



Multiplicity in a Medical Device Trial: Considerations to Control Type I Error

Kevin Najarian, MS
najariak@bsci.com

April 16, 2008
2:00PM

Views expressed in this presentation are those of the speaker and are not associated with Boston Scientific Corporation or AdvaMed

Overview

- Multiplicity
- Data
- Common Testing Options
- Correlation
- Composite Endpoints
- False Discovery Rate

Multiplicity

- **Multiplicity**: A view based on the theory that if you test long enough, you will inevitably find something statistically significant – false-positives due to random variability, even if no real effects exist.
- Some Sources of Multiplicity in Clinical Trials
 - **Multiple endpoints**
 - Multiple studies and/or multiple active arms
 - Multiple analyses and/or tests
 - Interim analysis
 - Preliminary tests
 - Subgroup analysis
 - Selection of covariates in an analysis model

Multiplicity

K independent test statistics with significance level α

\Rightarrow Prob(at least 1 test is falsely rejected): $1-(1-\alpha)^K = \alpha_K$

K	α	$1-(1-\alpha)^K$
1	0.05	0.0500
2	0.05	0.0975
3	0.05	0.1426
5	0.05	0.2262
10	0.05	0.4013

\Rightarrow Control α_K with a multiple test procedure

Dilemma

	<i>H₀ True</i>	<i>H₀ False</i>
<i>Reject H₀</i>	Type I Error, α Multiplicity inflates α	Correct, $1-\beta$ (Power)
<i>Accept H_a</i>	Correct, $1-\alpha$	Type II Error, β Decreasing α will inflate β

Controlling both α and β leads to increased sample size

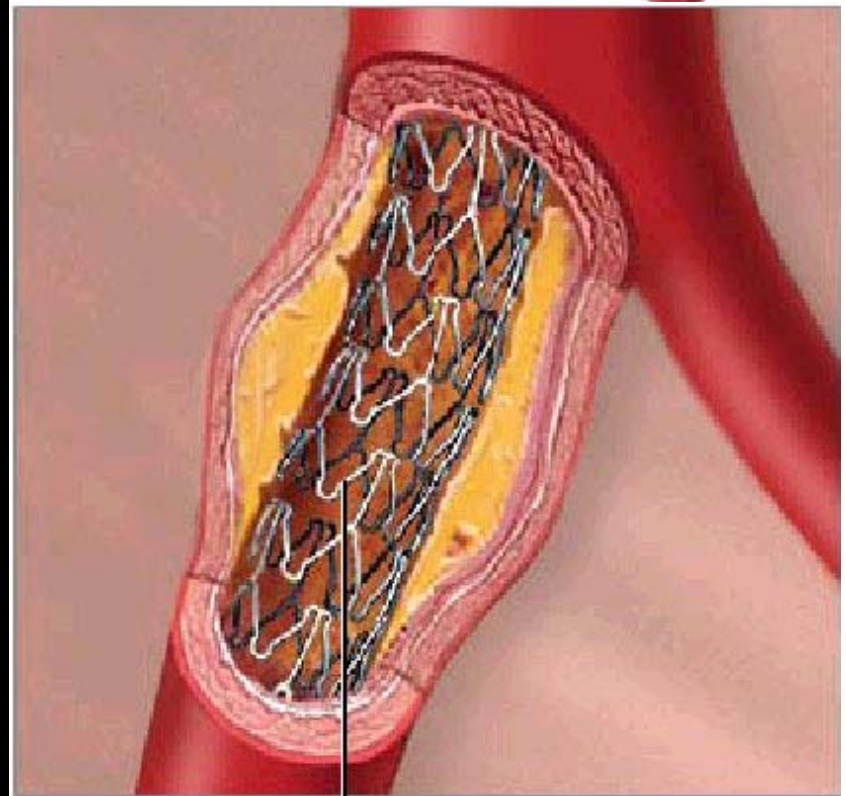
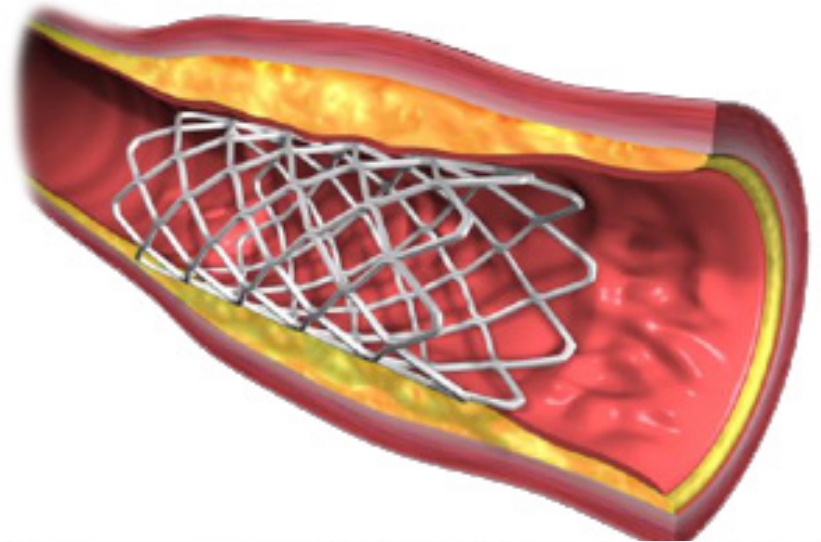
Data

Test Product:

Coronary Artery Stent

Analysis Design:

Comparison of two
stent products
(non-inferiority)



Coronary artery stent

Data

- **Primary Endpoint:**
 - Target Vessel Revascularization (TVR) – **binary**
- **Secondary Endpoints:**
 - Percent Diameter Stenosis - **continuous**
 - Binary Restenosis - **binary**
 - Minimum Lumen Diameter (MLD) - **continuous**
 - Late Loss - **continuous**
 - Percent Net Volume Obstruction - **continuous**

Data

Planning of Analyses:

- Data exploration
- Restrictions
- Consideration of endpoints
- Regulatory concerns

Data

Test of Non-Inferiority:

	Difference [Upper 1-Sided 95% CI]	p-value	Non-Inferiority Delta
Primary Endpoint			
TVR (%)	0.94 [2.98]	0.0487	3.0
Secondary Endpoints			
% Dia. Stenosis*	2.24 [4.44]	0.0006	6.6
Bin. Restenosis (%)	2.74 [5.98]	0.0354	6.3
MLD (mm)	-0.09 [-0.16]	0.0316	-0.17
Late Loss (mm)	-0.01 [0.04]	0.0001	0.18
% Net Vol. Obs.*	1.66 [3.88]	0.0021	5.67

* mean percentage

Global Alpha

- Useful for non-specific global claims; the results can be difficult to interpret; and the Type I error rate can remain inflated (weaker control of FWER).
- The focus of the p-value is on the most significant endpoint

Single Step

Bonferroni Procedure:

- For K endpoints, one accepts as statistically significant all those p-values $\leq \alpha/K$
- Too conservative when the endpoints are highly correlated
- Controls FWER

Single Step

Bonferroni Procedure: How was power compromised?

5 Secondary Endpoints

One-Sided Level of Significance = 0.05

Power = 95%

Bonferroni Level of Significance = $0.05/5 = 0.01$

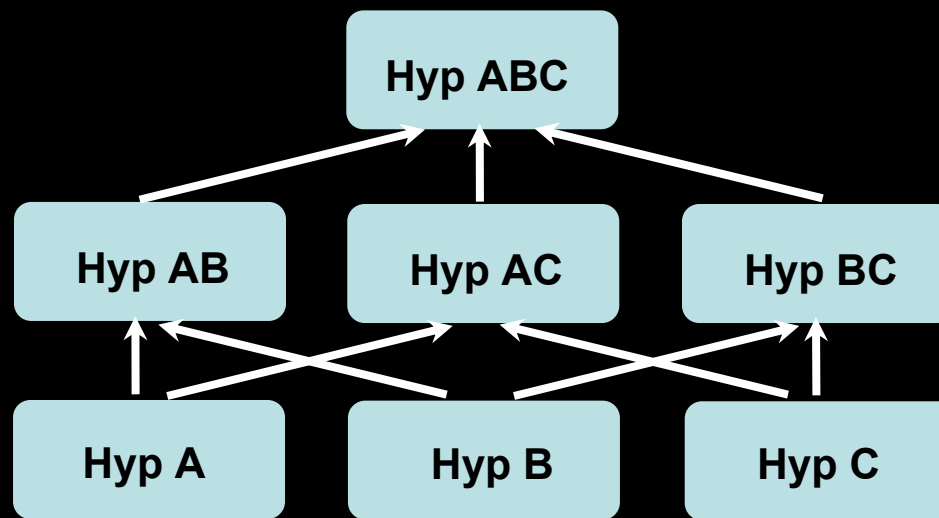
Secondary Endpoint	Planned Non-Inferiority Margin	Analysis Margin	Analysis P-value	Bonferroni Adjusted Power
% Dia. Stenosis	6.6	4.44	0.0006	84%
Bin. Restenosis	6.3%	5.98%	0.0354	83%
MLD	-0.17	-0.16	0.0316	85%
Late Loss	0.18	0.04	0.0001	83%
% Net Vol. Obs.	5.7	3.88	0.0021	83%

Single Step

Šidák Procedure:

- For K endpoints, one accepts as statistically significant all those p-values $\leq 1-(1-\alpha)^{1/K}$
- The adjusted p-values are $1-(1-p_k)^K$, $k=1,2, \dots, K$
- Does not always preserve FWER

Closed Testing



- Test each hypothesis in the closed family using a suitable α .
- A hypothesis is rejected if its associated test and all tests associated with hypotheses implying it are significant.

Closed Testing

Holm Procedure (step-down):

- Let $p_1 \leq p_2 \leq \dots \leq p_K$
- Reject the null if $p_k \leq \alpha / (K - k + 1)$, $k = 1, 2, \dots, K$
- In general, significance testing continues in decreasing order (most significant downward) until a null is not rejected and the value is retained for all remaining null hypotheses.

Closed Testing

Hochberg Procedure (step-up):

- Let $p_1 \geq p_2 \geq \dots \geq p_K$
- Reject the null if $p_k \leq \alpha/k$
- In general, significance testing continues in increasing order (least significant upward) until one rejects a null hypothesis, then one rejects all remaining null hypotheses without further testing.

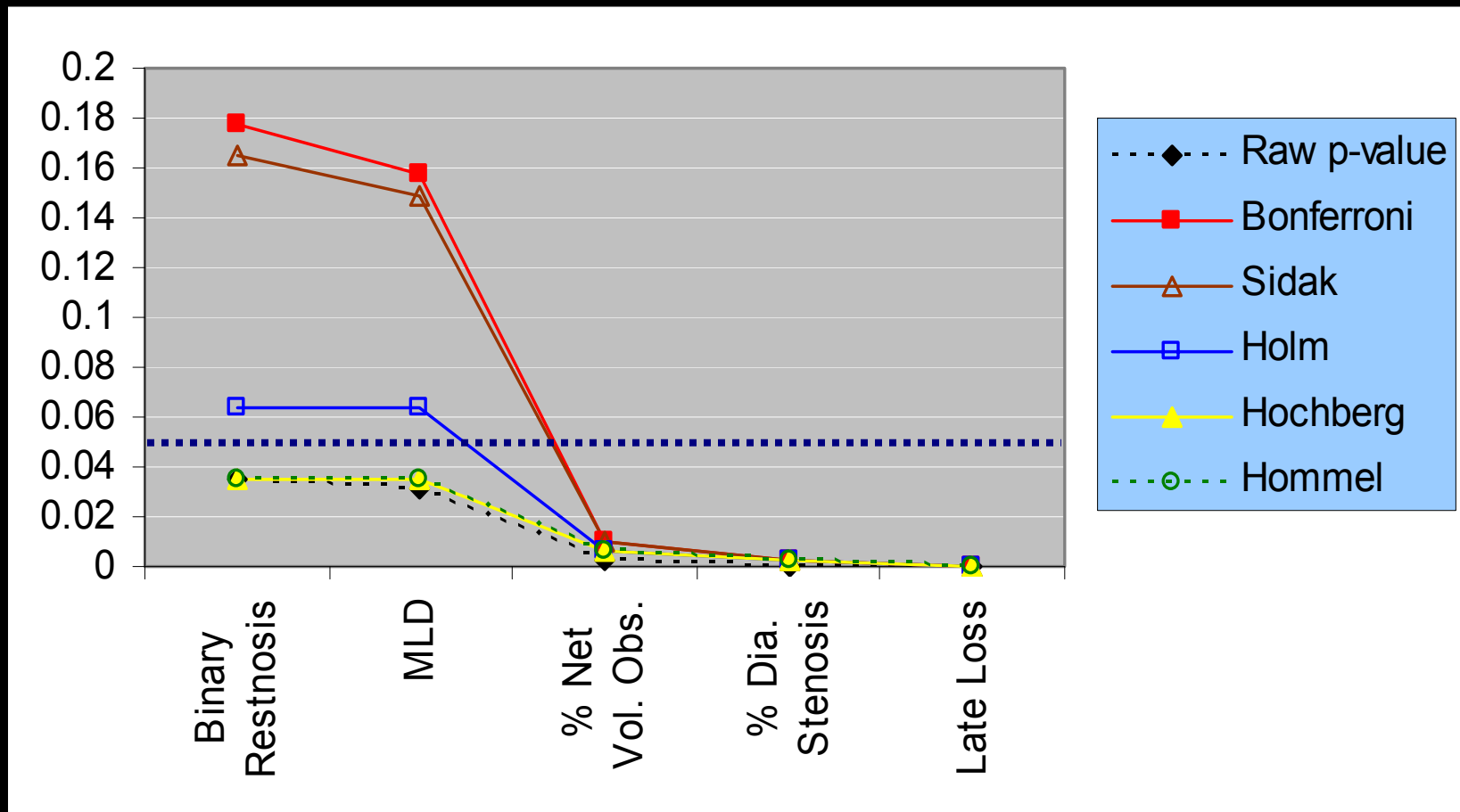
Closed Testing

Hommel Procedure (decision matrix):

- Starts with a global hypothesis and steps down to examine individual hypotheses.
- Rejects all null hypotheses whenever all raw p-values are significant.

Single Step and Closed Testing

Adjusted P-values vs. Endpoints



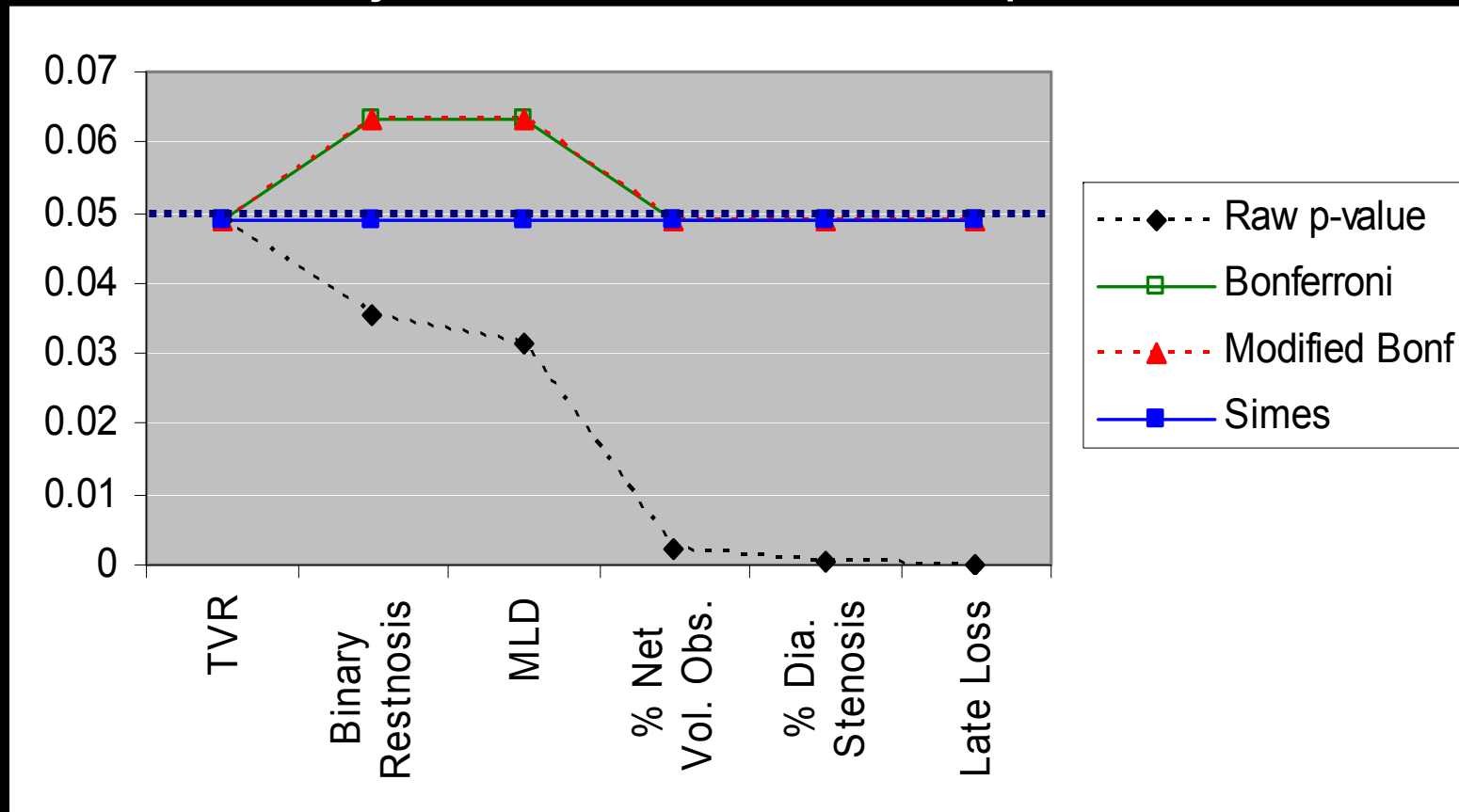
Note: Šidák, Hochberg and Hommel do not always control FWER

Gatekeeping Strategies

- **Testing families of hypotheses with hierarchical ordered endpoints.**
- **Acceptance or rejection of hypotheses in a particular family depend on the outcome of the tests from preceding families.**
- **Earlier families serve as gatekeepers.**
 - **Serial: all hypotheses in a family must be rejected**
 - **Parallel: one hypothesis within a family must be rejected**

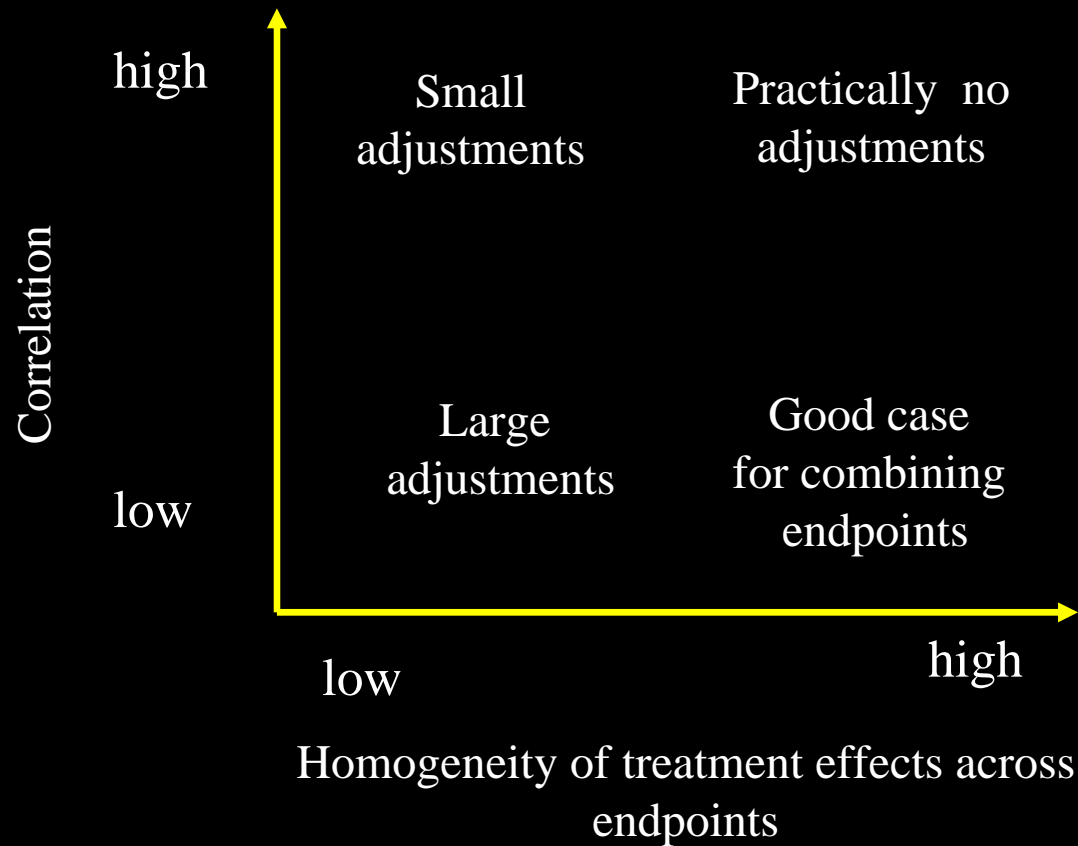
Gatekeeping Strategies

Bonferroni, Modified Bonferroni, and Simes' Adjusted P-values vs. Endpoints



Correlated Endpoints

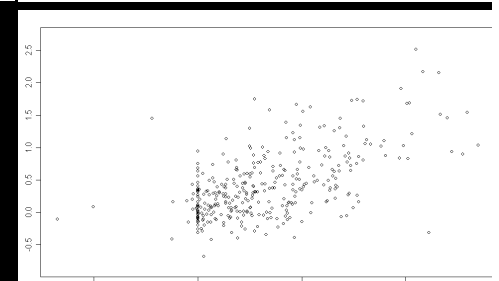
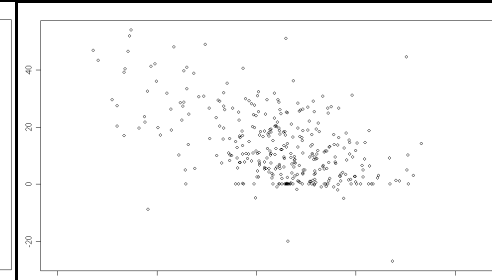
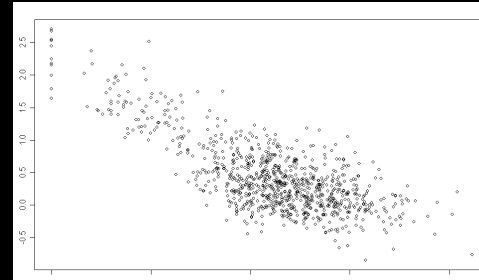
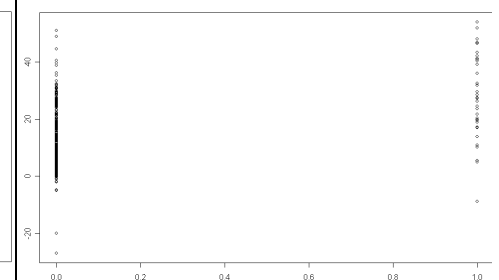
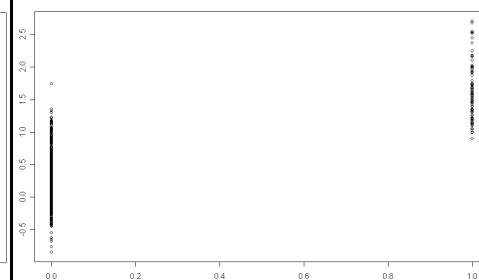
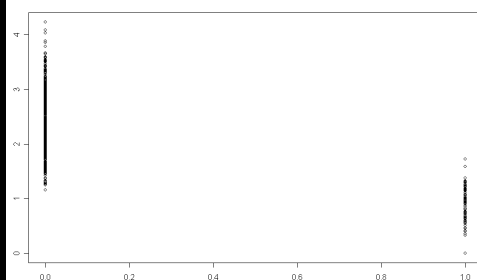
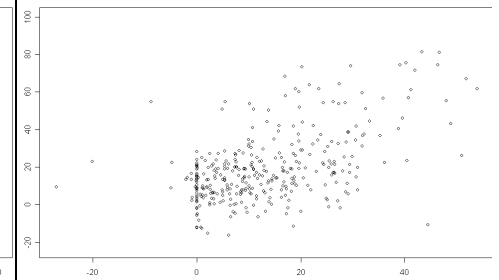
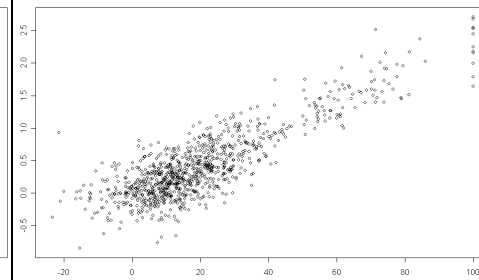
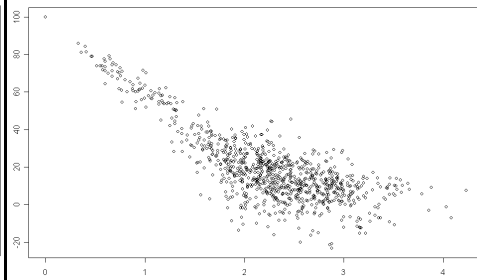
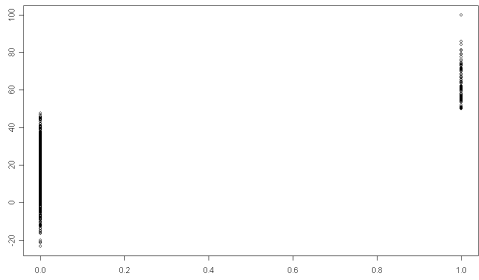
Can we take advantage of correlation?



Correlation Coefficients

	% Dia. Stenosis	Binary Restenosis	MLD	Late Loss	% Net Vol. Obs.
% Dia. Stenosis		0.51863	-0.68239	0.73754	0.45969
Binary Restenosis			-0.51634	0.51496	0.32932
MLD				-0.59840	-0.44749
Late Loss					0.52884
% Net Vol. Obs.					

Data Plots



Selected Correlation Procedures

Based on the Šidák Procedure:

- Slightly less conservative than Bonferroni
 - For non-correlated K endpoints
 - Adjusted p: $p_{ak} = 1 - (1 - p_k)^K$, for $k=1, \dots, K$

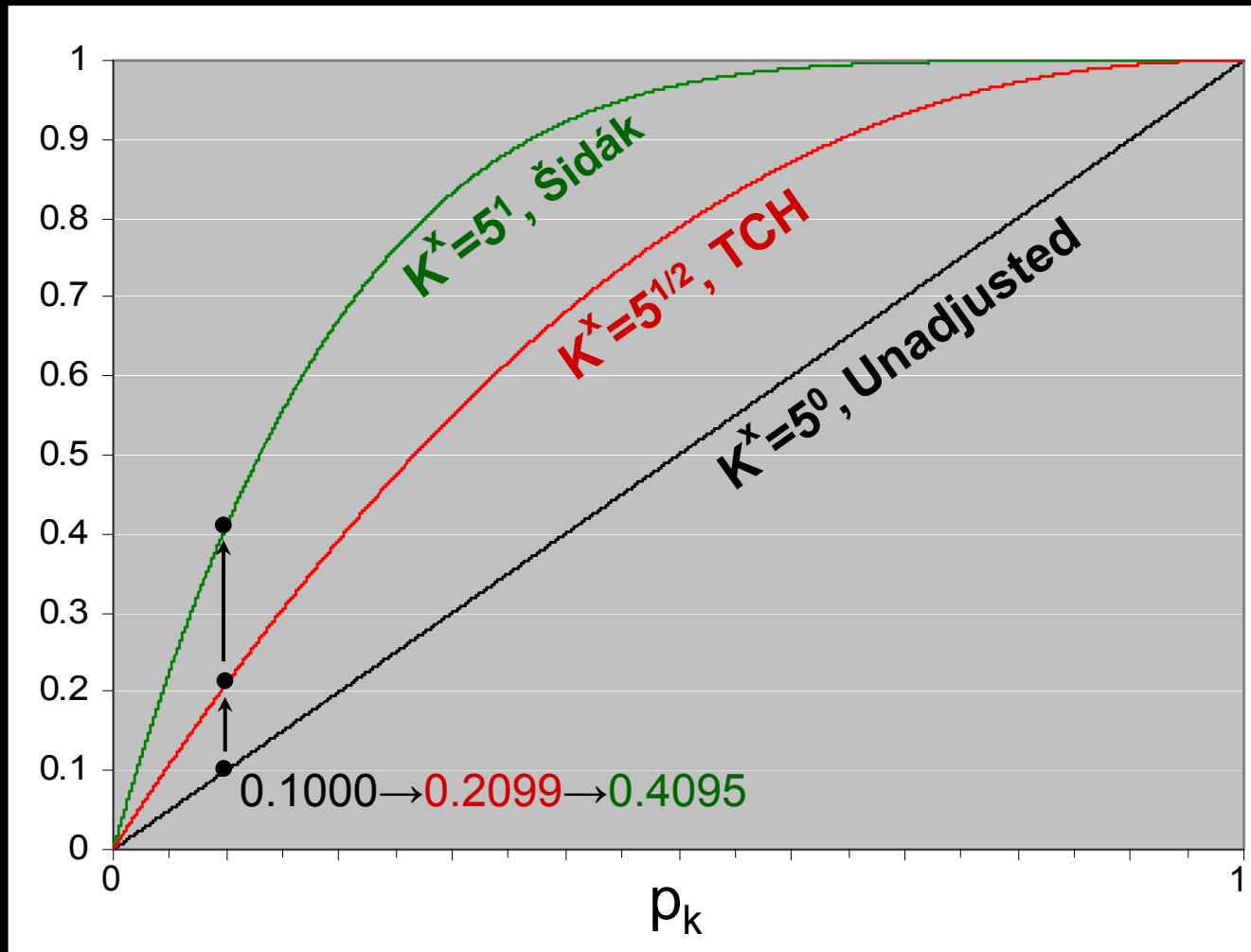


Adjustment factor that can
be influenced by a
correlation coefficient

Selected Correlation Procedures

$$p_{ak} = 1 - (1 - p_k)^{K^x}, \text{ for } k=1, \dots, K$$

Adjusted p: $1 - (1 - p_k)^{K^x}$



Selected Correlation Procedures

- TCH (Tukey, Ciminera and Heyse) Procedure:
 - For Strongly correlated K endpoints
 - Adjusted p: $p_{ak} = 1 - (1 - p_k)^{K^{1/2}}$, for $k=1, \dots, K$
- D/AP (Dubey and Armitage-Parmar) Procedure:
 - Adjusted p-values: $p_{ak} = 1 - (1 - p_k)^{m_k}$
 - where $m_k = K^{1-r_{.k}}$ and $r_{.k} = (K-1)^{-1} \sum |r_{jk}|$
 - r_{jk} is the correlation coefficient between the jth and kth endpoints
 - When the average of r_{jk} is 0 (Šidák Procedure)
 - When the average of r_{jk} is 1 (adjusted p = unadjusted p)
 - When the average of r_{jk} is 0.5 (TCH)

Selected Correlation Procedures

Endpoint	Raw	Šidák	TCH	D/AP
% Dia. Stenosis	0.0006	0.0030	0.0013	0.0011
Binary Restenosis	0.0354	0.1649	0.0774	0.0811
MLD	0.0316	0.1483	0.0693	0.0630
Late Loss	0.0001	0.0005	0.0002	0.0002
% Net Vol. Obs.	0.0021	0.0105	0.0047	0.0052

Composite Endpoints

- Reduce multiple endpoints into a single measurement
- Concerns:
 - Difficult to interpret the composite endpoint results
 - Difficult to characterize the benefits of the component endpoints

Composite Endpoints

- MACE: Cardiac Death, MI, TVR
 - Analysis Difficulties: Analysis of a specific individual component may not reflect the composite result.
 - Clinical Implications: The composite endpoint may be driven by softer components or all components may not trend positively, leading to misinterpretation or unsubstantiated claims.

False Discovery Rate (FDR)

- Multiple test controls: $\text{Prob}(V \geq 1) \leq \alpha$, where V is the number of true hypotheses falsely rejected

Fixed error rate

Estimated rejection area

- Number of multiple tests is huge: False rejections are more likely to occur

- A better control:

$$\frac{\# \text{ falsely rejected}}{\# \text{ rejected in total}}$$

- False Discovery Rate:

$$FDR = E \left[\frac{V}{R} \right]$$

Fixed rejection area

Estimated error rate

False Discovery Rate (FDR)

- As $E[V/R] \leq \text{Prob}(V \geq 1)$, FDR control is less stringent than FWE control and therefore promises higher powers.
- The FDR is much different from a p-value: a much higher FDR can be tolerated and still be quite meaningful.
- Useful for correlated testing
- FDR has been shown to be more beneficial than multiple testing for large sets of hypotheses

Concluding Remarks

- Plan ahead - Multiplicity is a factor in most trials.
- Non-Inferiority vs. Superiority (D'Agostino and Hereen).
- Know your endpoints.
- Consider multiplicity and adjustment options at the design phase.
 - Correlation procedures (there is little theory).
 - Sequential testing.
 - Flexible fixed-sequence approach.

References

- Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. Analysis of Clinical Trials Using SAS®: A Practical Guide. Cary, NC: SAS Institute Inc.
- Feise RJ. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology* 2002, 2:8.
- Horn M, Dunnett CW. Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. Recent Developments in Multiple Comparison Procedures, Institute of Mathematical Statistics, Lecture Notes – Monograph Series, Vol. 47 (2004) 48-64.
- <http://www.endovasc.com/images/graphics/stent.jpg>
- <http://www.northshorelij.com/images/phyarticles/stent.jpg>
- Huque M. Multiple Endpoint Testing in Clinical Trials – Some Issues & Considerations. FDA/Industry Workshop, Washington, DC, 2005.
- Huque M. Multiplicity Challenges and Problems in the Design and Review of Controlled Clinical Trials Regulatory Experiences and Concerns. PhRMA 2001 Biostatistics and Clinical Data Management Workshop, Hyatt Regency, Bethesda, MD, November 5-7, 2001.
- Huque M. On Some Statistical Considerations in Testing for Multiple Endpoints in Clinical Trials. ASA Biopharm Section FDA/Industry Workshop, Washington, DC, September 21-23, 2004.
- Sankoh AJ, Huque MF, Dubey SD. Some Comments on Frequently Used Multiple Endpoint Adjustment Methods in Clinical Trials. *Statistics in Medicine*, Vol. 16, 2529-2542 (1997).
- Scheid S. Multiple testing: False discovery rate and the q-value. <http://compdiag.molgen.mpg.de/docs/storeypp4.pdf>
- Tamhane AC, Logan BR. Multiple Endpoints: An Overview and New Developments. Division of Biostatistics, Medical College of Wisconsin, Technical Report 43, September 2003.
- Westfall PH, Wolfinger RD. Closed Multiple Testing Procedures and PROC MULTTEST. <http://support.sas.com/documentation/periodicals/obs/obswww23/>.
- Zhang J, Quan H, Ng j, Stepanavage ME. Some Statistical Methods for Multiple Endpoints in Clinical Trials. *Controlled Clinical Trials* 18:204-221 (1997).

Thank You!