Multiple Comparisons for Multiple Endpoints and Multiple Doses

Ajit C. Tamhane
Northwestern University

Design and Analysis of Clinical Trials Workshop
Institute of Mathematical Sciences
National University of Singapore
October 25, 2011

Outline I

Outline II

Key References

- Book: *Multiple Testing Problems in Pharmaceutical Statistics*, co-edited by A. Dmitrienko, A. C. Tamhane and F. Bretz, Taylor & Francis (2009)
- Website: http://multxpert.com

1. Introduction

- 1.1 Classical Single Hypothesis Testing
- 1.2 Sources of Multiplicity in Clinical Trials
- 1.3 Regulatory Guidelines

1.1 Classical Single Hypothesis Testing

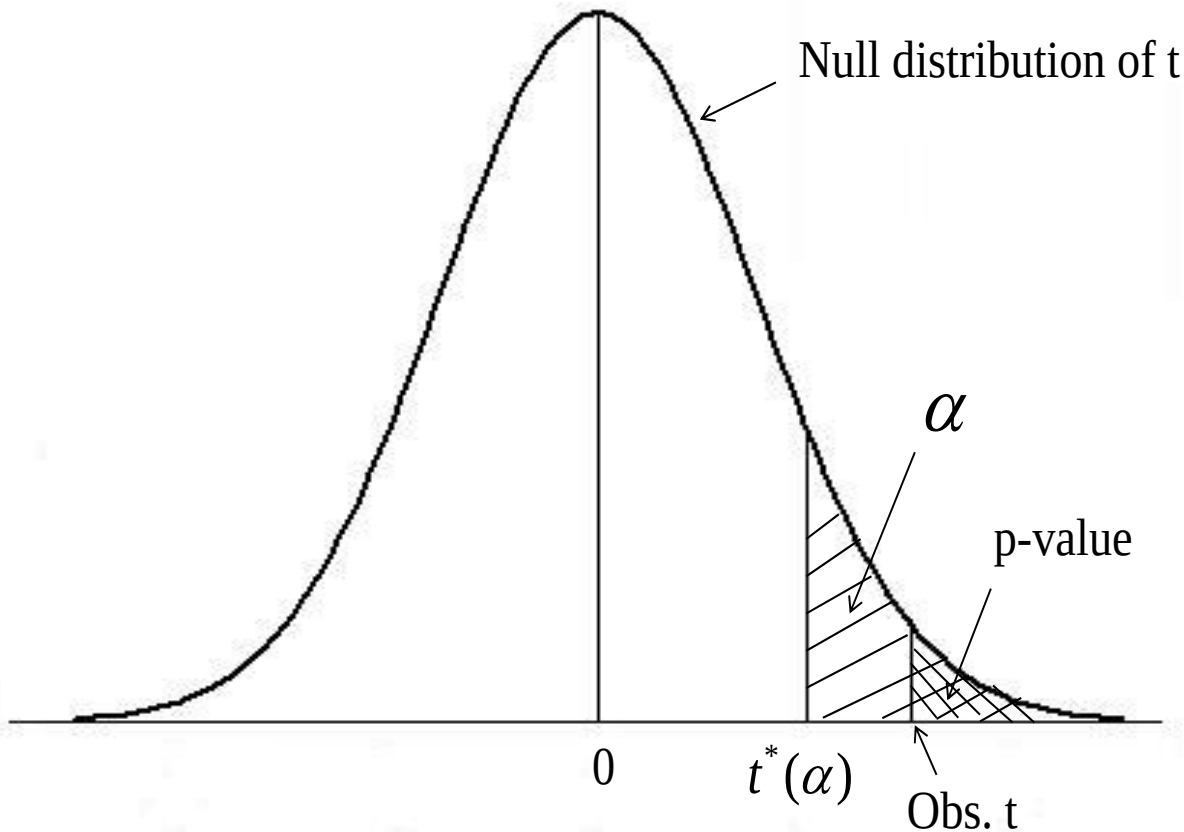
- A single a priori specified null hypothesis $H_0 : \delta = 0$ against an alternative hypothesis $H_1 : \delta > 0$, where $\delta =$ treatment effect.
- Type I error = Reject H_0 when H_0 is true = False positive, Type II error = Accept H_0 when H_1 is true = False negative
- $P(\text{Type I error}) \leq \alpha =$ Significance Level.
- Power $= 1-$ $P(\text{Type II error})$. Maximize power subject to α-requirement.
- Based on the data compute a test statistic

$$t = \frac{\widehat{\delta}}{\text{std. dev.}(\widehat{\delta})}$$

and its p-value.

- Reject H_0 if $t > t^*(\alpha)$ or equivalently if p-value $< \alpha$.

α-Level Test and p-Value



1.2 Sources of Multiplicity in Clinical Trials

- Multiple endpoints (efficacy and safety)
- Multiple treatment arms or doses of a drug
- Interim analyses (group sequential trials)
- Subgroup analyses
- Data-snooping or data-fishing
- Chance of false positives increases if no adjustment for multiplicity is made.
- Multiple test procedures (MTPs) control frequency of false positives.

1.3 Regulatory Guidelines

- FDA Multiplicity Guidance Document (due early 2012).
- ICH E9: "in confirmatory analyses, any aspects of multiplicity . . . should be identified in the protocol; adjustment should always be considered and the details of any adjustment procedure . . . should be set out in the analysis plan."

2. Error Rates

- 2.1 Familywise Error Rate (FWER)
- 2.2 False Discovery Rate (FDR)

2.1 Familywise Error Rate (FWER)

- A family is a collection of a priori stated null hypotheses

$$F = \{H_1, \ldots, H_n\}.$$

- Test statistics $t_1, t_2, \ldots, t_n$.
- p-values: $p_1, p_2, \ldots, p_n$.
- MTPs are commonly designed to control the Type I Familywise Error Rate (FWER):

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

for any combination of true and false H_i: Strong control (Hochberg and Tamhane 1987).

2.2 False Discovery Rate (FDR)

- Benjamini & Hochberg (1995)
- Let $R = \sharp$ of rejections, $V = \sharp$ of false rejections. Then

$$\text{FDR} = E\left(\frac{V}{R}\right).$$

- Used in exploratory studies involving a very large number of hypotheses, e.g., microarrays.
- FDR not applicable in confirmatory trials with a few endpoints. We will use FWER throughout.

3. Examples

All clinical trial examples are from Eli Lilly (courtesy Dr. Alex Dmitrienko), but with modified data

- 3.1 Cardiovascular Trial
- 3.2 Alzheimer's Trial
- 3.3 Extensions and Other Problems

3.1 Cardiovascular Trial

- Trial to evaluate the effects of lisinopril on mortality and morbidity of patients with heart disease (similar to Packer et al. studies (1996. 1999) on amlodipine and lisinopril).
- Two co-primary endpoints:
  - All-cause mortality
  - All-cause mortality or all-cause hospitalization
- Win criterion: Win on *at least one* endpoint (classical multiple comparisons problem)

3.2 Alzheimer's Trial

- Trial to evaluate the effects of donepezil on cognition and global changes in patients with mild to moderate Alzheimer's disease.
- Two co-primary endpoints:
  - Alzheimer's disease assessment scale-Cognitive subscale (ADAS-Cog)
  - Clinician global impression change (CGIC)
- Win criterion: Win on *both* endpoints

3.3 Extensions and Other Problems

- Primary and secondary endpoints with logical restrictions
- Multiple doses of a drug tested on multiple endpoints
- Non-inferiority/superiority tests
- Gatekeeping procedures
- Overall efficacy (global tests, O'Brien 1985)

4. Methods for Constructing MTPs

- 4.1 Union-Intersection (At Least One) Method
- 4.2 Intersection-Union (All or None) Method
- 4.3 Closure Method

4.1 Union-Intersection (At Least One) Method

- c: critical constant to control FWER at α level.

$$\text{Reject at least one } H_i$$
$$\Longleftrightarrow$$
$$\text{At least one } t_i > c$$
$$\Longleftrightarrow$$
$$t_{\max} > c$$

- FWER is controlled if c is the upper α critical point of $t_{\max}$ under H_0: MAX test (Roy 1953):
- Bonferroni procedure: Rejects H_i if $t_i > c = t^*\left(\frac{\alpha}{n}\right)$ or $p_i < \frac{\alpha}{n}$.

4.2 Intersection-Union (All or None) Method

- c: critical constant to control FWER at α level.

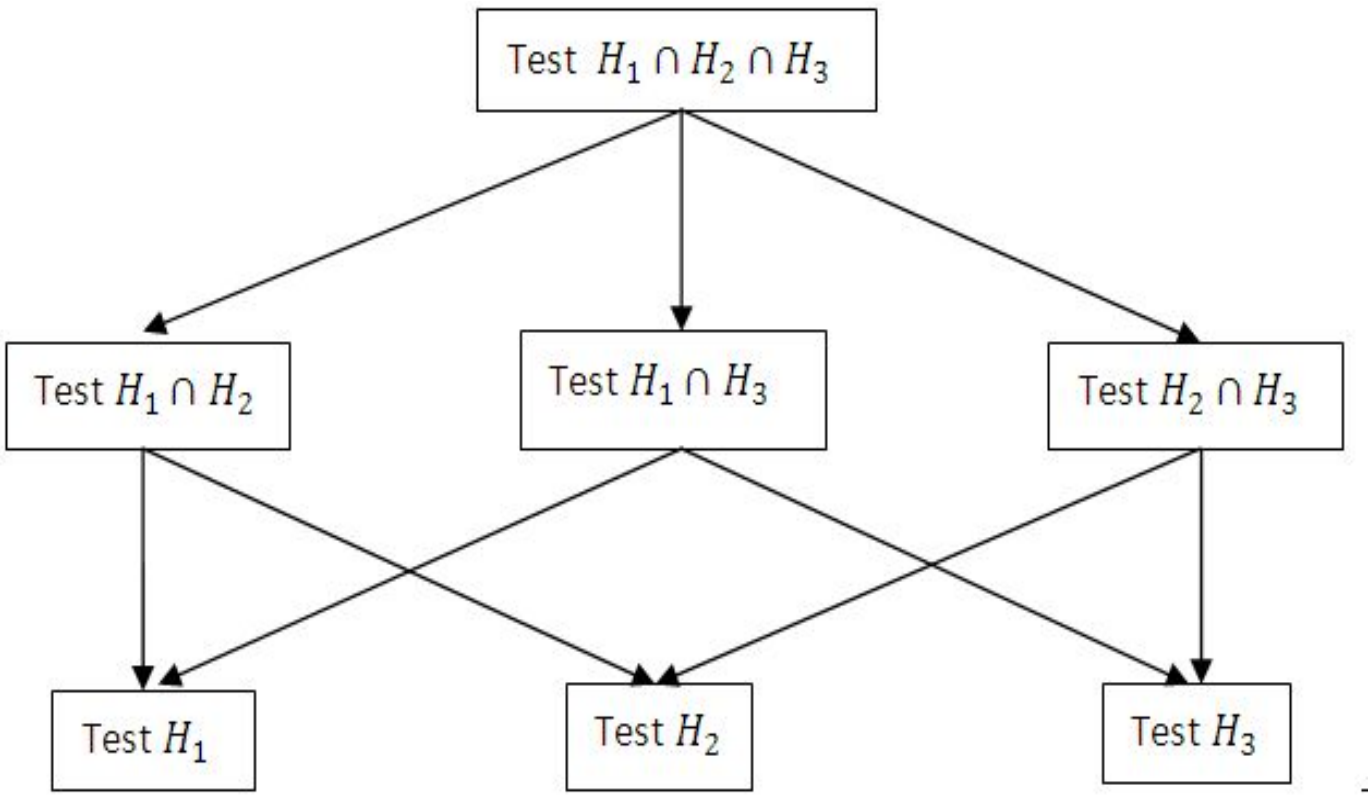
$$\text{Reject all } H_i$$
$$\Longleftrightarrow$$
$$\text{All } t_i > c$$
$$\Longleftrightarrow$$
$$t_{\min} > c$$

- FWER is controlled if c is the upper α critical point of a single t_i, i.e. $c = t^*(\alpha)$.
- Reject all H_i if $t_{\min} > t^*(\alpha)$ or $p_{\max} < \alpha$: MIN test (Berger 1982,Laska and Meisner 1989)
- No α-adjustment.

4.3 Closure Method

- Useful for constructing more powerful stepwise MTPs (Marcus, Peritz & Gabriel 1976).
- Test each intersection hypothesis using any α-level test starting with the full intersection of all elementary hypotheses.
- If any intersection hypothesis is accepted, accept all intersection hypotheses implied by it (ensures coherence).
- Strongly controls FWER $\leq \alpha$.
- In many cases shortcut stepwise procedures exist: Easy to use and transparent.

Closure Method: Example



5. Common p-Value Based MTPs

- 5.1 Holm Procedure
- 5.2 Simes Test
- 5.3 Hochberg Procedure

Reasons for Using p-Value Procedures

- Correlations between endpoints are unknown, so parametric procedures based on multivariate test statistics can't be exactly used.
- Marginal p-values are readily available (but ignore correlations).
- Marginal p-values may come from diverse tests, e.g., t-tests, χ^2-tests, logrank tests, etc.

5.1 Holm Procedure

- Holm (1979): Stepwise shortcut to a closed procedure that uses the Bonferroni test for each intersection hypothesis.
- Step-down algorithm

$$\begin{array}{ccccccc} H_{(1)} & & H_{(2)} & & & & H_{(n)} \\ p_{(1)} & \leq & p_{(2)} & \leq & \cdots & \leq & p_{(n)} \\ \frac{\alpha}{n} & & \frac{\alpha}{n-1} & & & & \frac{\alpha}{1} \end{array}$$

- Begin testing with $p_{(1)}$ & continue as long as you get rejections. If at the ith step $p_{(i)} > \frac{\alpha}{n-i+1}$ then accept $H_{(i)}$ and all the remaining hypotheses.
- Adjusted p-values:
  $\widetilde{p}_{(i)} = \max(\widetilde{p}_{(i-1)}, (n-i+1)p_{(i)}),\ i = 1, \ldots, n.$

5.2 Simes Test

- Simes (1986) Identity: If the p-values are independent then

$$P_{H_0}\left(p_{(1)} > \frac{\alpha}{n}, p_{(2)} > \frac{2\alpha}{n}, \ldots, p_{(n)} > \frac{\alpha}{1},\right) = 1 - \alpha.$$

- Reject H_0 at level α if at least one $p_{(i)} \leq \frac{i\alpha}{n}$.
- More powerful than the Bonferroni test, but requires the independence assumption (or positively correlated p-values, Sarkar and Chang 1997).
- Does not control FWER if used as an MTP, i.e., if we reject $H_{(i)}$ when $p_{(i)} \leq \frac{i\alpha}{n}$.

5.3 Hochberg Procedure

- Hochberg (1988): Conservative stepwise shortcut to a closed procedure that uses the Simes test for each intersection hypothesis. Exact shortcut: Hommel (1988).
- Step-up algorithm

$$\begin{array}{ccccccc} H_{(1)} & & H_{(2)} & & & & H_{(n)} \\ p_{(1)} & \leq & p_{(2)} & \leq & \cdots & \leq & p_{(n)} \\ \frac{\alpha}{n} & & \frac{\alpha}{n-1} & & & & \frac{\alpha}{1} \end{array}$$

- Begin testing with $p_{(n)}$ & continue as long as you get acceptances. If at the ith step $p_{(i)} < \frac{\alpha}{n-i+1}$ then reject $H_{(i)}$ and all the remaining hypotheses.
- Adjusted p-values:
  $\widetilde{p}_{(i)} = \min(\widetilde{p}_{(i+1)}, (n-i+1)p_{(i)}),\ i = 1, \ldots, n.$

6. MTPs for a priori Ordered Hypotheses

- 6.1 Fixed Sequence Procedure
- 6.2 Fallback Procedure

6.1 Fixed Sequence Procedure

- In some problems hypotheses are a priori ordered based on importance, e.g., ordered doses.

$$H_1 \to H_2 \to \cdots \to H_n.$$

- Fixed sequence procedure: Starting with H_1, reject each H_i if $p_i \leq \alpha$. Continue testing as long as rejections occur. Stop testing and accept all the remaining hypotheses if an acceptance occurs.
- No α-adjustment (Maurer, Hothorn & Lehmacher 1995).

6.2 Fallback Procedure

- Proposed by Wiens (2003).
- Assign weights $w_i > 0$ to hypotheses H_i $(i = 1, \ldots, n)$; $\sum_{i=1}^{n} w_i = 1$.
- At any step i test H_i at level α_i where

$$\alpha_i = \begin{cases} w_i \alpha & \text{if } H_{i-1} \text{ is accepted} \\ w_i \alpha + \alpha_{i-1} & \text{if } H_{i-1} \text{ is rejected} \end{cases}$$

- Fixed sequence procedure: Special case of fallback for $w_1 = 1, w_2 = \cdots = w_n = 0$.
- "Use it or lose it" principle.

Summary of Procedures

- In terms of power, Hochberg $>$ Holm $>$ Bonferroni.
- Hochberg requires p-values to be independent or positively correlated; no such restriction on Holm and Bonferroni.
- Fixed sequence and fallback used for a priori ordered hypotheses.
- Fallback is more flexible than fixed sequence.
- Whether fallback or fixed sequence is more powerful depends on the true effect sizes for ordered hypotheses and weights used by fallback.

7. Back to Examples

- 7.1 Cardiovascular Trial
- 7.2 Alzheimer's Disease Trial

7.1 Cardiovascular Trial

Placebo: $n = 1596$, Treatment: $n = 1568$, $\alpha = 0.025$

Endpoint	Event Rate (%)		z-statistic	1-sided p-value
	Placebo	Treatment		
E1	44.8	41.1	2.102	0.018
E2	83.8	80.8	2.211	0.014

7.1 Cardiovascular Trial

- Bonferroni Procedure: Both $p_1 = 0.018$ and $p_2 = 0.014 > \alpha/2 = 0.0125$, so declare both not significant.
- Holm Procedure: $p_{(1)} = p_2 = 0.014 > 0.0125$, so stop testing and declare both not significant.
- Hochberg Procedure: $p_{(2)} = p_1 = 0.018 < \alpha = 0.025$, so stop testing and declare both significant.
- Fixed Sequence Procedure: $p_1 = 0.018 < \alpha = 0.025$ and $p_2 = 0.014 < \alpha = 0.025$, so declare both significant.
- Fallback Procedure: Suppose $w_1 = 0.8, w_2 = 0.2$. $p_1 = 0.018 < 0.8\alpha = 0.020$ and $p_2 = 0.014 < 0.2\alpha + 0.8\alpha = 0.025$, so declare both significant.

7.1 Cardiovascular Trial

Adjusted p-Values

(Raw p-Values: $p_1 = 0.018, p_2 = 0.014$)

Procedure	Endpoint	
	E1	E2
Bonferroni	0.036	0.028
Holm	0.028	0.028
Hochberg	0.018	0.018
Fixed Sequence	0.018	0.018
Fallback	0.023	0.014

7.2 Alzheimer's Disease Trial

Placebo: $n = 161$, Treatment: $n = 167$, $\alpha = 0.025$

Endpoint	Mean and Std. Error		t-statistic	1-sided p-value
	Placebo	Treatment		
E1	2.1 (0.583)	0.4 (0.573)	2.080	0.0192
E2	4.4 (0.087)	4.1 (0.085)	2.469	0.0135

$p_{\max} = 0.0192 < 0.025$. So the treatment is effective.

8. Dose Comparisons with a Placebo

- 8.1 Dunnett Procedure
- 8.2 Major Depressive Disorder Trial Example

8.1 Dunnett Procedure

- Compare $m \geq 2$ increasing doses with a zero dose (control).
- Assume that data from Dose i is distributed $N(\mu_i, \sigma^2)$.
- Hypotheses $H_i : \mu_i - \mu_0 = 0$ vs. $\overline{H}_i : \mu_i - \mu_0 > 0$.
- Assume common sample size n per dose.
- Test statistics:

$$t_i = \frac{\overline{y}_i - \overline{y}_0}{\widehat{\sigma}\sqrt{2/n}} \ (i = 1, \ldots, m).$$

- p-values: $p_i \ (i = 1, \ldots, m)$.

8.1 Dunnett Procedure

- Dunnett Procedure: A parametric alternative to the Bonferroni procedure. Exploits known correlations (depend on the sample sizes) between the test statistics.
- Based on the union-intersection method: Reject H_i if $t_i > c$ where $c =$ upper α critical point of $t_{\max}$.
- To evaluate c, need the joint distribution of $t_1, \ldots, t_n$: Multivariate t-distribution.
- Table of Multivariate t and Bonferroni c-values ($\alpha = 0.05$, degrees of freedom $= \infty$)

m	1	2	3	4	5
Mult. c	1.645	1.916	2.062	2.160	2.234
Bonf. c	1.645	1.960	2.127	2.241	2.326

8.2 Major Depressive Disorder Trial

- Placebo-controlled parallel arm trial
- Four dose levels (D1: 10 mg/day, D2: 20 mg/day, D3: 40 mg/day, D4: 60 mg/day) + D0: Placebo
- 432 patients randomized to 5 dose groups
- Endpoint: Mean reduction from baseline in 17-item Hamilton Depression Scale (HAMD-17)
- Data

Dose	D0	D1	D2	D3	D4
n	85	87	88	87	85
$\overline{y}$	6.12	6.85	7.08	7.50	7.22

$$\widehat{\sigma} = 6.6$$

8.2 Major Depressive Disorder Trial

t-Statistics and p-Values, $\alpha = .05$

	D1 vs. D0	D2 vs. D0	D3 vs. D0	D4 vs. D0
t_i	1.464	1.925	2.767	2.206
p_i	0.072	0.027	0.003	0.017

- Dunnett procedure: Compares t-statistics with $c = 2.160$ (for $m = 4$). Declares doses D3 and D4 significantly better than D0.
- Step-down Dunnett procedure: $t_3 > 2.160, t_4 > 2.062, t_2 > 1.916, t_1 < 1.645$. Stops testing and declares doses D2, D3 and D4 significantly better than D0.

8.2 Major Depressive Disorder Trial

t-Statistics and p-Values, $\alpha = .05$

	D1 vs. D0	D2 vs. D0	D3 vs. D0	D4 vs. D0
t_i	1.464	1.925	2.767	2.206
p_i	0.072	0.027	0.003	0.017

- Bonferroni procedure: Compares t-statistics with $c = 2.241$ (for $m = 4$) or p-values with $.05/4 = .0125$. Declares dose D3 significantly better than D0.
- Holm procedure: $p_3 < .0125, p_4 < .0167, p_2 > .025$. Stops testing and declares doses D3 and D4 significantly better than D0 (same result with Hochberg procedure).

9. Gatekeeping Procedures

- 9.1 Why Gatekeeping?
- 9.2 Assumptions and Notation
- 9.3 Types of Gatekeeping
- 9.4 Serial Gatekeeping
- 9.5 Parallel Gatekeeping
- 9.6 General Gatekeeping

9.1 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 (due early 2012) is expected to have a similar requirement.

9.2 Assumptions and 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:

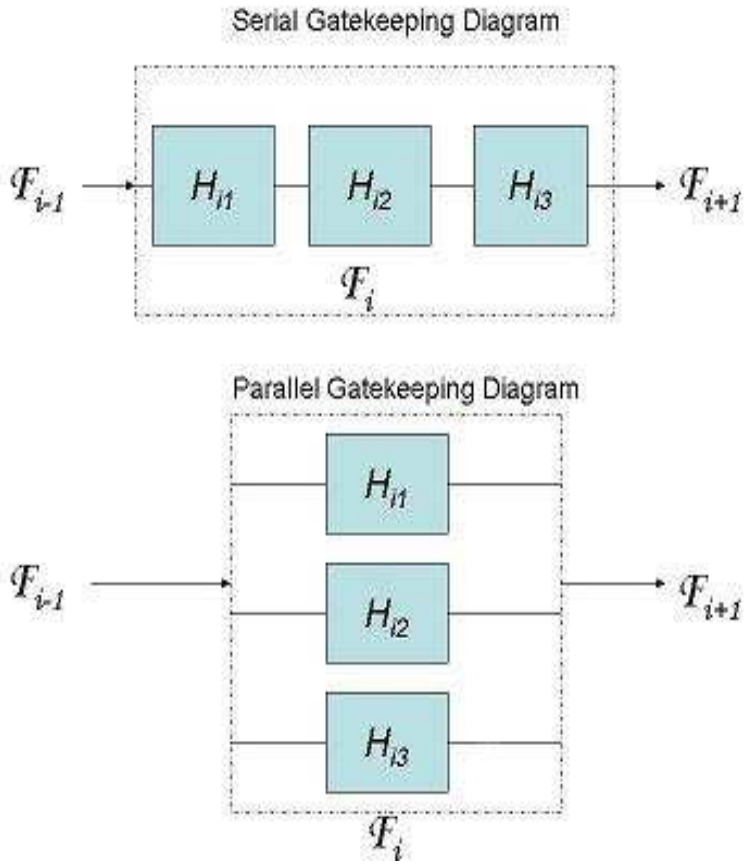
$$\text{FWER} = P\{\text{Reject at least one true } H_i\} \leq \alpha.$$

- 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).

9.3 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: Gatekeeper F_j is passed iff all $H_i \in F_j$ are rejected (Maurer, Hothorn & Lehmacher 1995).
- Parallel gatekeeping: Gatekeeper F_j is passed iff at least one $H_i \in F_j$ is rejected (Dmitrienko, Offen & Westfall 2003).
- General gatekeeping (Dmitrienko, Wiens, Tamhane & Wang 2007, Dmitrienko and Tamhane 2011a,b, Dmitrienko, Kordzkhia and Tamhane 2011).

Serial and Parallel Gatekeeping



Serial and Parallel Gatekeeping: 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

9.4 Serial Gatekeeping

- Maurer, Hothorn & Lehmacher (1995).
- Test each F_j at local level α using any procedure. Proceed to test F_{j+1} iff all $H_i \in F_j$ are rejected.
- Use the Intersection-Union (IU) procedure which tests each $H_i \in F_j$ at level α.
- Generalization of the fixed sequence test.
- This procedure can be derived using the closure method, hence controls FWER strongly.

9.5 Parallel Gatekeeping

- Dmitrienko, Westfall & Offen (2003), Dmitrienko, Tamhane, Wang & Chen (2006), Guilbaud (2007), Dmitrienko, Tamhane & Wiens (2008).
- Stepwise procedure based on the "use it or lose it" principle (underlies the fixed sequence and fallback tests).
  - If a hypothesis is rejected then the α allocated to it is not spent and can be reused to test other hypotheses.
  - If a hypothesis is accepted then the α allocated to it is spent and cannot be reused to test other hypotheses.
- The error rate function quantifies the "unused" α that can be carried forward from one family to the next.

Error Rate Function: Definition

- Consider a single family $F = \{H_1, H_2, \ldots, H_n\}$ and let $N = \{1, 2, \ldots, n\}$.
- Let $H(I) = \bigcap_{i \in I} H_i$ for $I \subseteq N$ be an intersection hypothesis.
- For fixed α, the error rate function of a procedure is

$$e(I|\alpha) = \sup_{H(I)} P\left\{\text{Reject at least one } H_i,\ i \in I \middle|\ H(I)\right\}.$$

- For the Bonferroni procedure: $e(I|\alpha) = \frac{|I|}{n}\alpha$.
- $e(\emptyset|\alpha) = 0$, $e(I|\alpha) \leq e(J|\alpha)$ if $I \subseteq J$, $e(N|\alpha) = \alpha$.

Separable Procedures

- A procedure is separable if its error rate function satisfies

$$e(I|\alpha) < \alpha \text{ if } I \subset N \text{ and } e(N|\alpha) = \alpha.$$

- Single-step procedures, e.g., Bonferroni and Dunnett, are separable; stepwise procedures, e.g., Holm, Hochberg, Hommel and fallback, are non-separable.
- These stepwise procedures can be made separable by using a convex combination of the critical constants of stepwise and single-step procedures (e.g., Bonferroni or Dunnett).
- We call such hybrid stepwise procedures *truncated procedures*.
- In terms of power, non-separable procedures $>$ truncated procedures $>$ separable procedures.

Truncated Holm Procedure

- Compare ordered p-values, $p_{(j)}$, with the critical constants

$$\left[\frac{\gamma}{n-j+1}+\frac{1-\gamma}{n}\right]\alpha \text{ for } j=1,2,\ldots,n.$$

- $\gamma=0$ gives Bonferroni and $\gamma=1$ gives Holm.
- The error rate function:

$$e(I|\alpha)=\left[\gamma+(1-\gamma)\frac{|I|}{n}\right]\alpha$$

  if $|I|>0$ and $e(I|\alpha)=0$ if $|I|=0$.
- Truncated Holm is less powerful than Holm but more powerful than Bonferroni.

Parallel Gatekeeping: Stepwise Procedure

- Specify procedures $\mathcal{P}_1, \ldots, \mathcal{P}_m$ for families $F_1, \ldots, F_m$. The first $m-1$ procedures must be separable. $\mathcal{P}_m$ can be non-separable.
- Step 0: Set $\alpha_1 = \alpha$.
- Step j: For $j = 1, \ldots, m-1$, test F_j using $\mathcal{P}_j$ at level α_j. Let A_j be the index set of accepted hypotheses. Set

$$\alpha_{j+1} = \alpha_j - e_j(A_j|\alpha_j).$$

- Step m: Test F_m using $\mathcal{P}_m$ at level α_m.
- Note $\alpha_{j+1} = 0$ if $A_j = N_j$: parallel gatekeeping condition.
- Can test F_{j+1} when $A_j \subset N_j$ iff $\alpha_{j+1} > 0$ iff $\alpha_j > e_j(A_j|\alpha_j)$: separability condition.

Parallel Gatekeeping Stepwise Procedure: Example 1

- Suppose $\mathcal{P}_1, \ldots, \mathcal{P}_{m-1}$ are Bonferroni and $\mathcal{P}_m$ is any FWER controlling procedure.
- Let $a_j = |A_j| = \sharp$ accepted hypotheses, $r_j = |R_j| = \sharp$ rejected hypotheses. Then

$$\begin{aligned} \alpha_{j+1} &= \alpha_j - e_j(A_j|\alpha_j) \\ &= \alpha_j - \left(\frac{a_j}{n_j}\right)\alpha_j \\ &= \left(\frac{r_j}{n_j}\right)\alpha_j \\ &= \prod_{i=1}^{j}\left(\frac{r_i}{n_i}\right)\alpha. \end{aligned}$$

- More rejections $\Rightarrow$ More α carried forward, $r_j = 0 \Rightarrow$ procedure stops.

Parallel Gatekeeping Stepwise Procedure: Example 2

- $\alpha = .05$
- $\mathcal{P}_1$ for F_1 (Primary Endpoints): Bonferroni
- $\mathcal{P}_2$ for F_2 (Secondary Endpoints): Hochberg

F_i	Endpoint	p	α_i	Crit. Value	Result
F_1	P_1	.031	.05	.025	NS
	P_2	.013	.05	.025	S
F_2	S_1	.039	.025	.025	NS
	S_2	.027	.025	.0125	NS

$$\alpha_2 = \alpha_1 - (a_1/n_1)\alpha_1 = \alpha - (1/2)\alpha = .025.$$

Parallel Gatekeeping Stepwise Procedure: Example 2

$\mathcal{P}_1$ for F_1 (Primary Endpoints): Truncated Holm ($\gamma = 0.5$)
$\mathcal{P}_2$ for F_2 (Secondary Endpoints): Hochberg

F_i	Endpoint	p	α_i	Crit. Value	Result
F_1	P_1	.031	.05	.0375	S
	P_2	.013	.05	.025	S
F_2	S_1	.039	.05	.05	S
	S_2	.027	.05	.025	S

1. Truncated Holm constant for comparing $p = .031$:

$$\left[\frac{0.5}{1} + \frac{1 - 0.5}{2}\right](.05) = .0375.$$

2. $\alpha_2 = \alpha = .05$.

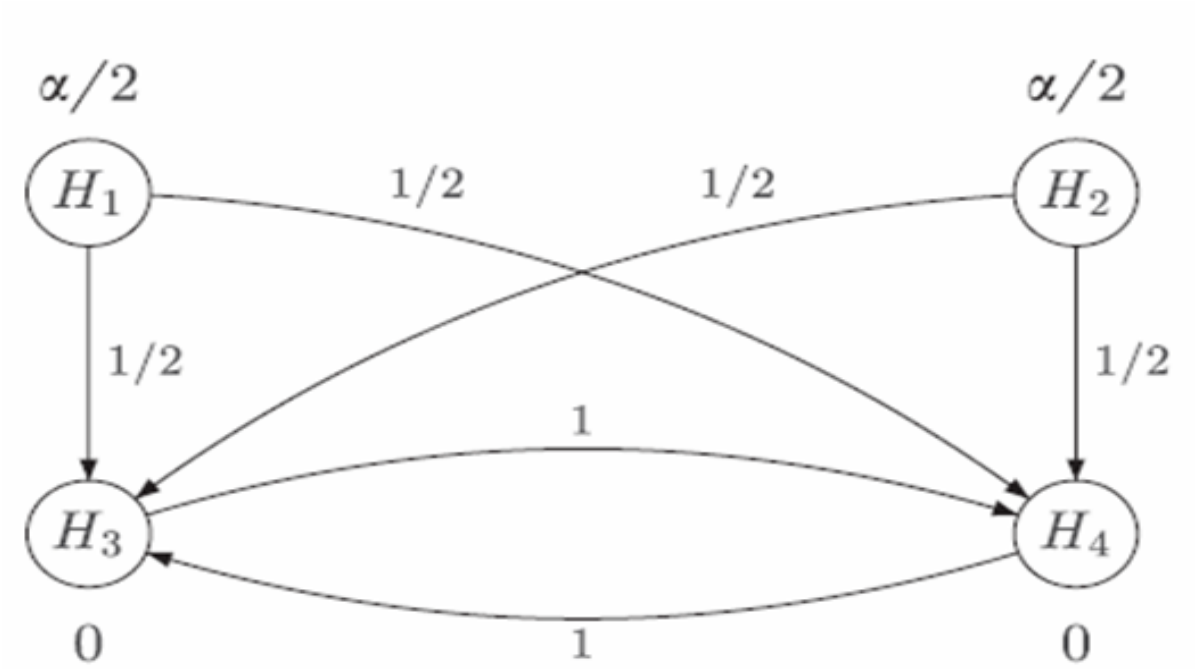
Choice of Truncation Parameter γ

- γ must be prespecified —not selected in light of the data.
- Higher the γ, higher the power of the procedure for the primary family.
- If higher power for the primary family does not result in rejection of more hypotheses then a smaller α carried to the next family $\Rightarrow$ less power for secondary family.
- If higher power for the primary family results in rejection of more hypotheses then a larger α could be carried to the next family $\Rightarrow$ higher power for secondary family.
- Generally, higher power in the primary family is at the expense of lower power for the secondary family.

Graphical Procedures

- Bretz, Maurer, Brannath and Posch (2009)
- $\mathcal{P}_1$: Bonferroni, $\mathcal{P}_2$: Holm

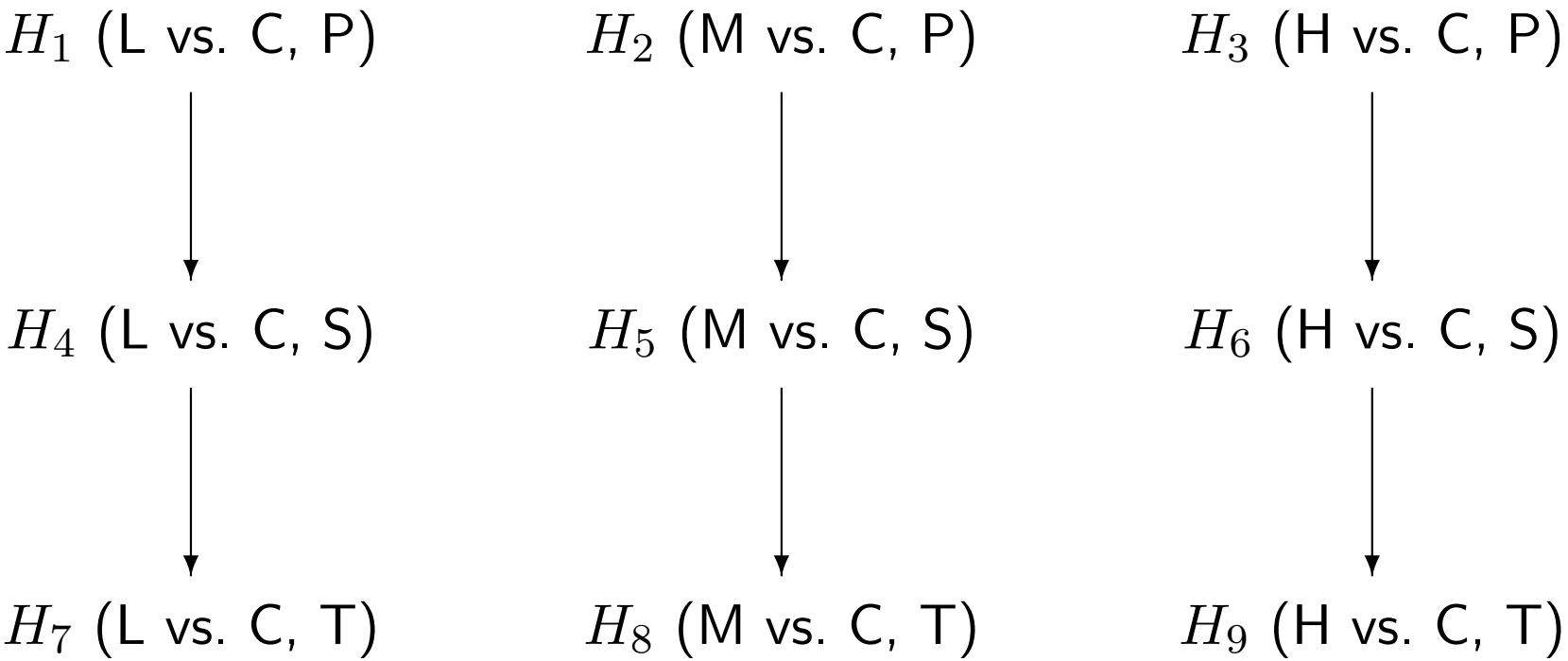
Parallel Gatekeeping Procedure for 2 Families



9.6 General Gatekeeping

- More complex clinical decision rules involving multiple objectives do not fit in the simple serial/parallel framework.
- Example 1: 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.

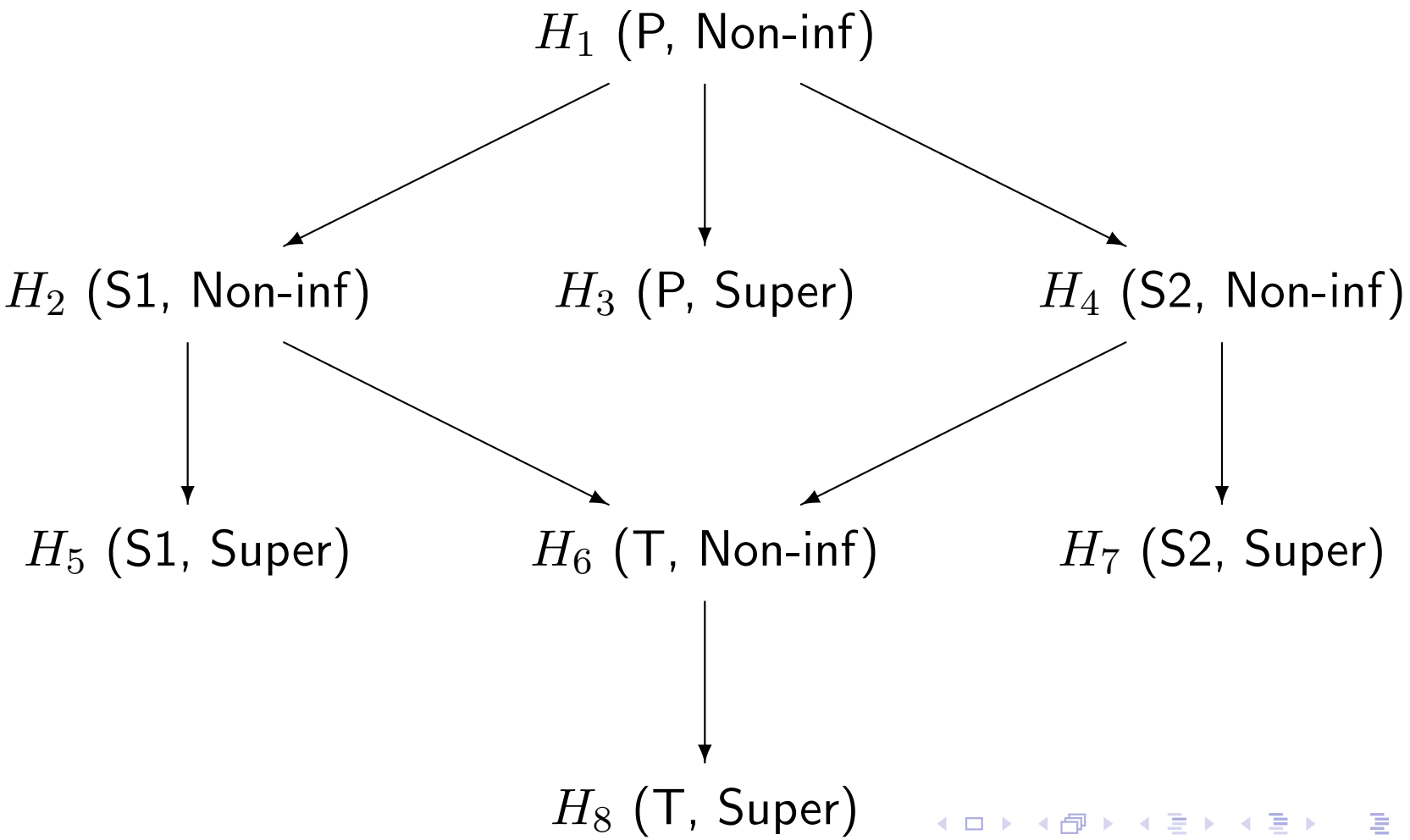
Diabetes Trial Example: Tree Diagram



Example 2: Hypertension Trial

- 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.

Hypertension Example: Tree Diagram



References on General Gatekeeping

- 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)

10. Take-Home Lessons

- 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.

References on Multiple Testing

- Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Royal Statist. Soc., Ser. B*, **57**, 289–300.
- Berger, R. L. (1982), "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, **24**, 295–300.
- Dmitrienko, A., Tamhane, A.C. and Bretz, F. (2009), *Multiple Testing Problems in Pharmaceutical Statistics*, Boca Raton, FL: Taylor & Francis/CRC Press.
- Gabriel, K.R. (1969), "Simultaneous test procedures—Some theory of multiple comparisons," *Ann. Math. Statist.*, **40**, 224–250.

References on Multiple Testing

- Hochberg, Y. (1988), "A sharper Bonferroni procedure for multiple significance testing," *Biometrika*, **75**, 800–802.
- Hochberg, Y. and Tamhane, A.C. (1987), *Multiple Comparison Procedures*, John Wiley and Sons: New York.
- Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian J. Statist.*, **6**, 65-70.
- Hommel, G. (1988), "A stagewise rejective multiple test procedure based on a modified Bonferroni test," *Biometrika*, **75**, 383–386.
- Laska, E. M. and Meisner, M. J. (1989), "Testing whether an identified treatment is best," *Biometrics*, **45**, 1139–1151.

References on Multiple Testing

- Marcus, R. Peritz, E. and Gabriel, K.R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, **63**, 655–660.
- Packer, M. et al. (1996), "Effect of amlodipine on morbidity and mortality in severe chronic heart failure. Prospective randomized amlodipine survival evaluation study group, *N. England J. Med.*, **335**, 1107-1114.
- Packer, M. et al. (1999), "Comparative effects of low and high doses of the angiotensin-converting enzyme inhibitor, lisinopril, on morbidity and mortality in chronic heart failure,"ATLAS Study Group, *Circulation*, **100**, 2312-2318.
- Roy, S. N. (1953), "On a heuristic method of test construction and its use in multivariate analysis," *Ann. Math. Statist.*, **24**, 220–238.

References on Multiple Testing

- Sarkar, S. and Chang, C.K. (1997), "Simes' method for multiple hypothesis testing with positively dependent test statistics," *J. Amer. Statist. Assoc.*, **92**, 1601–1608.
- Simes, R.J. (1986), "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, **63**, 655–660.
- Wiens, B. (2003), "A fixed-sequence Bonferroni procedure for testing multiple endpoints," *Pharma. Statist.*, **2**, 211–215.

References on Gatekeeping Procedures

- Bretz, F., Maurer, W., Brannath W. and Posch, M. (2009), "A graphical approach to sequentially rejective multiple test procedures," *Statistics in Medicine*, **28**, 586–604.
- Dmitrienko, A. and Kordzakhia, G. (2011), "Superchain procedures in clinical trials with multiple objectives,"submitted for publication.
- Dmitrienko, A., Kordzakhia, G., and Tamhane, A.C. (2011), "Multistage and mixture parallel gatekeeping procedures for clinical trials," *Journal of Pharmaceutical Statistics*, **21**, 726-747.
- 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**, 2387–2400.

References on Gatekeeping Procedures

- Dmitrienko, A. and Tamhane, A.C. (2007), "Gatekeeping procedures with clinical trial applications," *Journal of Pharmaceutical Statistics*, **6**, 171–180.
- Dmitrienko, A. and Tamhane, A.C. (2009), "Gatekeeping procedures in clinical trials," in *Multiple Testing Problems in Pharmaceutical Statistics* (eds. A. Dmitrienko, A.C.Tamhane and F. Bretz), Chapman and Hall/CRC: Boca Raton, pp.165–192.
- Dmitrienko, A. and Tamhane, A.C. (2011a), "Mixtures of multiple testing procedures for gatekeeping applications in clinical trial applications," *Statistics in Medicine*, **30**, 1473-1488.
- Dmitrienko, A. and Tamhane, A.C. (2011b), "General theory of mixture procedures for gatekeeping," submitted for publication.

References on Gatekeeping Procedures

- Dmitrienko, A., Tamhane, A.C., Wang, X., and Chen, X. (2006), "Stepwise gatekeeping procedures in clinical trial applications," *Biometrical Journal*, **48**, 984-991.
- Dmitrienko, A., Tamhane, A.C. and Wiens, B. (2008), "General multistage gatekeeping procedures," *Biometrical Journal*, **50**, 667–677.
- Dmitrienko, A. and Wiens, B. (2005), "The fallback procedure for evaluating a single family of hypotheses," *Journal of Biopharmaceutical Statistics*, **15**, 929–942.
- Dmitrienko, A., Wiens, B.L. Tamhane, A.C. and Wang, X. (2007), "Tree-structured-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives," *Statistics in Medicine*, **26**, 2465–2478.

References on Gatekeeping Procedures

- Guilbaud, O. (2007), "Bonferroni parallel gatekeeping - Transparent generalizations, adjusted p-values, and short direct proofs," *Biometrical Journal* **49**, 917–927.
- O'Neill, R.T. (1997), "Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance," *Controlled Clinical Trials*, **18**, 550-556.