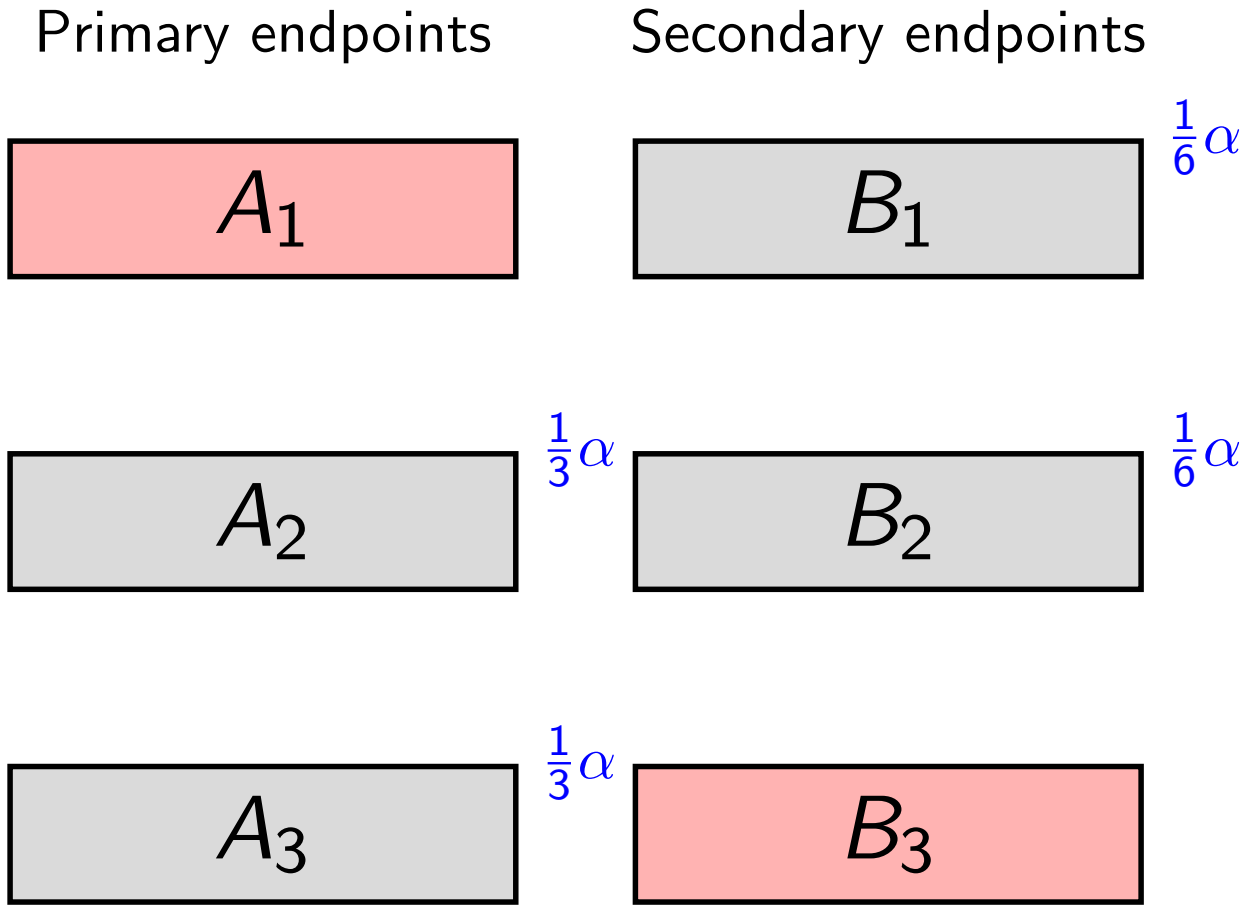
Parallel Gatekeeping

Suppose we are able to reject some of the secondary endpoints. . .



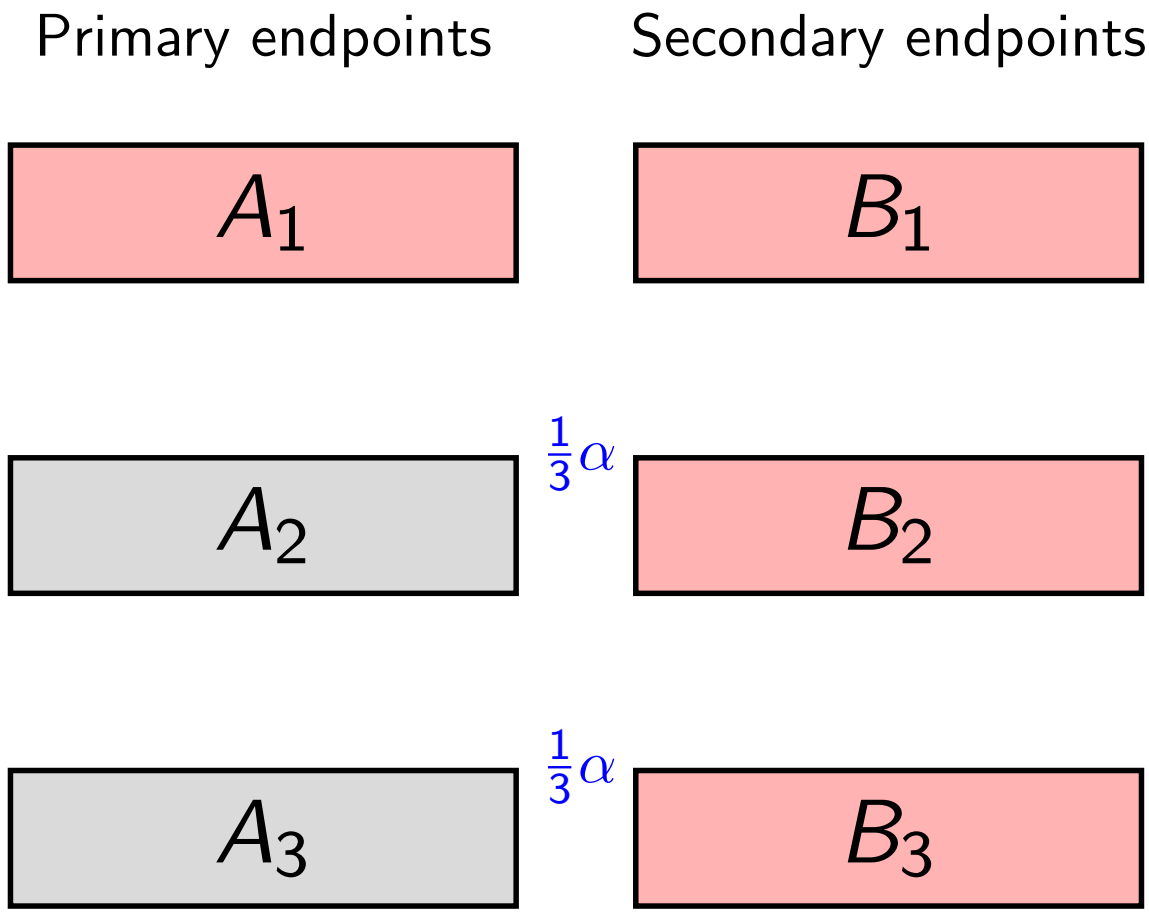
Parallel Gatekeeping

Go on doing Holm for the secondary endpoints



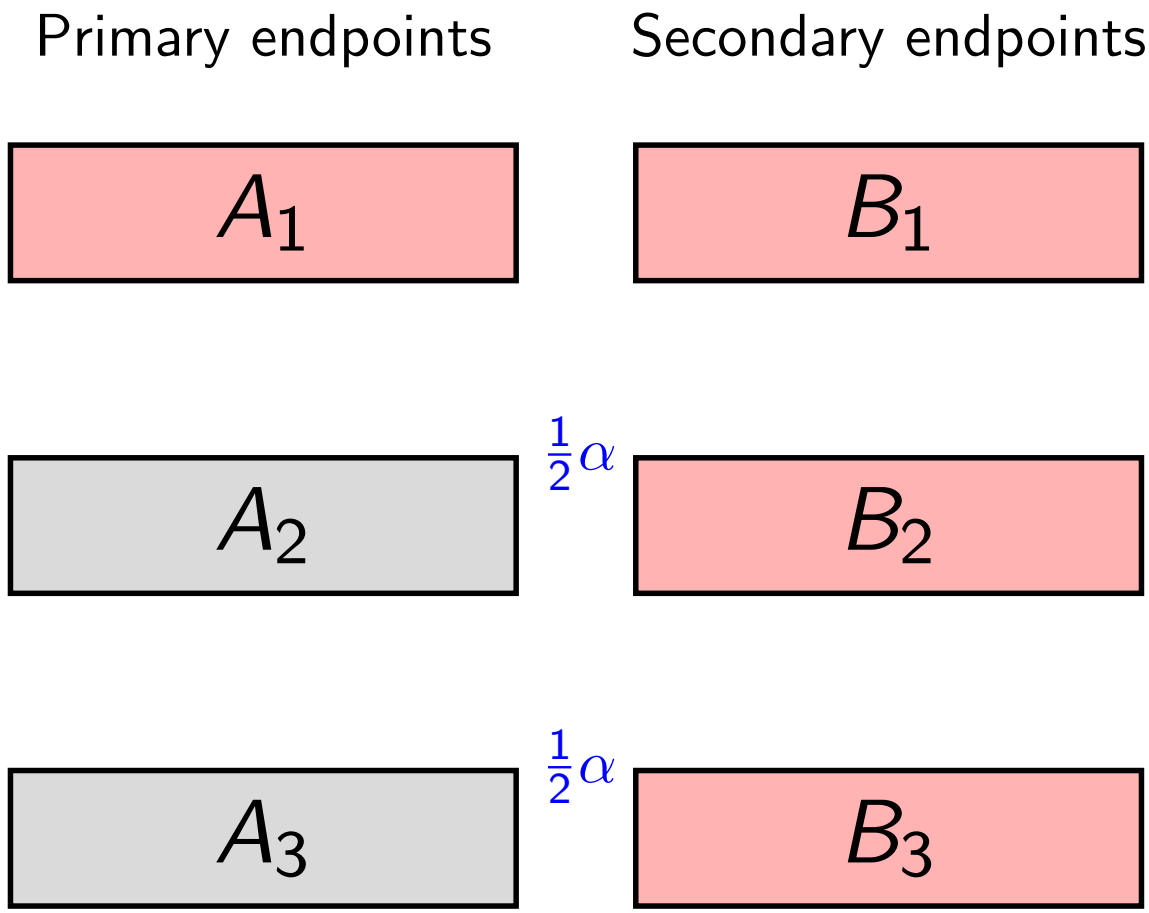
Parallel Gatekeeping

And if we reject all secondary ones. . .



Parallel Gatekeeping

Go on doing Holm for the primary endpoints



Intersection-Union test

Intersection-Union testing arises naturally in studies when a significant outcome with respect to two or more objectives is required in order to declare the study successful.

- **All** analyses must show benefit
- The trial's outcome is positive if all analyses produce a significant outcome

$$\boxed{H_I : \cup_{i=1}^{m} H_i \quad \text{versus} \quad H_U : \cap_{i=1}^{m} K_i}$$

$\rightarrow$ **no multiplicity adjustment!**

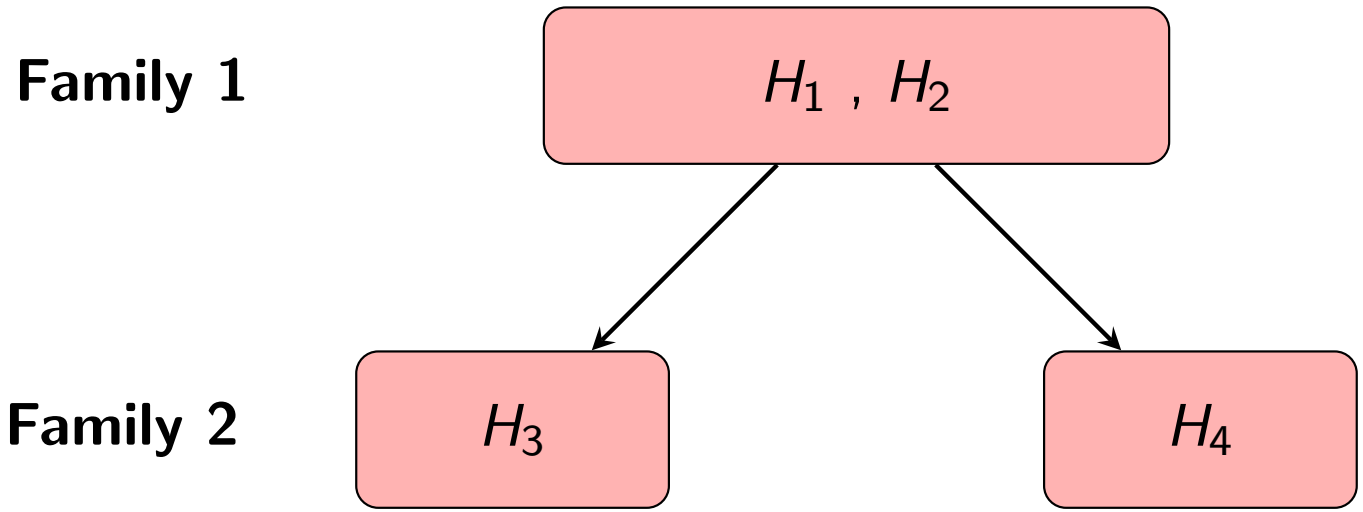
Serial Gatekeeping [6]

Family 1 is a **serial gatekeeper** for Family 2 $\rightarrow$ All hypotheses must be rejected in Family 1 to proceed to Family 2.

Example: Alzheimer's diseases trial
Objective: Evaluate the effects of a treatment on cognition and global changes in patients with mild to moderate Alzheimer's disease
Design: Treatment versus placebo
Primary endpoints: Endpoint 1: Cognition endpoint (ADAS-Cog), Endpoint 2: Clinical global scale (CIBIC plus)
Treatment effect on both endpoints must be significant
Secondary endpoints; Endpoint 3: Biochemical marker, Endpoint 4: Imaging marker

[6] Westfall, P. H., Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. Journal of Statistical Planning and Inference, 99(1), 25-40.

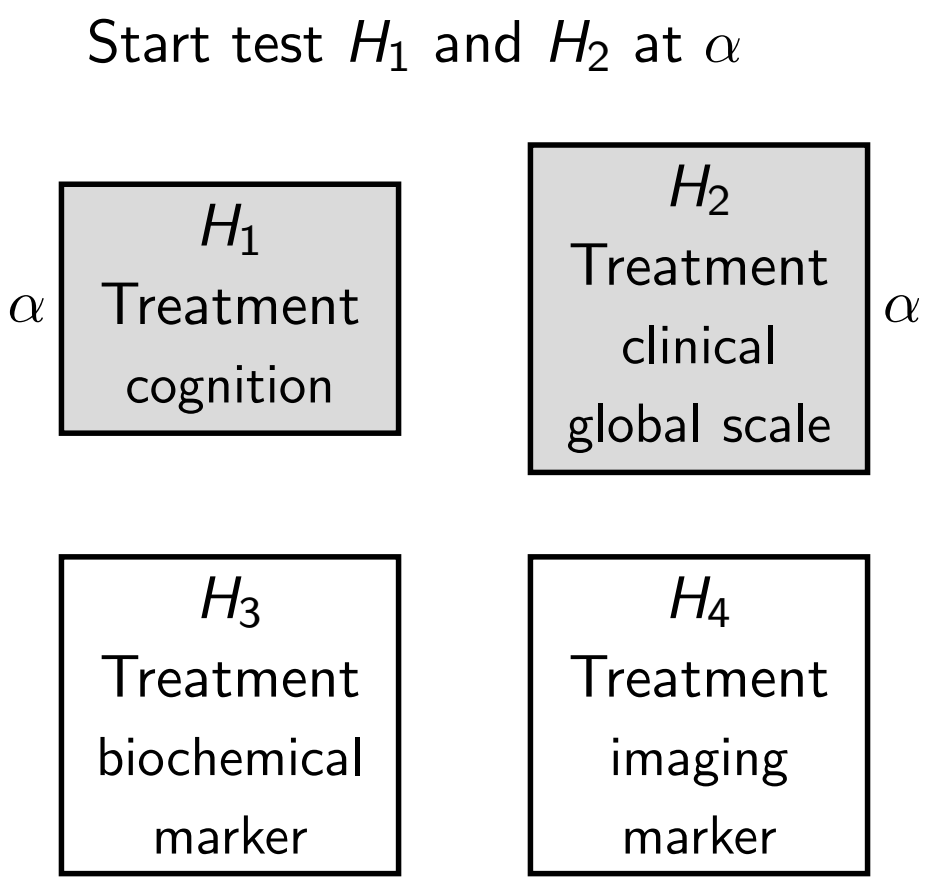
Serial Gatekeeping



Family 1: $\{H_1, H_2\}$ Treatment versus Placebo on cognition and clinical global scale

Family 2: $\{H_3, H_4\}$ Treatment versus Placebo on biochemical and imaging marker

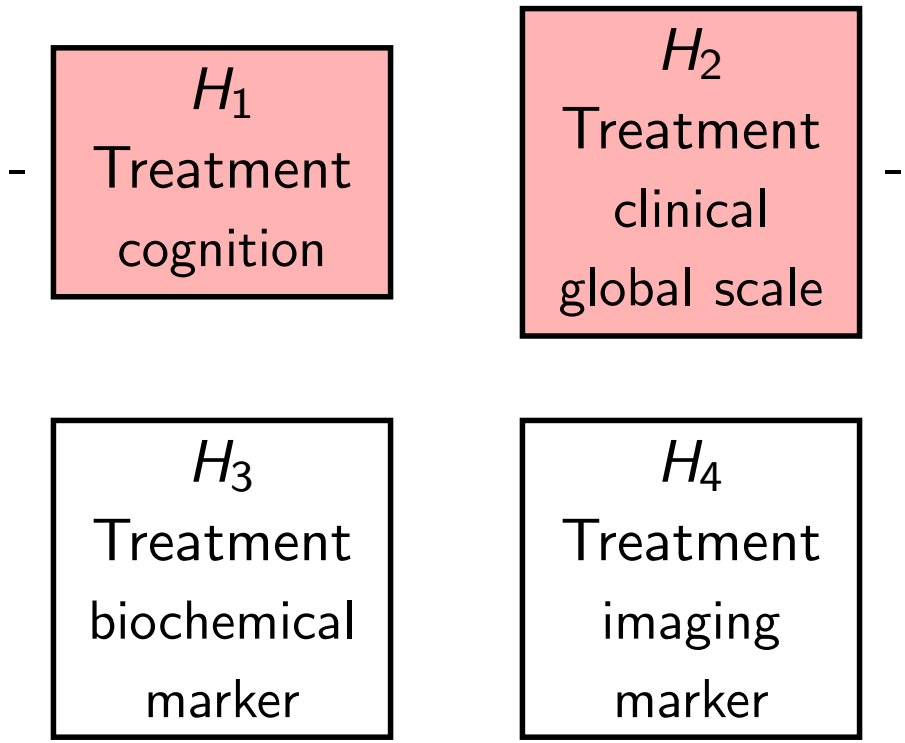
Serial Gatekeeping



secondary endpoints are tested if all primary tests are significant

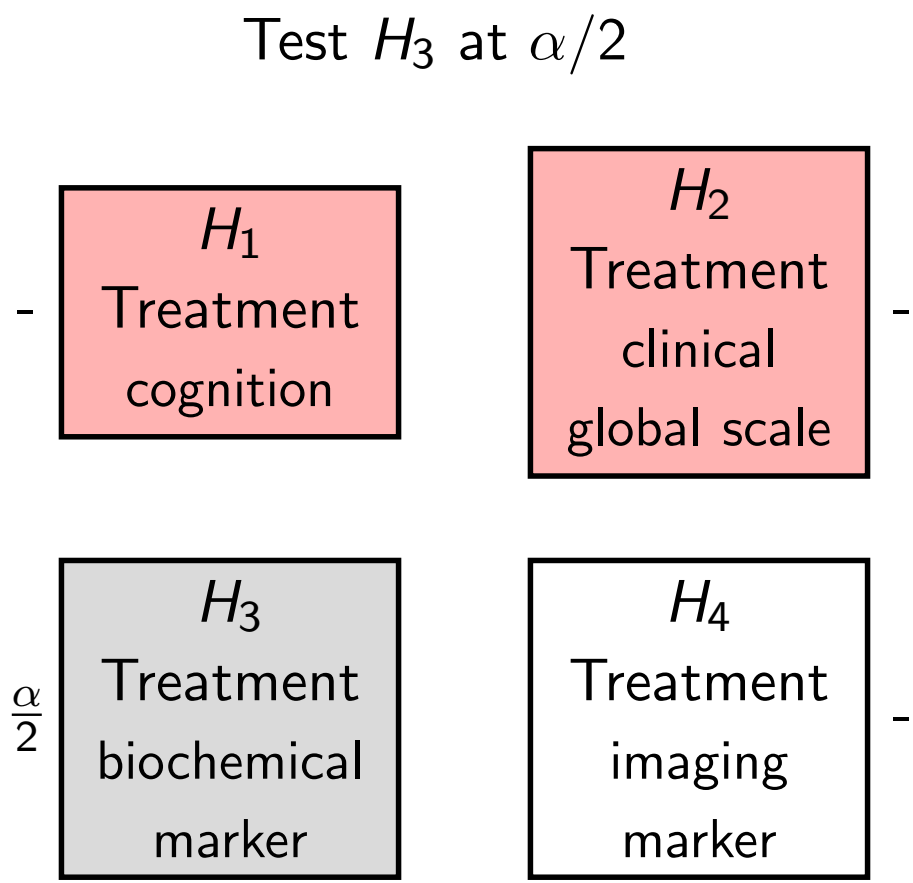
Serial Gatekeeping

Suppose $p_1 < \alpha$ and $p_2 < \alpha$



secondary endpoints are tested if all primary tests are significant

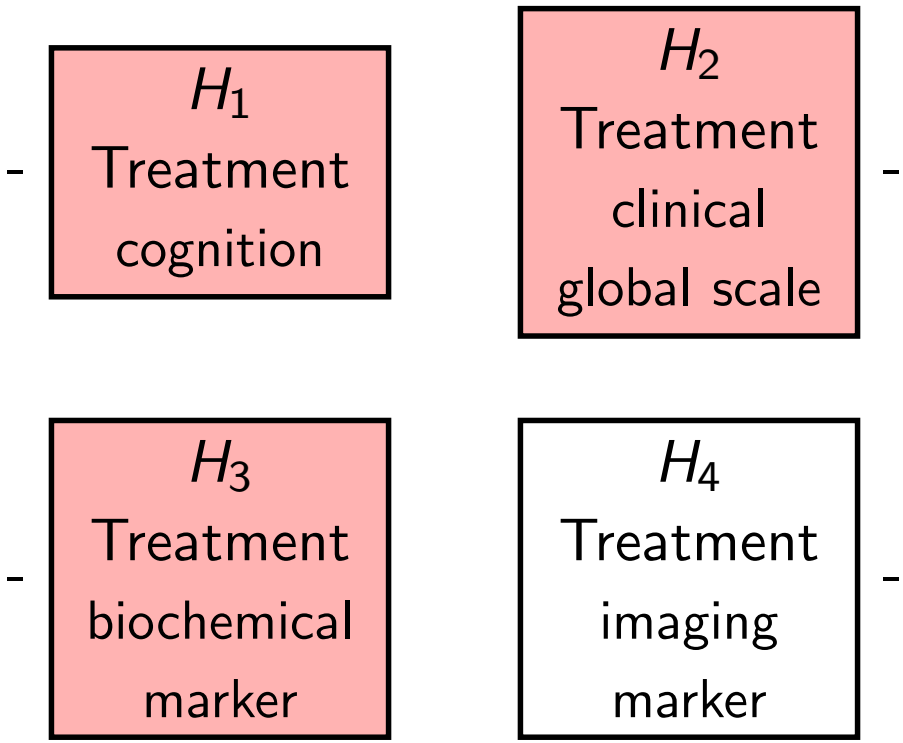
Serial Gatekeeping



secondary endpoints are tested if all primary tests are significant

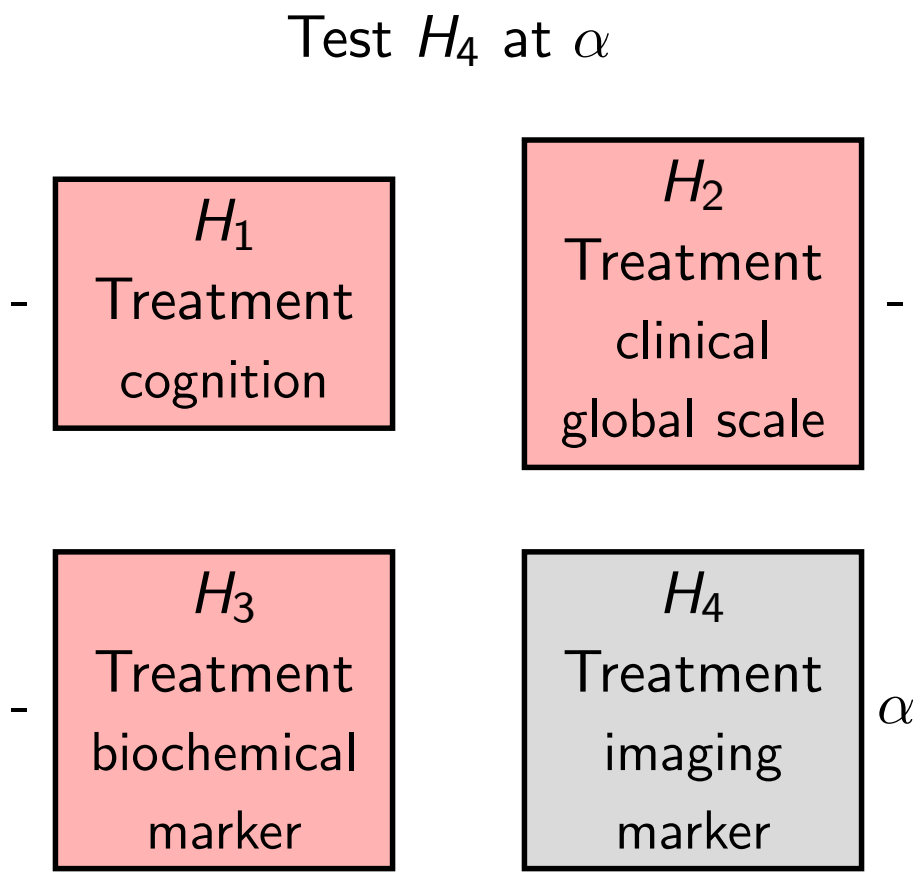
Serial Gatekeeping

Suppose $p_3 < \alpha/2$



secondary endpoints are tested if all primary tests are significant

Serial Gatekeeping



secondary endpoints are tested if all primary tests are significant

Outline

1. FamilyWise Error Rate (FWER)
2. Bonferroni (single-step)
3. Holm (step-wise)
4. Closed Testing
5. Gatekeeping strategies
6. **Summary**

Summary

Family-Wise Error

- Generalizes Type I errors to the case of **multiple hypotheses**

Summary

Family-Wise Error

- Generalizes Type I errors to the case of **multiple hypotheses**
- Controls the probability of **at least one** false rejection among all rejections

Summary

Family-Wise Error

- Generalizes Type I errors to the case of **multiple hypotheses**
- Controls the probability of **at least one** false rejection among all rejections
- **Adjusts p-values** (adjusted p-values are always equal to or worse than the unadjusted p-values)

Summary

Family-Wise Error

- Generalizes Type I errors to the case of **multiple hypotheses**
- Controls the probability of **at least one** false rejection among all rejections
- **Adjusts p-values** (adjusted p-values are always equal to or worse than the unadjusted p-values)
- Understand the **hierarchical/sequential structure** of hypotheses testing $\rightarrow$ increase power

Summary

Family-Wise Error

- Generalizes Type I errors to the case of **multiple hypotheses**
- Controls the probability of **at least one** false rejection among all rejections
- **Adjusts p-values** (adjusted p-values are always equal to or worse than the unadjusted p-values)
- Understand the **hierarchical/sequential structure** of hypotheses testing $\rightarrow$ increase power

R Software

- Bonferroni and Holm: `library(stats); p.adjust()`
- Closed Testing: `library(cherry); closed()`
- Post-hoc and more: `library(multcomp); glht()`