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 $\alpha$

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

<table>
<thead>
<tr>
<th>$K$</th>
<th>$\alpha$</th>
<th>$1-(1-\alpha)^K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.0500</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.0975</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>0.1426</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.2262</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>0.4013</td>
</tr>
</tbody>
</table>

$\Rightarrow$ Control $\alpha_K$ with a multiple test procedure
## Dilemma

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ True</th>
<th>$H_0$ False</th>
</tr>
</thead>
</table>
| **Reject $H_0$** | Type I Error, $\alpha$  
**Multiplicity inflates $\alpha$** | Correct, 1-$\beta$  
(Power) |
| **Accept $H_a$** | Correct, 1-$\alpha$ | Type II Error, $\beta$  
Decreasing $\alpha$ will inflate $\beta$ |

Controlling both $\alpha$ and $\beta$ leads to increased sample size.

---

AdvaMed 2008
Data

Test Product:
Coronary Artery Stent

Analysis Design:
Comparison of two stent products (non-inferiority)
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:

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

* 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**

<table>
<thead>
<tr>
<th>Secondary Endpoint</th>
<th>Planned Non-Inferiority Margin</th>
<th>Analysis Margin</th>
<th>Analysis P-value</th>
<th>Bonferroni Adjusted Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Dia. Stenosis</td>
<td>6.6</td>
<td>4.44</td>
<td>0.0006</td>
<td>84%</td>
</tr>
<tr>
<td>Bin. Restenosis</td>
<td>6.3%</td>
<td>5.98%</td>
<td>0.0354</td>
<td>83%</td>
</tr>
<tr>
<td>MLD</td>
<td>-0.17</td>
<td>-0.16</td>
<td>0.0316</td>
<td>85%</td>
</tr>
<tr>
<td>Late Loss</td>
<td>0.18</td>
<td>0.04</td>
<td>0.0001</td>
<td>83%</td>
</tr>
<tr>
<td>% Net Vol. Obs.</td>
<td>5.7</td>
<td>3.88</td>
<td>0.0021</td>
<td>83%</td>
</tr>
</tbody>
</table>
Single Step

Šidák Procedure:

– For K endpoints, one accepts as statistically significant all those p-values ≤ \(1-(1-\alpha)^{1/K}\)

– The adjusted p-values are \(1-(1-p_k)^K\), k=1,2, … ,K

– Does not always preserve FWER
Closed Testing

• Test each hypothesis in the closed family using a suitable $\alpha$.
• 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 $p_1 \leq p_2 \leq \ldots \leq p_K$

– Reject the null if $p_k \leq \alpha/(K-k+1)$, $k=1,2,\ldots,K$

– 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 $p_1 \geq p_2 \geq \ldots \geq p_k$

- 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

- Raw p-value
- Bonferroni
- Modified Bonf
- Simes

Endpoints:
- TVR
- Binary Restenosis
- MLD
- % Net Vol. Obs.
- % Dia. Stenosis
- Late Loss
Correlated Endpoints

Can we take advantage of correlation?

- High correlation:
  - Homogeneity of treatment effects across endpoints:
    - Low: Small adjustments
    - High: Practically no adjustments

- Low correlation:
  - Good case for combining endpoints
## Correlation Coefficients

<table>
<thead>
<tr>
<th>% Dia. Stenosis</th>
<th>Binary Restenosis</th>
<th>MLD</th>
<th>Late Loss</th>
<th>% Net Vol. Obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Dia. Stenosis</td>
<td></td>
<td>0.51863</td>
<td>-0.68239</td>
<td>0.73754</td>
</tr>
<tr>
<td>Binary Restenosis</td>
<td></td>
<td>-0.51634</td>
<td>0.51496</td>
<td>0.32932</td>
</tr>
<tr>
<td>MLD</td>
<td></td>
<td></td>
<td>-0.59840</td>
<td>-0.44749</td>
</tr>
<tr>
<td>Late Loss</td>
<td></td>
<td></td>
<td></td>
<td>0.52884</td>
</tr>
<tr>
<td>% Net Vol. Obs.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 \( k=1, \ldots, K \)

Adjustment factor that can be influenced by a correlation coefficient
Selected Correlation Procedures

\[ p_{ak} = 1 - (1 - p_k)^{K^x} \], for \( k = 1, \ldots, K \)

\( K^x = 5^0, \) Unadjusted

\( K^x = 5^{\frac{1}{2}}, \) TCH

\( K^x = 5^1, \) Šidák

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

0.1000 → 0.2099 → 0.4095
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 \( k=1, \ldots ,K \)

• **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

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Raw</th>
<th>Šídák</th>
<th>TCH</th>
<th>D/AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Dia. Stenosis</td>
<td>0.0006</td>
<td>0.0030</td>
<td>0.0013</td>
<td>0.0011</td>
</tr>
<tr>
<td>Binary Restenosis</td>
<td>0.0354</td>
<td>0.1649</td>
<td>0.0774</td>
<td>0.0811</td>
</tr>
<tr>
<td>MLD</td>
<td>0.0316</td>
<td>0.1483</td>
<td>0.0693</td>
<td>0.0630</td>
</tr>
<tr>
<td>Late Loss</td>
<td>0.0001</td>
<td>0.0005</td>
<td>0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td>% Net Vol. Obs.</td>
<td>0.0021</td>
<td>0.0105</td>
<td>0.0047</td>
<td>0.0052</td>
</tr>
</tbody>
</table>
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

- 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:

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

- 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!