Statistical Give and Take: Power Implications on Strong Control of the Type I Error Rate

CHRIS HOLLAND
DIRECTOR, BIOSTATISTICS
MACROGENICS, INC.

Abstract

- New gatekeeping statistical procedures provide strong control of the Type I error rate for increasingly complex clinical trial study designs.

- These procedures allow for greater assurance that positive results will make it into an approved product’s labeling.

- An increased sample size is needed in order to provide adequate power for meeting a study’s key objectives.

- However, true power implications between competing procedures or approaches are not always well understood.
Abstract (cont.)

- Researchers need to balance competing objectives with respect to budgets, timelines, and endpoints that are required for approval and those that are “nice to have” in the labeling.

- We will look at one such procedure, the truncated Holm test, applied to a situation that involves multiple treatment groups and co-primary endpoints.

- The properties of this "separable" test will be examined with respect to varying truncation fractions and Type II error rates compared to a standard Holm procedure.

Background -- Definitions

- **Gatekeeping:**
  - A rule or set of rules to control how one moves from testing one hypothesis to the next.

- **Serial gatekeeping:**
  - Gatekeeping in a sequential manner, where one hypothesis is tested at a time.

- **Parallel gatekeeping:**
  - Gatekeeping in a parallel manner, where multiple hypotheses may be tested at a time.
**Background -- Definitions**

- **Family:**
  - A set of hypotheses. Families are usually ordered based on importance.

- **Separable:**
  - A trait of a multiple testing procedure (MTP) that allows one to carry forward the Type I error rate for any rejected hypotheses in a given family to test hypotheses in the next ordered family.

---

**Background**

- There are many MTPs that function as either serial or parallel gatekeeping procedures that control the Type I error within a family:
  - Bonferonni
  - Holm
  - Hochberg
  - Fallback
  - Dunnett

- However, with the exception of the Bonferonni test, these tests are not separable.
Example: The Holm step-down MTP

- Order the acquired p-values, denoted as $p_{(1)}$ to $p_{(n)}$, so that:
  $$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$$

- Each hypothesis associated with each ordered p-value is accepted and testing stops if:
  $$p_{(i)} > \frac{\alpha}{(n - i + 1)}$$

Otherwise reject the hypotheses and move on to the next one

Non-Separable

- The Holm procedure is not separable. Consider a family of three hypotheses: $H_{(i)}$: $\mu_{(i)} = 0$ for $i=1,2,$ and 3.

- Suppose $\mu_{(1)} \rightarrow \infty$ and $\mu_{(3)} = 0$. Therefore $p_{(1)} \rightarrow 0$ and the procedure will incorrectly reject $H_{(3)}$ with probability $\alpha$

- No alpha is left for the next family
The Truncated Holm Procedure

- The truncated Holm procedure is separable.
- One of many general multi-stage gatekeeping procedures proposed by Dmitrienko, Tamhane, Wiens (2008)
- To implement, rank each p-value and compare \( p(i) \) to:
  \[
  w_i \alpha = \left[ \gamma \left( \frac{n - i + 1}{n} \right) + \left( 1 - \gamma \right) /n \right] \alpha,
  \]
  Where \( 0 \leq \gamma < 1 \)
- Gamma (\( \gamma \)) is called the truncation fraction

Note that for \( \gamma = 0 \), the truncated Holm reduces to a basic Bonferonni procedure

For \( \gamma = 1 \), the procedure reduces to the standard Holm procedure and is no longer separable

Using the truncated Holm test, one can test subsequent families provided at least 1 significant result is obtained from the given family
The Truncated Holm Procedure

- The amount of alpha that can be carried over to the next family depends on the number of rejected hypotheses from the current family
  \[ \alpha_{i+1} = \alpha_i - \left[ \gamma + (1 - \gamma)\left(\frac{a_i}{n}\right) \right] \alpha_i \]
  When \( a_i > 0 \), where \( a_i \) represents the number of accepted hypotheses
  - Thus, if all hypotheses are accepted, then all of the alpha will have been used up and the procedure stops
- If all hypotheses are rejected, then all of the alpha can be carried to the next family. That is:
  \[ \alpha_{i+1} = \alpha_i \text{ when } a_i = 0 \]

Multiplicity Adjusted p-Values

- Adjusted p-values are calculated so that statistical significance can be ascertained
- The definition of adjusted p-values by Westfall and Young (1993) is used to calculate the adjusted p-values
  - The multiplicity adjusted p-value for a given null hypothesis and an MTP is defined as the significance level at which the procedure rejects the hypothesis
- One can therefore cycle through a set of significance levels to determine the smallest one that results in the rejection of the given hypothesis
Questions...

- How does one choose $\gamma$?
  - What are the operating characteristics?
  - For the first family, power decreases as gamma decreases,
  - But for subsequent families, “it depends”

- What are the power and sample size implications for this strong Type I error control?

Simulations

- To help answer these questions, simulations were used to examine the operating characteristics of a truncated Holm test compared to a full Holm test for one specific study design.
Simulations

- Consider the following:
  - A Phase 3 (registration) study
  - With 3 experimental treatment regimens compared to a placebo (100 subjects per arm)
  - And 2 primary endpoints.
- Thus, 6 hypotheses across 2 families exist
- Simulated scenarios include:
  - Those where 0, 1, 2, and 3 treatment groups are different than placebo with respect to primary endpoints 1 and 2

Two joint multivariate normal distributions were created
  - One representing no response
  - The other designed to provide approximately 90% power for showing a difference in the first endpoint compared to the no effect group (with alpha = 0.0167) and >99% power for the second endpoint
- 1,000 iterations generated for each scenario
Simulations

- Error rates calculated as:
  - Type I: Any time the procedure incorrectly found a statistically significant difference (across all hypotheses and families) between groups that were drawn from the same (no effect) distribution.
  - Type II: Any time the procedure failed to detect a statistically significant difference (across all hypotheses and families) between groups that were drawn from different distributions.

- Comparisons made to a full Holm procedure where one may move to the next family as long as one significant difference is found in the current family (which is known to not provide strong Type I error control).

Two-Stage Testing Procedure

![Diagram of Two-Stage Testing Procedure]

Notes: $\alpha_1 = \alpha$; $\alpha_2$ is calculated after Family 1 testing is complete; $w_i$'s are functions of $\gamma$ and $\gamma = 1$ for Family 2.
Simulation Results – Two-Stage Design

Error Rates with One Effective Treatment
Two-Stage Design

Type 1 Error
Type 2 Error

Gamma

Error Rates

0.00 0.04 0.08 0.12 0.16 0.20

0.00 0.04 0.06 0.08 0.10

0.06 0.063

0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00
Error Rates with Two Effective Treatments
Two-Stage Design

Type 1 Error with Zero Effective Treatments
and Type 2 Error with 3 Effective Treatments
Two-Stage Design
Two-Stage Design Simulation Findings

- Full Holm procedure testing results in alpha inflation and a Type I error rate that, in some cases, approached to 8%.
- For all three relevant scenarios (1, 2, and 3 effective treatment groups):
  - Truncated Holm tests, for gamma values in the 0.6-0.85 range, provided similar power to the full Holm procedure while also providing strong Type I error control
  - For the one effective treatment scenario, the Type II error spiked slightly with high gamma values (≥0.90)
- For the three effective treatments scenario, the Type II error decreased with truncation fractions (gammas)

Three-Stage Design

- Because of the strong power built into the simulation for the second endpoint, the two-stage testing results may mask, to some extent, some properties of the procedure.
- A third stage of testing was therefore added to further examine the characteristics of the truncated Holm test.
- The third endpoint was designed to have ~86% power at alpha = 0.0167 to show a difference from the control group
Three-Stage Testing Procedure

Family 1

$p_1(1) \leq w_1 \alpha_1$?

Yes $\rightarrow$ Yes $\rightarrow$

$p_1(2) \leq w_2 \alpha_2$?

Yes $\rightarrow$ Yes $\rightarrow$

$p_1(3) \leq w_3 \alpha_1$

No

STOP

Family 2

$p_2(1) \leq w_1 \alpha_2$?

Yes $\rightarrow$ Yes $\rightarrow$

$p_2(2) \leq w_2 \alpha_2$?

Yes $\rightarrow$

$p_2(3) \leq w_3 \alpha_2$

Family 3

$p_3(1) \leq w_1 \alpha_3$?

Yes $\rightarrow$

$p_3(2) \leq w_2 \alpha_3$?

Yes $\rightarrow$

$p_3(3) \leq w_3 \alpha_3$

Notes: $\alpha_1 = \alpha_2$; $\alpha_2$ and $\alpha_3$ are calculated after testing from the previous family is complete; $w_i$'s are functions of $\gamma$ and $\gamma = 1$ for Family 3.
Error Rates with One Effective Treatment
Three-Stage Design

Error Rates with Two Effective Treatments
Three-Stage Design
Three-Stage Design Simulation Findings

- Full Holm procedure testing results in, as expected, further alpha inflation with each additional stage
  - Type I error rate exceeded 11%
- For the three effective treatments scenario, power reductions with gamma > 0.8 are minimal (compared to the full Holm procedure).
- However, for the one and two effective treatment scenarios, power reductions are rather significant, especially with increasing gamma values.
  - Potentially due to fewer rejections from prior families and more alpha being spent with increasing gamma
Simulation Findings

- It is unclear whether the increased Type II error rates are the result of the third stage of testing only, or if power was decreased for the second stage of testing also.

- Plots to examine Type II error by Family/Stage were created.
Type 2 Error Rates with Two Effective Treatments
By Stage and Family

Type 2 Error Rates with Three Effective Treatments
By Stage and Family
Three-Stage Design Simulation Findings

- Power reductions in the three-stage design appear to be the result of there being less alpha available for the third stage of testing.

- Power for the second stage of testing was largely unaltered.

Sample Iteration – 2 Effective Trt Groups

<table>
<thead>
<tr>
<th>Family</th>
<th>Null Hypothesis</th>
<th>Raw p-value</th>
<th>2-Stage Design</th>
<th>3-Stage Design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \gamma = 0.3 )</td>
<td>( \gamma = 0.8 )</td>
</tr>
<tr>
<td>F1</td>
<td>H11</td>
<td>0.386</td>
<td>0.724</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>H12</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>H12</td>
<td>0.007</td>
<td>0.018</td>
<td>0.014</td>
</tr>
<tr>
<td>F2</td>
<td>H21</td>
<td>0.770</td>
<td>0.770</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>H22</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>H23</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>F3</td>
<td>H31</td>
<td>0.649</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H32</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H33</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample Iteration – 2 Effective Trt Groups

- This sample shows that lower gamma values can pay large dividends by the third stage of testing.

- One acceptance of a hypothesis at the second stage has a big effect on the alpha left over for the third stage

Conclusions

- Separaible tests, such as the truncated Holm procedure, provide strong experiment-wise Type I error control to complex study designs that involve multiple families of hypotheses

- They allow for parallel testing across families, rather than strict sequential testing, as long as one hypothesis is rejected in each family.

- In general, the choice for the truncation fraction, or gamma, is not a clear-cut decision.
  - Smaller gammas are more conservative for the first stage of testing, but leave more alpha for subsequent stages
Conclusions

- For the two-stage design tested via simulations, the truncated Holm test provided power that was comparable to that from the full Holm test
  - Gamma’s in the range of 0.8-0.9 appeared to provide the best power overall
  - (this was also true for the three-stage design with 3 effective treatment groups)
- For the other three-stage design scenarios (1 and 2 effective treatments) Type II errors increased significantly
  - Most Type II errors were from the 3rd stage of testing
  - Gammas ≤ 0.30 still, however, provided reasonable power

Suggested Strategies for Truncated Tests

- Choose larger gammas when wanting to control Type I error for a smaller number of endpoints
- Choose smaller gammas when wanting to control Type I error for a larger number of endpoints
  - Unless, perhaps, you expect all treatment groups to be effective
- Perform simulations and power your study accordingly
- Another approach might be as follows:
  - Test all treatment regimens for your co-primary or primary and key secondary endpoints
  - For additional secondary endpoints, test only the lowest effective dose sequentially, with the remaining alpha
Acknowledgements

- Thanks to the following individuals for their assistance with this presentation:
  - David Carlin, PhD (MacroGenics, Inc.)
  - Alex Dmitrienko, PhD (Eli Lilly & Co.)
  - Vladimir Mats, PhD (Quartesian, Inc.)
  - Wei Xu, PhD (MacroGenics, Inc.)
  - Leonid Zeitlin, PhD (Quartesian, Inc.)

References


Questions?

P. Chris Holland, MS
hollandc@macrogenics.com