Gatekeeping Testing Strategies in Clinical Trials

March, 2020

Gatekeeping

Why Gatekeeping?

Assumptions and Notation

Types of Gatekeeping

Serial Gatekeeping

Parallel Gatekeeping

General Gatekeeping

Multiplicity

- Multiplicity is omnipresent in clinical trials and causes Type I error inflation.
- Proper multiplicity adjustment is necessary to control Type I error inflation via control of FWER especially in confirmatory clinical trials.
- Single-step and more powerful stepwise multiple test procedures are easy to use to deal with standard multiple endpoints/multiple dose comparisons.
- Complex multiple test procedures, called gatekeeping procedures, are required when hypotheses are hierarchically ordered and logically related.

Closed test procedures

- In the category "closed testing“ lie:

1. O’Brien-type
2. fixed-sequence methods
3. gatekeeping methods,
4. dose-response methods
5. co-primary multiple endpoints
6. simultaneous multiple doses
7. graphical methods.

Closed test procedures

- In the simple case of three hypotheses H_1, H_2, and H_3, e.g. comparing three groups or considering three co-primary endpoints.
- There are eight of states of nature that are possible. All three hypotheses could be true, any of the three pairs could be true, any of the three individual hypotheses may be true, or none could be true.
- Closure-based testing begins by forming the set of all intersections null hypotheses of the individual (or elementary) hypotheses H_i.
- Rejection of an elementary hypothesis requires rejection of all intersection hypotheses H_I that "include" H_i in the intersection.
- Any α-level test may be used to test the intersections H_I.

Example: three hypotheses

1. Test each hypothesis H_1, H_2, H_3 using an appropriate α-level test.
2. Create the "closure" of the set, which is the set of all possible intersections among H_1, H_2, H_3, in this case the hypotheses H_{12}, H_{13}, H_{23}, and H_{123}.
3. Test each intersection using an appropriate α-level test. These tests could be F-tests, MANOVA tests, or in general any test that is valid for the given intersection. (There are many possibilities for testing these intersection hypotheses, and each method for testing intersections results in a different closed testing procedure. We present and compare seven such procedures below.)
4. You may reject any hypothesis $H_{i,}$ when the following conditions both hold

- The test of H_i itself yields a statistically significant result, and
- The test of every intersection hypothesis that includes H_i is statistically significant.

We illustrate the method using

```
data mult;
   input G Y1 Y2 Y3;
   datalines;
0 14.4 7.00 4.30
0 14.6 7.09 3.88
0 13.8 7.06 5.34
0 10.1 4.26 4.26
0 11.1 5.49 4.52
0 12.4 6.13 5.69
0 12.7 6.69 4.45
1 11.8 5.44 3.94
1 18.3 1.28 0.67
1 18.0 1.50 0.67
1 20.8 1.51 0.72
1 18.3 1.14 0.67
1 14.8 2.74 0.67
1 13.8 7.08 3.43
1 11.5 6.37 5.64
1 10.9 6.26 3.47
;
```

Our first closed testing method uses basic t-tests for the component hypotheses, and Hotelling's T^2 test (refer, for example, to Johnson and Wichern, 1998, p. 302-306) for the composites, computed as follows using PROC REG:

```
proc reg data=mult;
   model Y1 Y2 Y3 = G;
   H1: mtest Y1;
   H2: mtest Y2;
   H3: mtest Y3;
   H12: mtest Y1, Y2;
   H13: mtest Y1, Y3;
   H23: mtest Y2, Y3;
   H123: mtest Y1, Y2, Y3;
run;
```

Each MTEST statement produces a test statistic and p-value. The following diagram lists the p-values for the hypotheses, arranged in a hierarchical fashion to better illustrate the closed testin method.

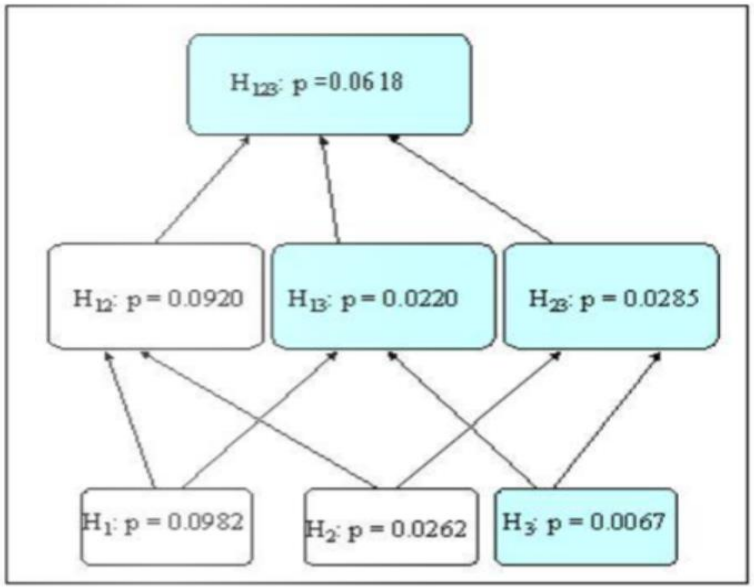
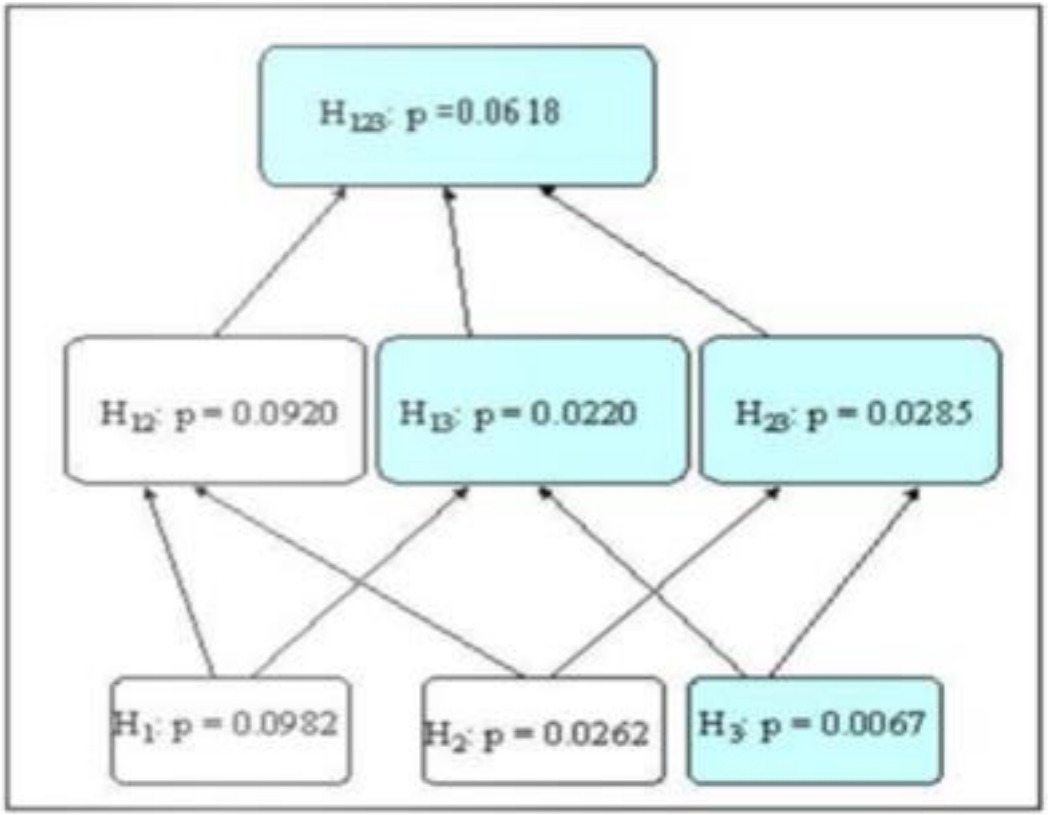


Illustration of the closed testing method using Hotelling's T^2 tests

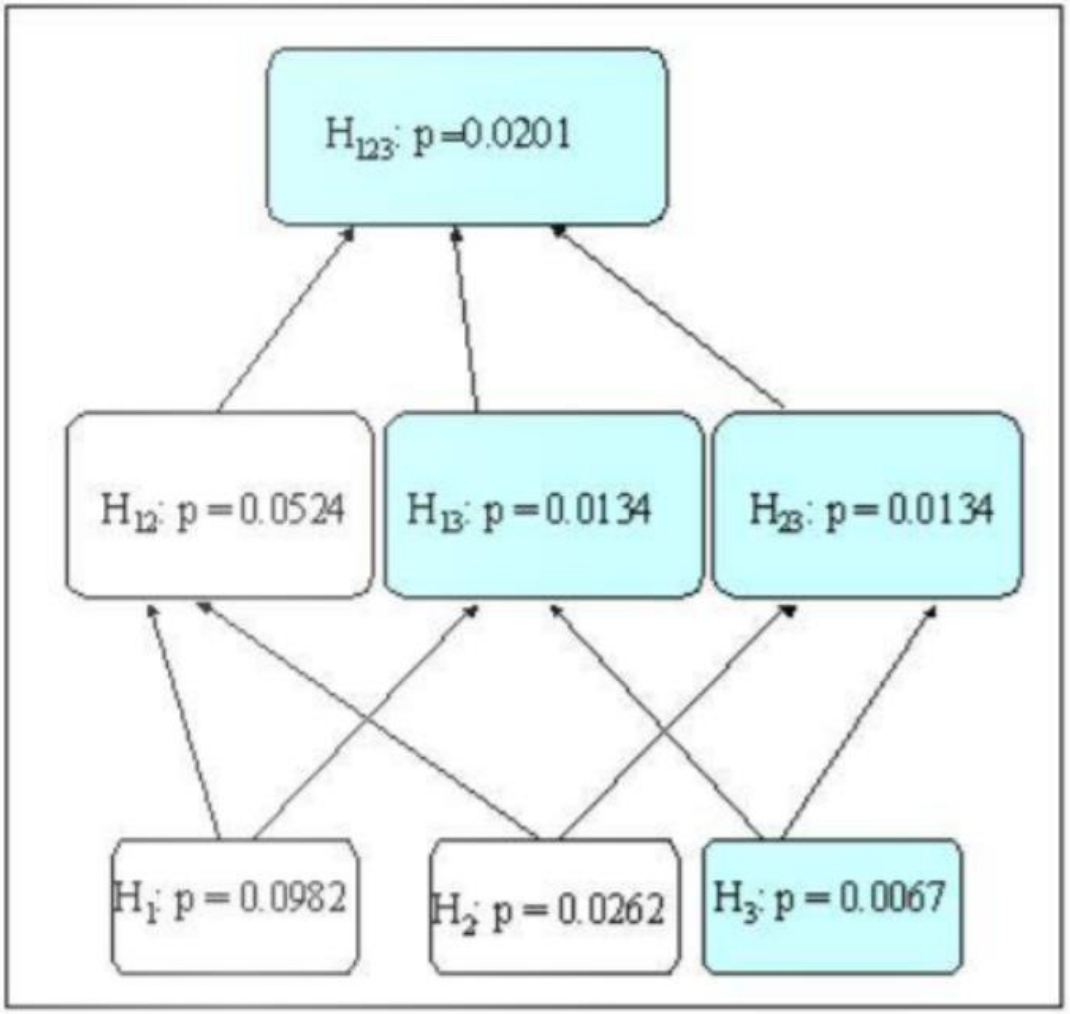
Definition: When using a closed testing procedure, the *adjusted p-value* for a given hypothesis H_i is the maximum of all p-values for tests that include H_i as a special case (including the p-value for the H_i test itself).

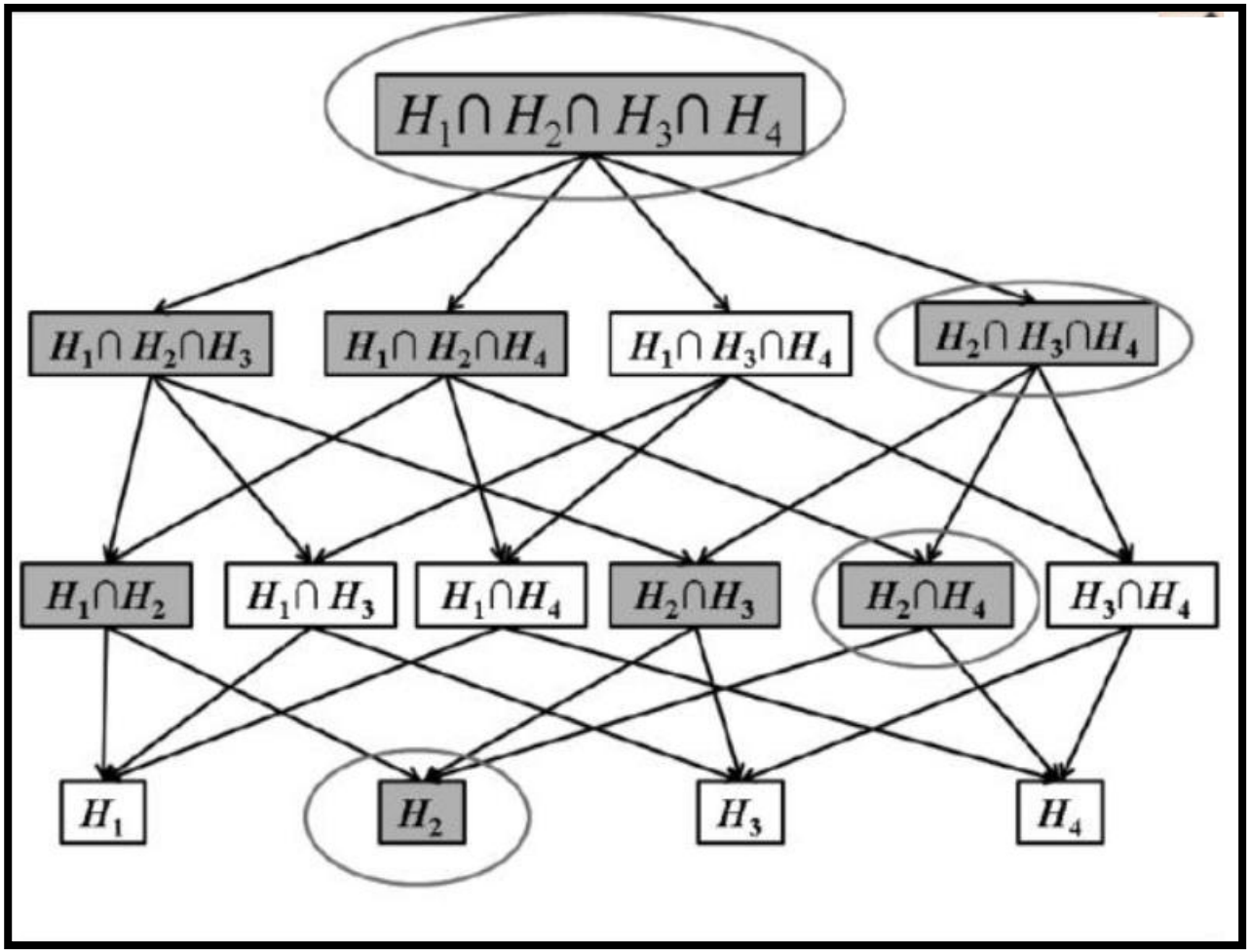
The adjusted p-value for testing H_3 is, therefore, formally computed as max(0.0067, 0.0220, 0.0285, 0.0618) = 0.0618.

Hotelling T3



Bonferroni-Holm





The closure hierarchy for $m = 4$ hypotheses illustrating the shortcut. All circled hypotheses must be rejected if H_2 is to be rejected

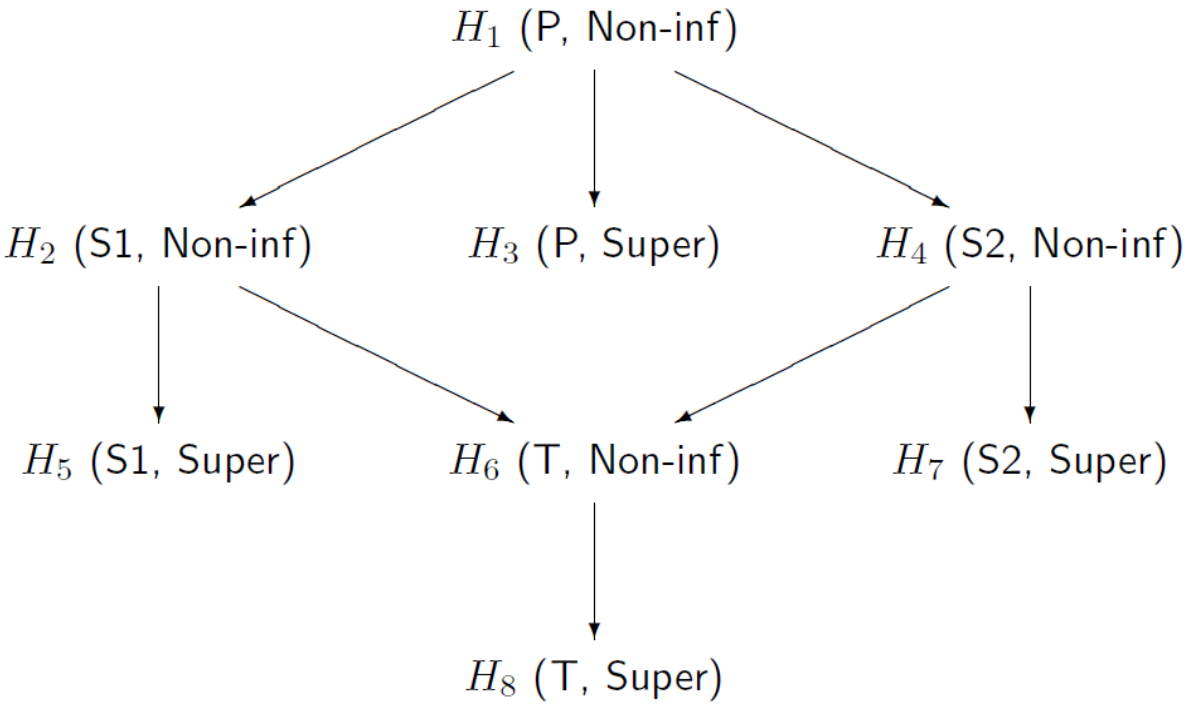
Why Gatekeeping?

- Clinical trials often involve multiple hierarchically ordered hypotheses with logical restrictions, e.g., multiple endpoints, multiple patient subgroups, noninferiority-superiority tests.
- Sponsors like to enrich product labels by additional claims.
- O'Neill (1997): "Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance."
- CPMP Points to Consider Document (2002): "Additional claims... [for] secondary variables... are possible only after the primary objective of the clinical trial has been achieved, and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy."
- FDA Multiplicity Guidance Document 2017.

Introductary Example

- Primary endpoint (P): Mean reduction in systolic blood pressure.
- Two secondary endpoints (S1 and S2): Mean reduction in diastolic blood pressure and proportion of patients with controlled systolic/diastolic blood pressure.
- Tertiary endpoint (T): Average blood pressure based on ambulatory blood pressure monitoring.
- Test superiority conditional on showing noninferiority for each endpoint subject to their hierarchical ordering.

Introductary Example



- Obviously, we want the global end result to be significant at some predefined alpha leven e.g. 5%
- How do we achieve that? The answer is Gate keeping!

Gatekeeping Procedures

- Gatekeeping procedures are used in the situation when there are multiple analyses (e.g. endpoints) and these are grouped into different families.

- With a gatekeeping procedure, the families are tested in a sequential manner and the tests for subsequent families will be performed only if the tests for the previous family are significant. In other words, the families of hypotheses examined earlier serve as GATEKEEPERS.

- Gatekeeping procedures preserve the overall false positive rate

- While the term ‘gatekeeping procedure’ may not have been used, this approach has been implemented in many clinical trials, especially in the regulatory setting. It is e.g. very typical that the secondary endpoints will only be tested only if the primary endpoint is tested significantly.

- In this way, the alpha-level for primary efficacy endpoints will be tested at alpha=0.05 level and not be compromised due to the consideration of the secondary endpoints.

Primary versus secondary findings

Dilemma

- regulatory agencies and pharmaceutical companies have long debated what secondary findings should be included in the product label
- regulatory agencies are concerned that pharmaceutical companies tend to present favorable data and ignore unfavorable data

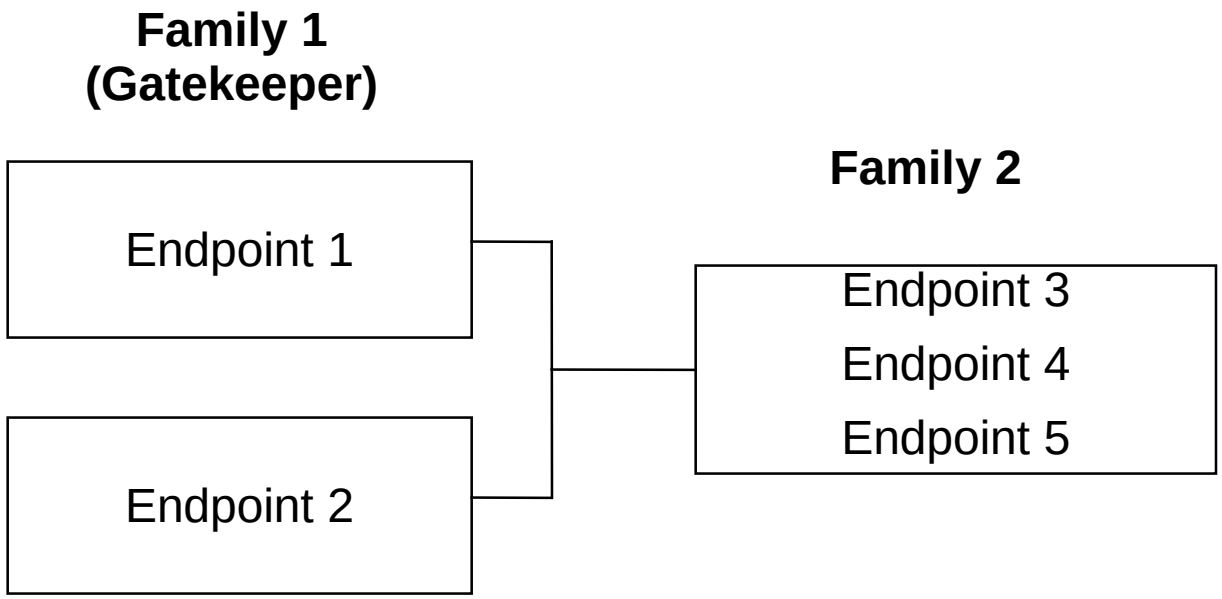
Gatekeeper strategies offer one potential solution to the dilemma

Example primary/Secondary

- A clinical trial will typically have one or more primary endpoints (family for primary endpoints) and have multiple secondary endpoints (family for secondary endpoints).
- If there are many secondary endpoints, the secondary endpoints can be further divided into multiple secondary different families.
- In this way, the alpha-level for primary efficacy endpoints will be tested at alpha=0.05 level and not be compromised due to the consideration of the secondary endpoints.

Gatekeeping Procedures

- An example with 2 primary and one secondary endpoints



Notation

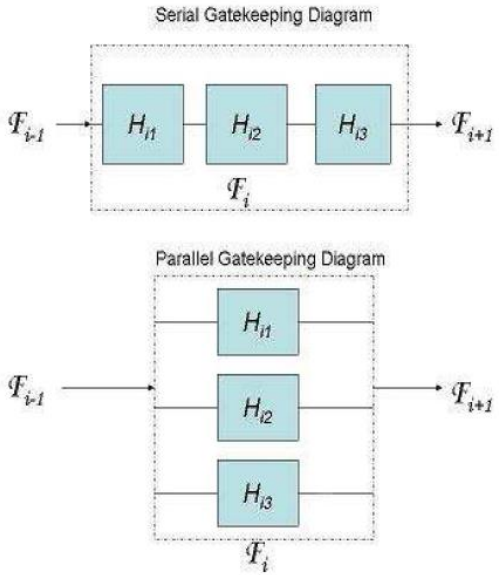
- $n \geq 2$ hypotheses, $H_1, \ldots, H_n$, grouped into $m \geq 2$ ordered families $F_1, \ldots, F_m$.
- Family $F_j = \{H_i : i \in N_j\}$ where $N_1 = \{1, \ldots, n_1\}, N_j = \{n_1 + \ldots + n_{j-1} + 1, \ldots, n_1 + \ldots + n_j\}$.
- Family F_j consists of n_j hypotheses with $\sum_{j=1}^{m} n_j = n$.
- F_j is a gatekeeper for F_{j+1}, $j = 1, 2, \ldots, m-1$.
- Strong control of FWER:

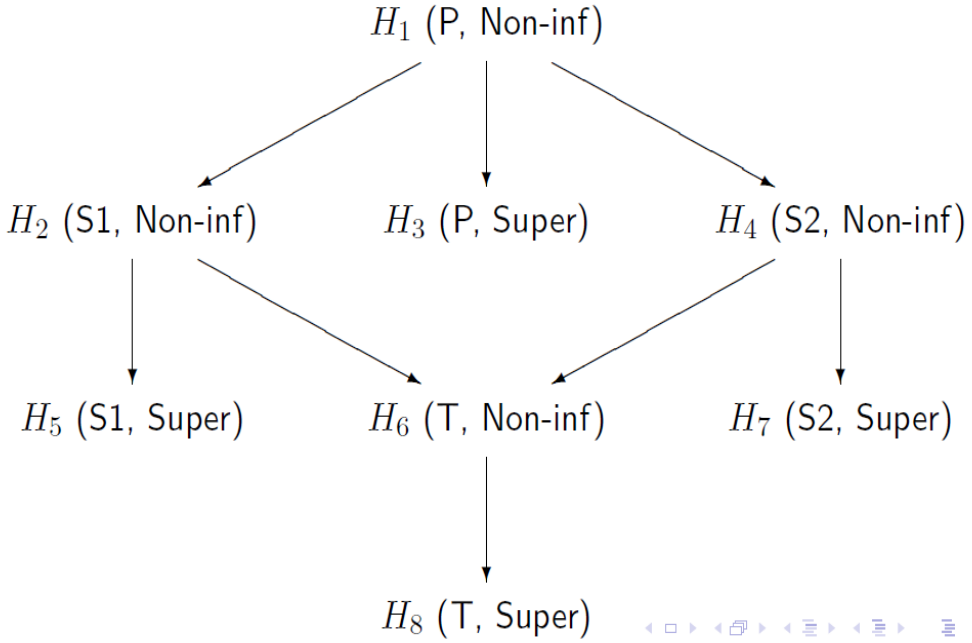
$$\text{FWER} = P\{\text{Reject at least one true } H_i\} \leq \alpha.$$

- Independence Condition: Inferences on $H_i \in F_j$ don't depend on inferences on $H_i \in F_k$ for $k > j$ (desirable but not essential).

Types of Gatekeeping

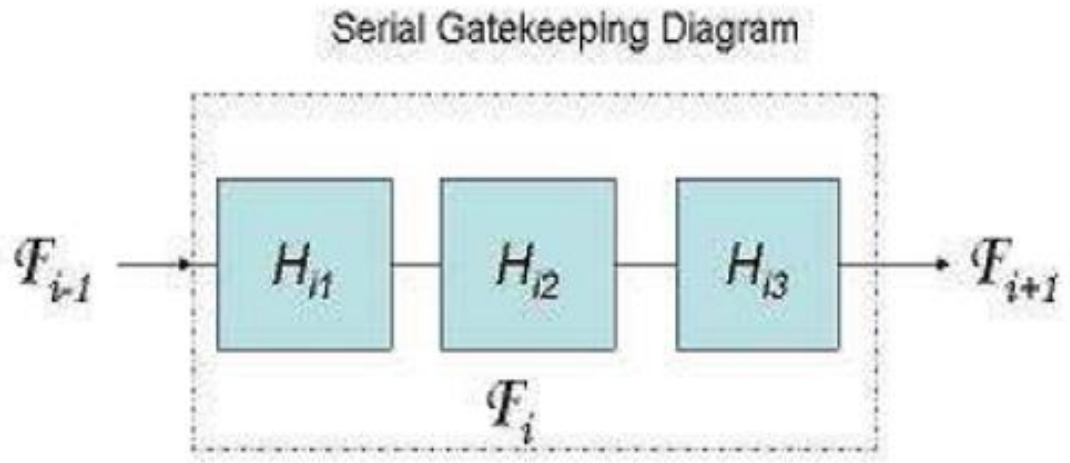
- If the gatekeeper F_j is passed then hypotheses in F_{j+1} are *testable* (i.e., they must be tested to make accept/reject decision); otherwise all hypotheses in F_k for $k > j$ are *non-testable* (i.e., are automatically accepted).





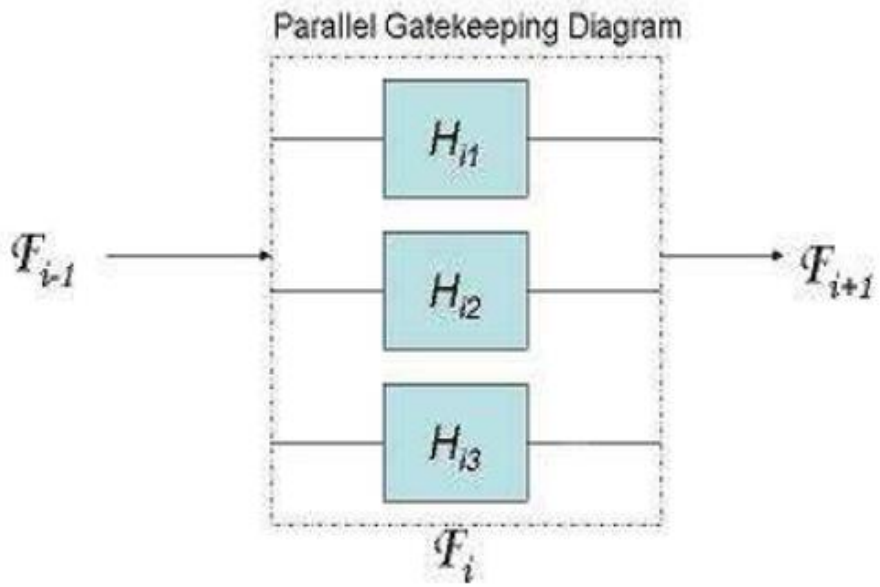
Serial Gatekeeping

- Serial gatekeeping: Gatekeeper F_j is passed iff all $H_i \in F_j$ are rejected (Maurer, Hothorn & Lehmacher 1995).



Parallel Gatekeeping

- Dmitrienko, Westfall & Offen (2003), Dmitrienko, Tamhane, Wang & Chen (2006), Guilbaud (2007), Dmitrienko, Tamhane & Wiens (2008).



- Parallel gatekeeping: Gatekeeper F_j is passed iff at least one $H_i \in F_j$ is rejected (Dmitrienko, Offen & Westfall 2003).

Examples

- Serial gatekeeping example: Alzheimer disease trial
  - Primary endpoints: (i) Alzheimer disease assessment scale - Cognitive subscale (ADAS-COG), (ii) Clinical global impression change (CGIC). Both must be significant.
  - Secondary endpoints: Biochemical and imaging markers
- Parallel gatekeeping example: Osteoporosis trial in post-menopausal women
  - Primary endpoints: (i) Incidence of new vertebral fractures, (ii) Incidence of new invasive breast cancer
  - Secondary endpoint: Incidence of new non-vertebral fractures

  At least one primary should be significant to proceed to secondary…

General Gatekeeping

- More complex clinical decision rules involving objectives that do not fit in simple serial/parallel framework
  - Based on the closed testing principle (Marcus et al, 1976)
  - Focus on strategies derived using Bonferroni's test
  - Easily extended to more powerful tests that account for the correlation among the endpoints (Dunnett's test, resampling tests)

Tree-structured gatekeeping: Dmitrienko, Wiens, Tamhane & Wang (2007).

Mixture gatekeeping: Dmitrienko & Tamhane, A.C. (2011a, 2011b), Dmitrienko, Kordzakhia & Tamhane (2011).

Superchain procedures: Dmitrienko & Kordzakhia (2011)

General gatekeeping (Dmitrienko, Wiens, Tamhane & Wang 2007, Dmitrienko and Tamhane 2011a,b, Dmitrienko, Kordzkhia and Tamhane 2011).

Example

Diabetes Trial

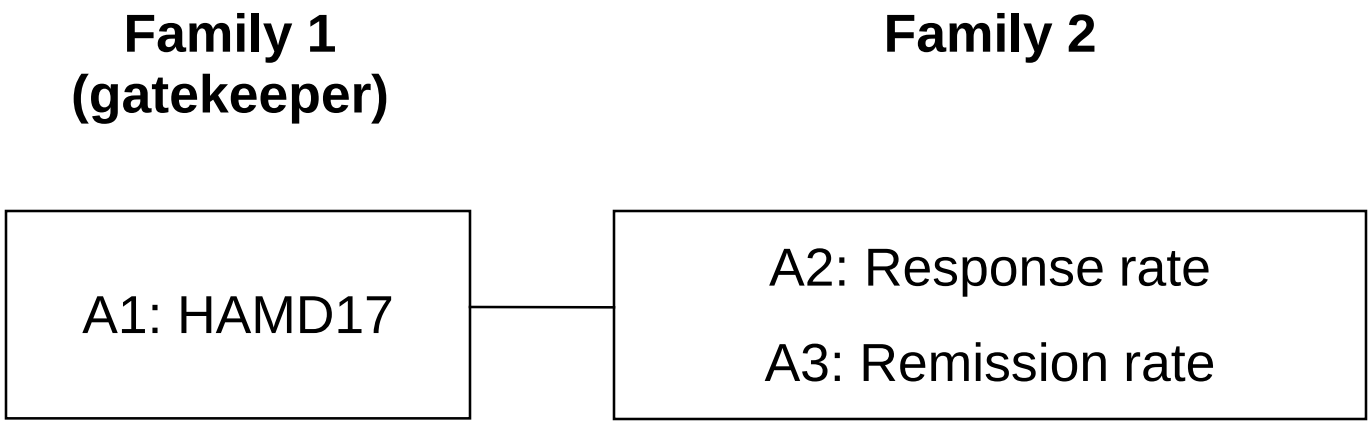
- Three Doses (High, Medium, Low) + Control with 3 Endpoints
- Primary endpoint: Hemoglobin A1c
- Secondary endpoint: Fasting serum glucose
- Tertiary endpoint: HDL cholesterol.
- For each dose, determine significant endpoints conditional on all higher-ranked endpoints being significant.

Example: One primary endpoint

Depression trial

- Experimental drug is compared to placebo
- Single primary endpoint
  - 17-item Hamilton depression rating scale (HAMD 17 score)
- Trial is declared successful if the drug is superior to placebo
- Two important secondary endpoints
  - response rate based on the HAMD 17 score
  - remission rate based on the HAMD 17 score
- Can the secondary findings be included in the product label?

Sequential gatekeeping strategy



Step 1: Perform the primary analysis

Step 2: Perform the secondary analyses with an adjustment for multiplicity if the primary analysis yielded a significant result

Sequential gatekeeping strategy

Endpoint	Raw p	Adjusted p
Primary: HAMD 17	0.046	0.046
Secondary: Response rate	0.048	0.048
Secondary: Remission rate	0.021	0.042

Primary analysis: No adjustment for multiplicity
Secondary analyses: Stepwise Holm's test

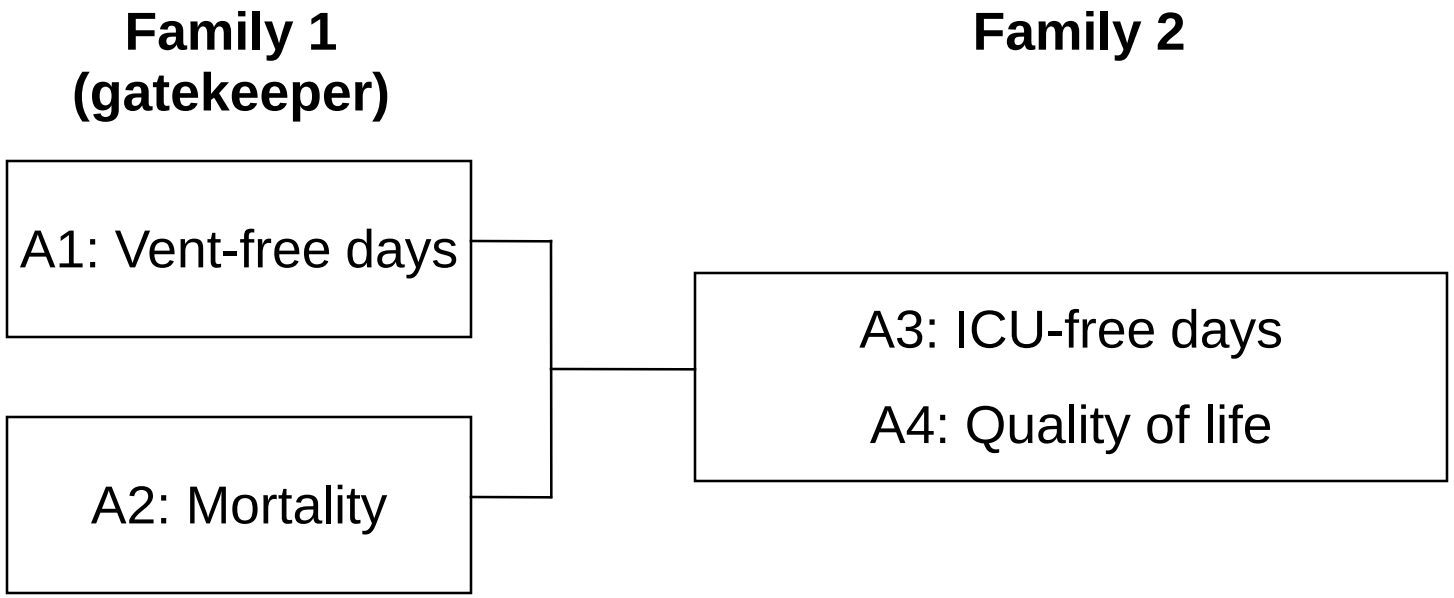
All primary and secondary findings are significant at 5% level

Example: Multiple primary endpoints

Clinical trial in patients with acute lung injury

- Experimental drug is compared to placebo
- Two primary endpoints
  - number of days patients are off mechanical ventilation (vent-free days)
  - 28-day all-cause mortality rate
- Trial is declared successful if the drug is superior to placebo with respect to either endpoint
- Two important secondary endpoints
  - number of days patients are out of Intensive Care Unit (ICU-free days)
  - overall quality of life at the end of the study
- Can the secondary findings be included in the product label?

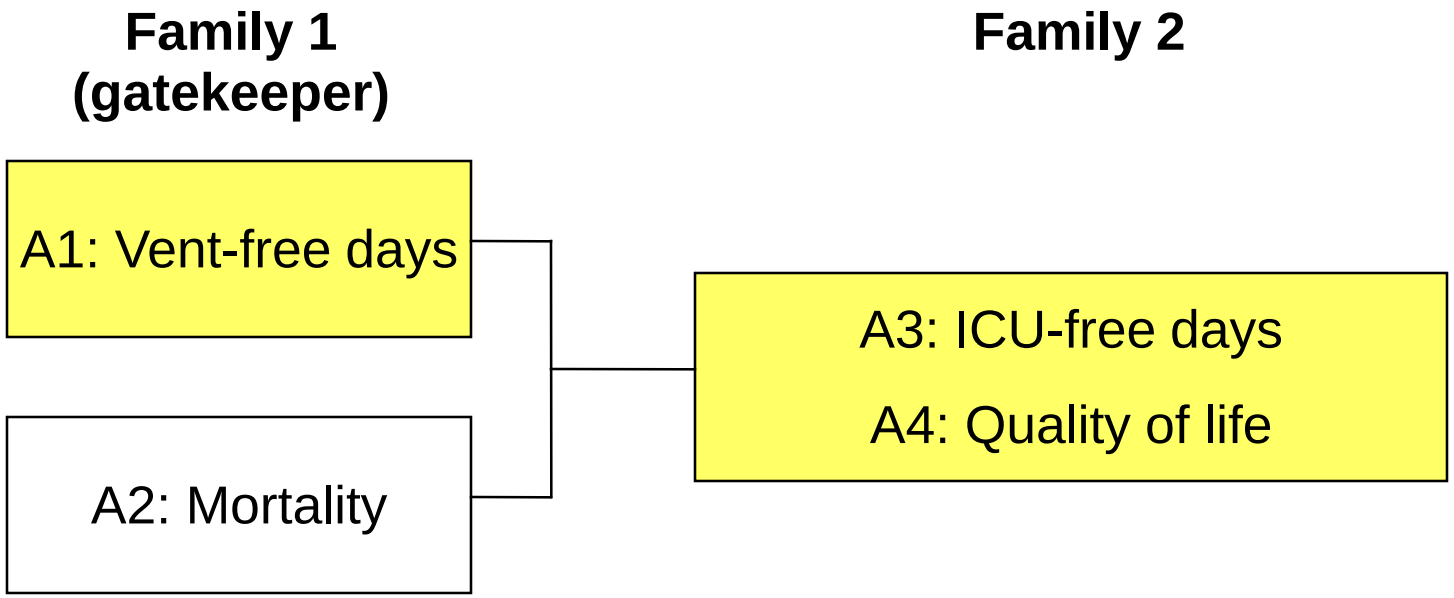
Parallel gatekeeping strategy



Step 1: Perform the primary analyses with an adjustment for multiplicity

Step 2: Perform the secondary analyses with an adjustment for multiplicity if at least one primary analysis yielded a significant result

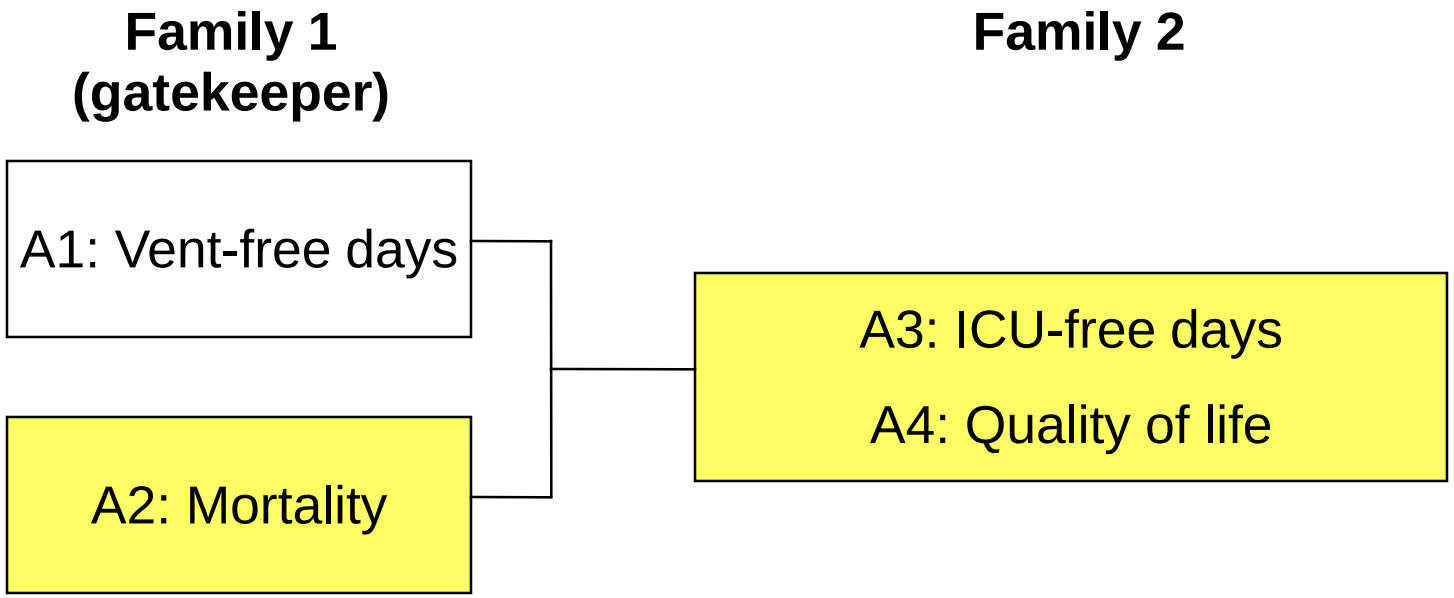
Parallel gatekeeping strategy



Step 1: Perform the primary analyses with an adjustment for multiplicity

Step 2: Perform the secondary analyses with an adjustment for multiplicity if at least one primary analysis yielded a significant result

Parallel gatekeeping strategy



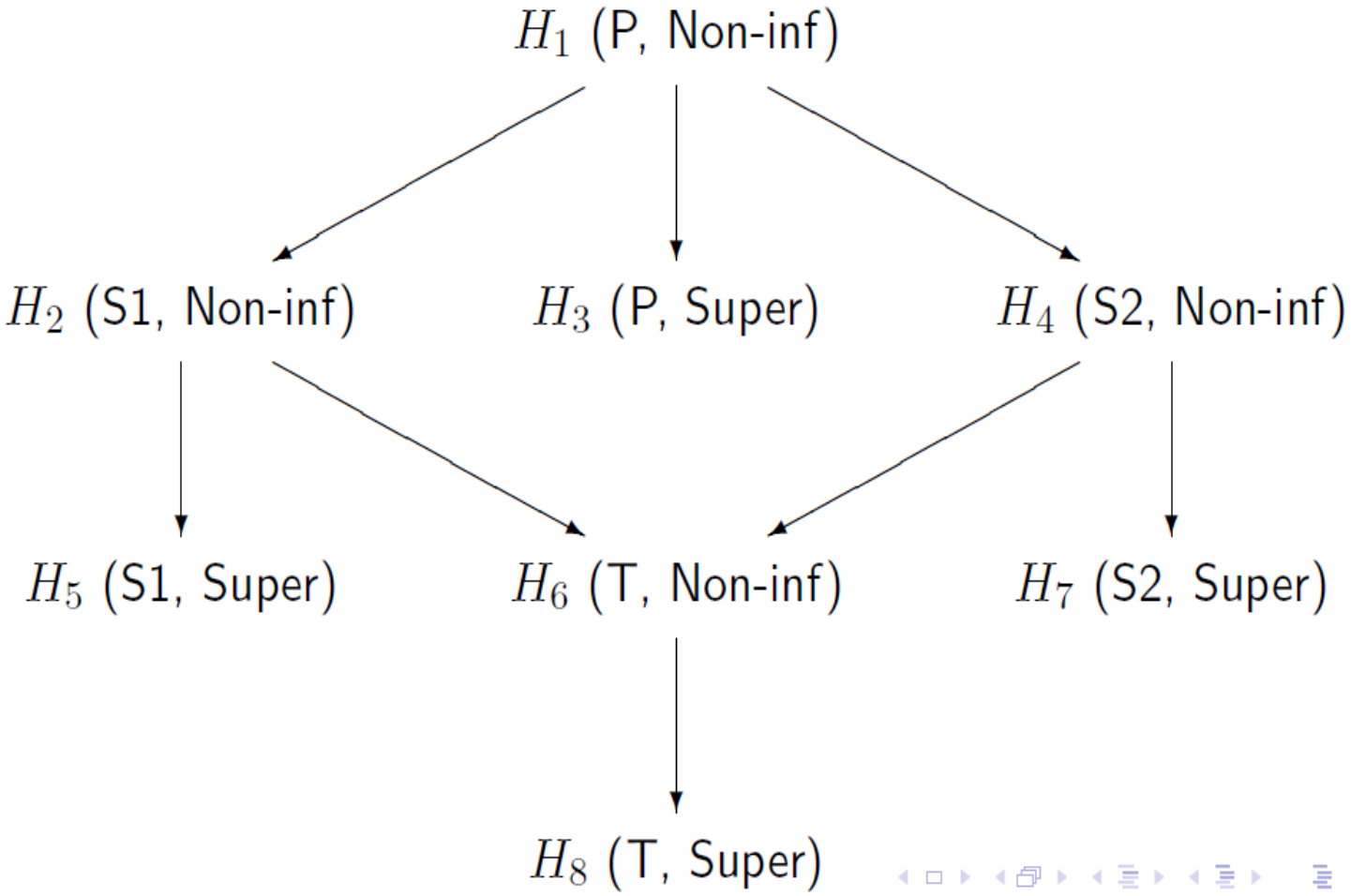
Step 1: Perform the primary analyses with an adjustment for multiplicity

Step 2: Perform the secondary analyses with an adjustment for multiplicity if at least one primary analysis yielded a significant result

Example

- Primary endpoint (P): Mean reduction in systolic blood pressure.
- Two secondary endpoints (S1 and S2): Mean reduction in diastolic blood pressure and proportion of patients with controlled systolic/diastolic blood pressure.
- Tertiary endpoint (T): Average blood pressure based on ambulatory blood pressure monitoring.
- Test superiority conditional on showing noninferiority for each endpoint subject to their hierarchical ordering.

Example

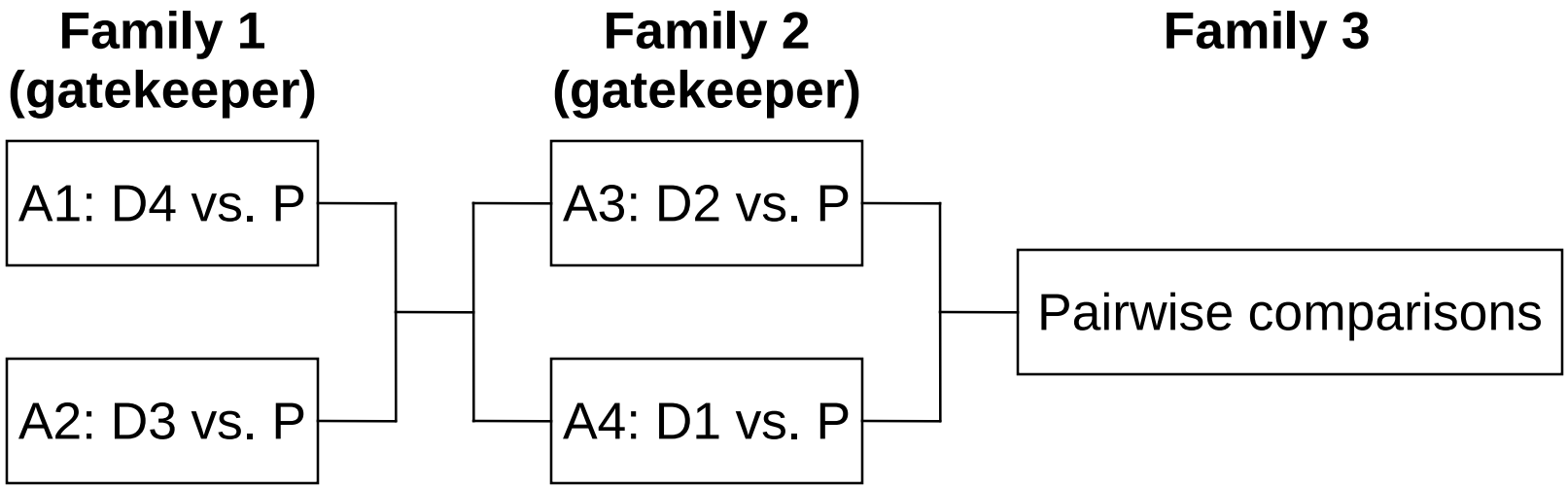

H_1 (P, Non-inf)
H_2 (S1, Non-inf)
H_3 (P, Super)
H_4 (S2, Non-inf)
H_5 (S1, Super)
H_6 (T, Non-inf)
H_7 (S2, Super)
H_8 (T, Super)

Example 3: Dose-finding study

Clinical trial in patients with hypertension

- Four doses of an experimental drug are compared to placebo
  - doses are labeled as D1, D2, D3 and D4
- Primary endpoint
  - reduction in diastolic blood pressure
- Objectives of the study
  - find the doses with a significant reduction in diastolic blood pressure compared to placebo
  - study the shape of the dose-response curve

Example 3: Dose-finding study

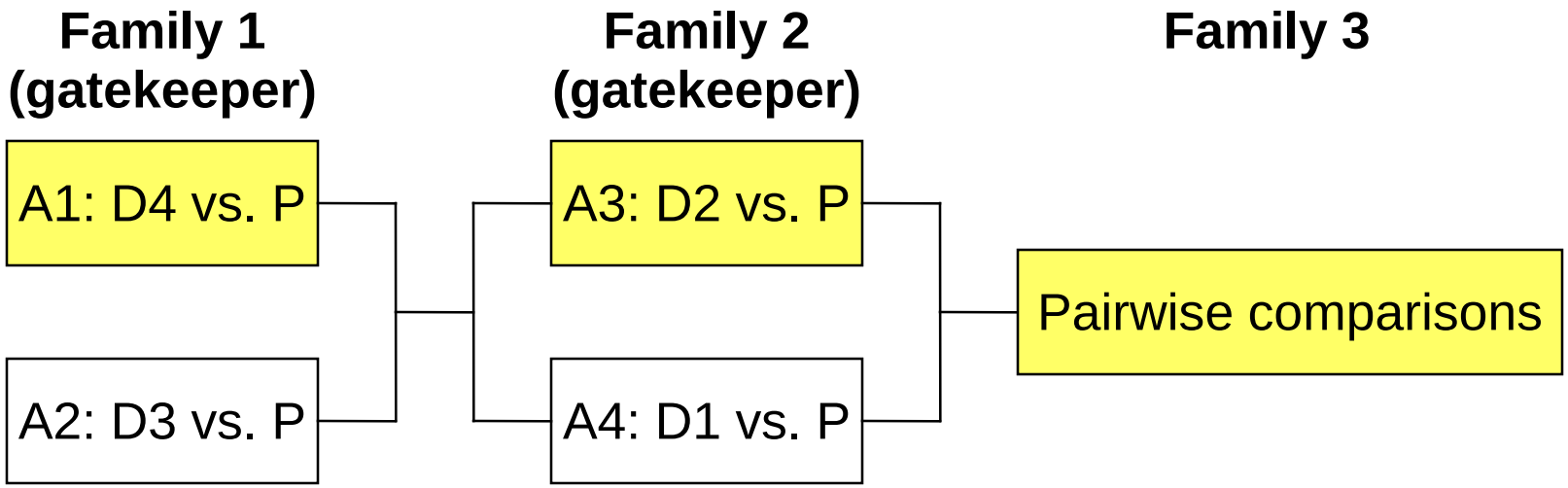


Step 1: Compare doses D3 and D4 to placebo

Step 2: Compare doses D1 and D3 to placebo if at least one comparison at Step 1 is significant

Step 3: Perform various pairwise dose comparisons if at least one comparison at Step 2 is significant

Example 3: Dose-finding study

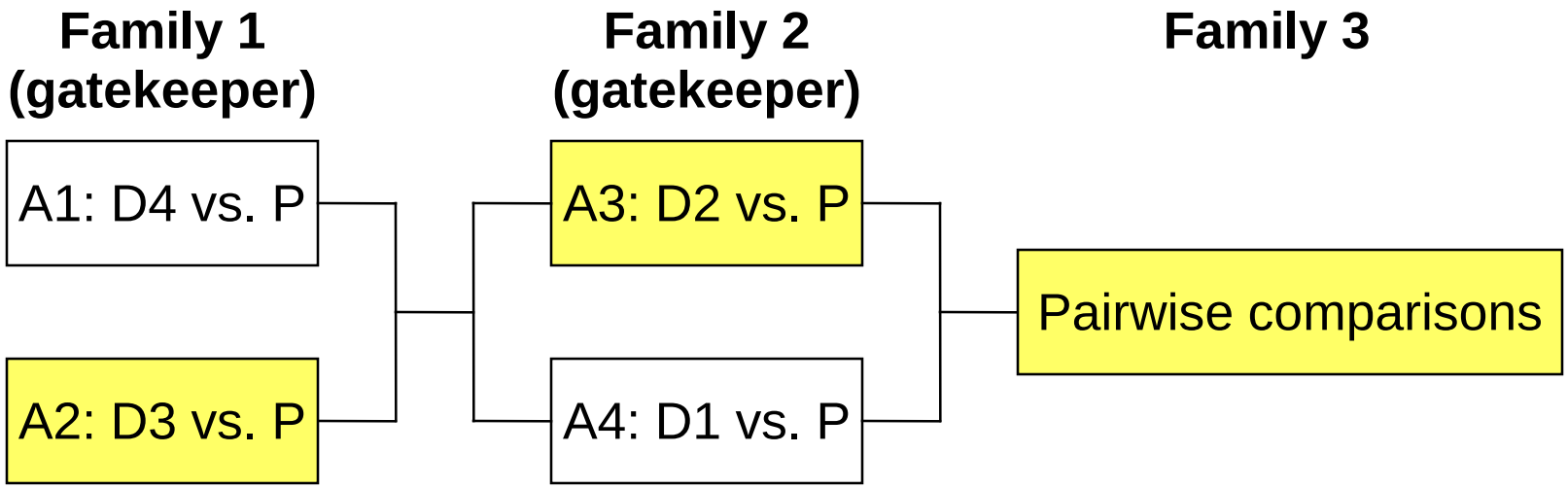


Step 1: Compare doses D3 and D4 to placebo

Step 2: Compare doses D1 and D3 to placebo if at least one comparison at Step 1 is significant

Step 3: Perform various pairwise dose comparisons if at least one comparison at Step 2 is significant

Example 3: Dose-finding study



Step 1: Compare doses D3 and D4 to placebo

Step 2: Compare doses D1 and D3 to placebo if at least one comparison at Step 1 is significant

Step 3: Perform various pairwise dose comparisons if at least one comparison at Step 2 is significant

Parallel gatekeeping strategy

Comparison	Raw p	Adjusted p	
		Gatekeeping procedure	Holm procedure
D4 vs. P	0.0008	0.0016	0.0055
D3 vs. P	0.0135	0.0269	0.0673
D2 vs. P	0.0197	0.0394	0.0787
D1 vs. P	0.7237	1.0000	1.0000
D4 vs. D1	0.0003	0.0394	0.0021
D4 vs. D2	0.2779	1.0000	0.8338
D3 vs. D1	0.0054	0.0394	0.0324
D3 vs. D2	0.8473	1.0000	1.0000

Doses D2, D3 and D4 are significantly different from placebo at 5% level

Summary

Gatekeeping strategies can be successfully used in

- pivotal trials with multiple primary and secondary endpoints
- dose-finding studies

Registration trials

- a priori designation of gatekeeping strategy allows additional data useful to physician and patient to be presented in the product label

Dose-finding studies

- efficient tests of dose-response relationship

Extensions

More powerful gatekeeping tests

- based on more powerful tests, e.g., Simes test
- based on tests accounting for the correlation among the endpoints (exact parametric tests such as Dunnett's test and approximate resampling-based Westfall-Young tests)

Software implementation

- SAS programs for gatekeeping tests can be found in Dmitrienko, Molenberghs, Chuang-Stein, Offen. (2004). *Analysis of Clinical Trials: A Practical Guide*. SAS Publishing, Cary, NC.

References

- 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.
- Marcus R, Peritz E, Gabriel KR. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika*. 63, 655-660.
- Westfall, Krishen. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference*. 99, 25-40.