Case Studies of Applying Multiple Testing Procedures in Neuroscience Late-Phase Clinical Trials

Haiyan Xu, Ph.D.

Janssen R&D, Johnson & Johnson

Acknowledge: **Yevgen Tymofyeyev, Ph.D. (Janssen)**
(major contribution to case studies and simulations)

ASA Biopharm Workshop September 13, 2018

Problems we face

One experiment

- Multiple dose comparisons
- Multiple endpoints
- Multiple subpopulations
- Non-inferiority + superiority
- Multiple looks, etc

Generate **multiple hypotheses**: H1, H2, H3, H4, etc

FDA, EMA, other Health authorities

Require controlling study-wise type I error at α level (two-sided 5%, or one-sided 2.5%)
{We often call this familywise error rate (FWER)}

Need to design efficient multiple testing procedures (MTPs)

(6% power difference, ~100 additional subjects, ~200-300 additional subjects in a program, e.g., ~$50k per subject, *$10-15MM increase*)

My *Naive* Classification of Multiple Testing Problems

Class 1: No prefixed testing sequence of hypotheses

("Decision paths")

Class 2: Simple (or relative simple) ordering

Class 3: More complex hierarchical structure (or "decision paths")

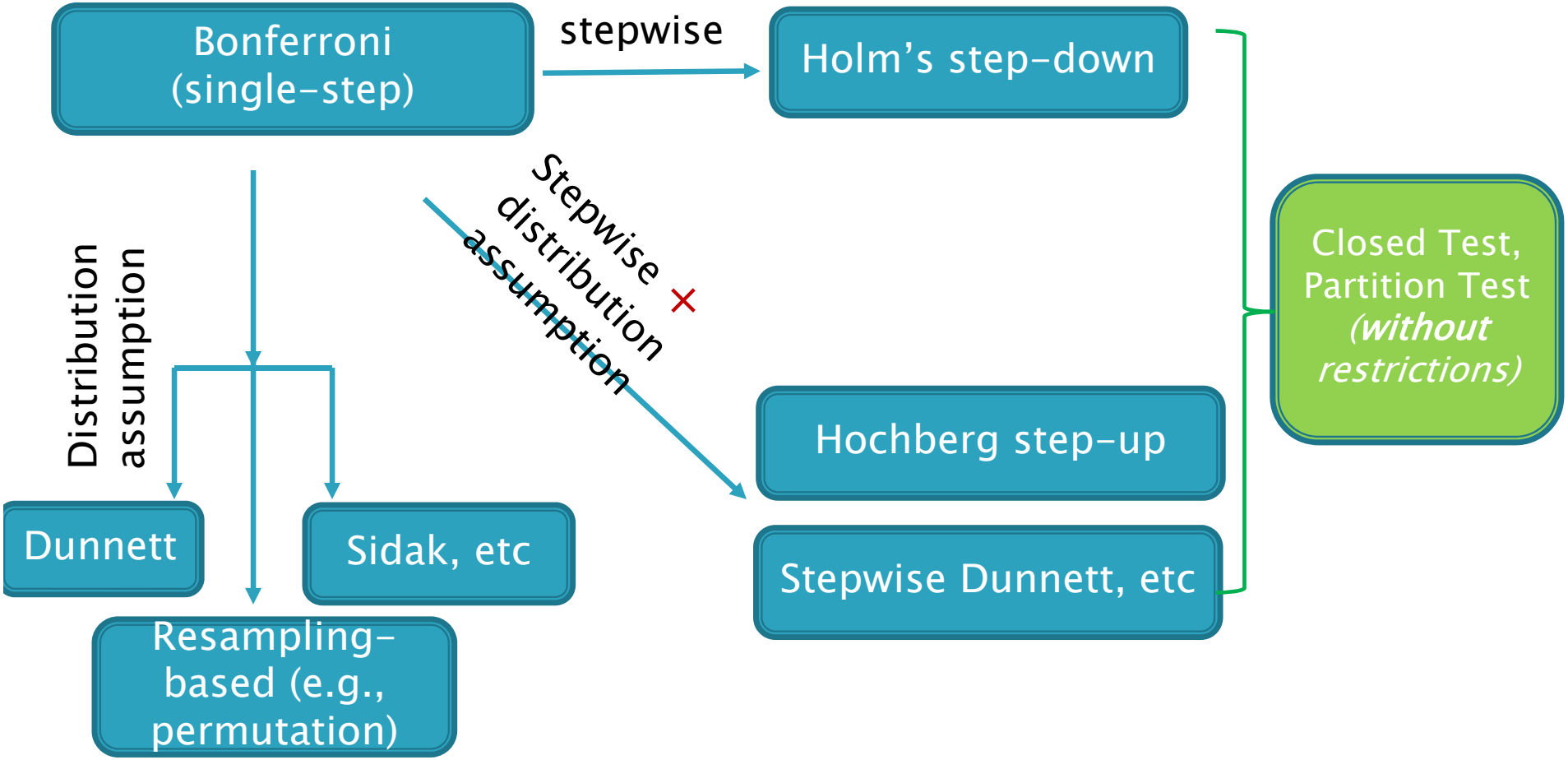
e.g., Multiple sources of multiplicity

- Multiple doses + multiple endpoints
- Multiple populations + multiple endpoints
- Etc

Class 4: multiplicities in group sequential/ adaptive design setting

(*will not discuss in detail today*)

Tools for Class 1 Problem: No prefixed testing sequence

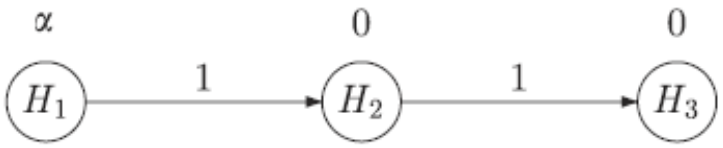


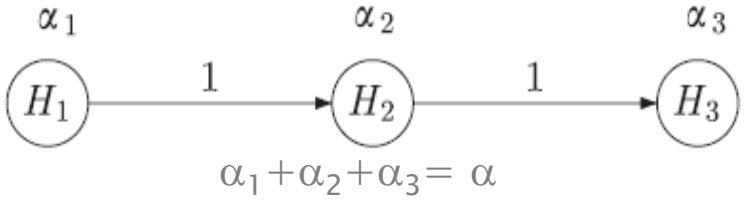
Tools for Class 2 Problem: Simple (or relatively simple) ordering

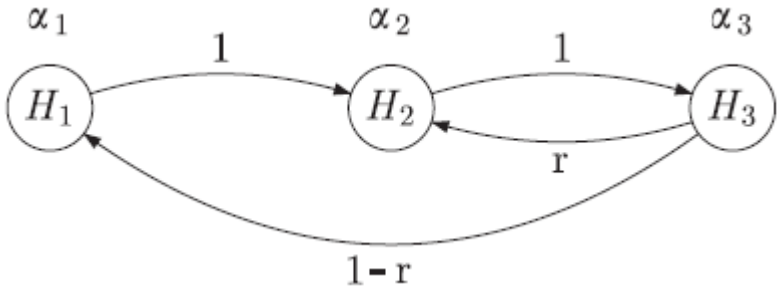
Fixed sequence

Original Fallback (Not recommended)

Improved fallback







Tools for Class 3 Problem: More complex hierarchical structure

Gatekeeping Procedures

- Serial Gatekeeping
- Parallel Gatekeeping
- Tree Structured Gatekeeping, General Mixture Gatekeeping , etc

Partitioning Decision Paths Approach (Liu and Hsu [JASA 2009])

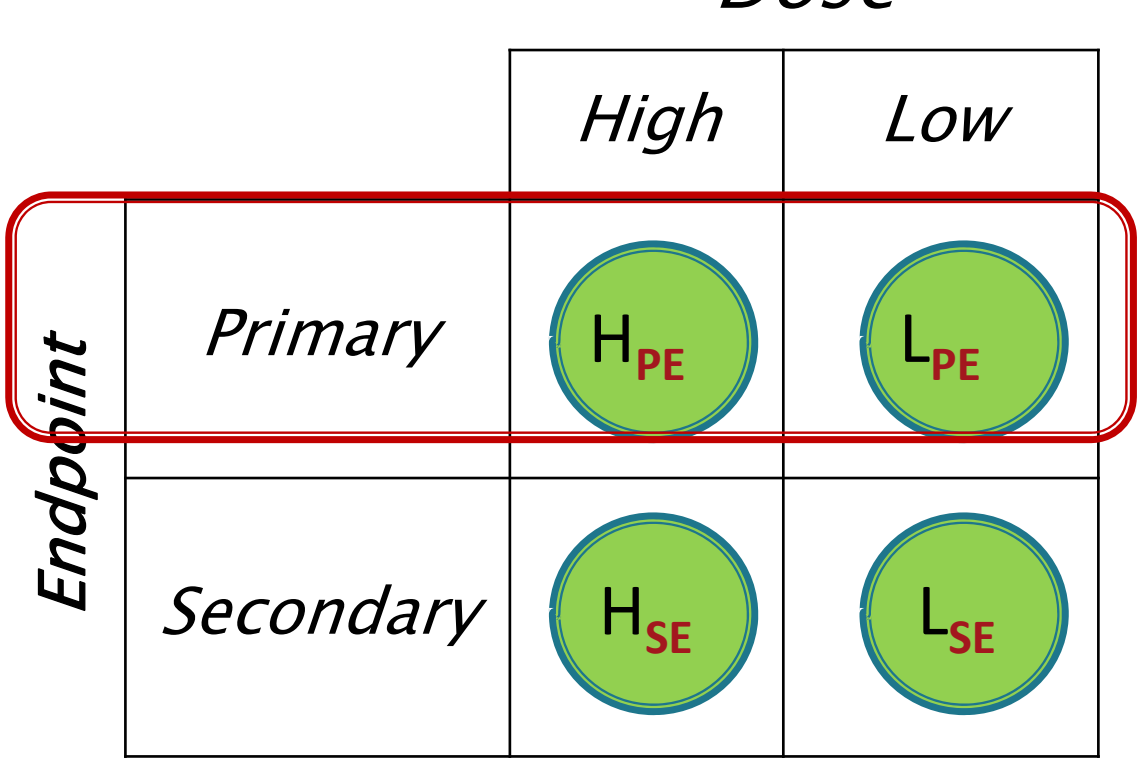
Follow the *Decision Path Principle*: *Null hypotheses should be formulated so that decision making naturally follows logical paths.*

Closed Test, Partition Test *(with restrictions)*

Case Study #1: schizophrenia/bipolar disorder in ph3

- **Primary hypothesis**: Compound X improves symptoms vs placebo as measured by a symptom scale (e.g., PANSS [Positive and Negative Syndrome Scale for Schizophrenia], YMRS for Bipolar disorder)
- **Key secondary hypothesis**: Compound X improves functioning vs placebo based on a functioning scale (e.g., PSP [Personal and Social Performance Scale] for schizophrenia, GAF for Bipolar disorder)
- **4 arms**: placebo, 3 doses of the new treatment

Case Study #1: schizophrenia/bipolar disorder in ph3 (cont'd)



Potential strategy: test primary endpoint first, if both doses are rejected, then test secondary (so-called serial gatekeeping)

However, it is of great interest to show effect in one dose, but in both endpoints.

Case Study #1: schizophrenia/bipolar disorder in ph3 (cont'd)

- Initially, Dunnett-based parallel gatekeeping testing procedure was proposed, based on Dmitrienko et al (2006)
- There was concern regarding utilizing sample based correlation between endpoints
- Dunnett-Bonferroni-based parallel gatekeeping procedure was utilized, based on Xu et al (2009)

Intersection hypotheses	Decision rule
H_{111111}, H_{111110}, H_{111101}, H_{111100}, H_{111011}, H_{111010}, H_{111001}, H_{111000}	$T_1^P > c_1$ or $T_2^P > c_1$ or $T_3^P > c_1$
H_{110110}, H_{110100}, H_{110010}, H_{110000}	$T_1^P > c_1$ or $T_2^P > c_1$
H_{101101}, H_{101100}, H_{101001}, H_{101000}	$T_1^P > c_1$ or $T_3^P > c_1$
H_{011011}, H_{011010}, H_{011001}, H_{011000}	$T_2^P > c_1$ or $T_3^P > c_1$
H_{100100}, H_{100000}	$T_1^P > c_1$
H_{010010}, H_{010000}	$T_2^P > c_1$
H_{001001}, H_{001000}	$T_3^P > c_1$
H_{110101}, H_{110011}, H_{110001}, H_{110111}	$T_1^P > c_1$ or $T_2^P > c_1$ or $T_3^S > c_2$
H_{101111}, H_{101110}, H_{101011}, H_{101010}	$T_1^P > c_1$ or $T_3^P > c_1$ or $T_2^S > c_2$
H_{011111}, H_{011110}, H_{011101}, H_{011100}	$T_2^P > c_1$ or $T_3^P > c_1$ or $T_1^S > c_2$
H_{100111}, H_{100011}	$T_1^P > c_1$ or $T_2^S > c_3$ or $T_3^S > c_3$
H_{010111}, H_{010101}	$T_2^P > c_1$ or $T_1^S > c_3$ or $T_3^S > c_3$
H_{001111}, H_{001110}	$T_3^P > c_1$ or $T_1^S > c_3$ or $T_2^S > c_3$
H_{100110}, H_{100010}	$T_1^P > c_1$ or $T_2^S > c_4$
H_{100101}, H_{100001}	$T_1^P > c_1$ or $T_3^S > c_4$
H_{010110}, H_{010100}	$T_2^P > c_1$ or $T_1^S > c_4$
H_{010011}, H_{010001}	$T_2^P > c_1$ or $T_3^S > c_4$
H_{001101}, H_{001100}	$T_3^P > c_1$ or $T_1^S > c_4$
H_{001011}, H_{001010}	$T_3^P > c_1$ or $T_2^S > c_4$
H_{000111}	$T_1^S > c_5$ or $T_2^S > c_5$ or $T_3^S > c_5$
H_{000110}	$T_1^S > c_6$ or $T_2^S > c_6$
H_{000101}	$T_1^S > c_6$ or $T_3^S > c_6$
H_{000011}	$T_2^S > c_6$ or $T_3^S > c_6$
H_{000100}	$T_1^S > c_7$
H_{000010}	$T_2^S > c_7$
H_{000001}	$T_3^S > c_7$

A closed testing based procedure, with 6 individual hypotheses

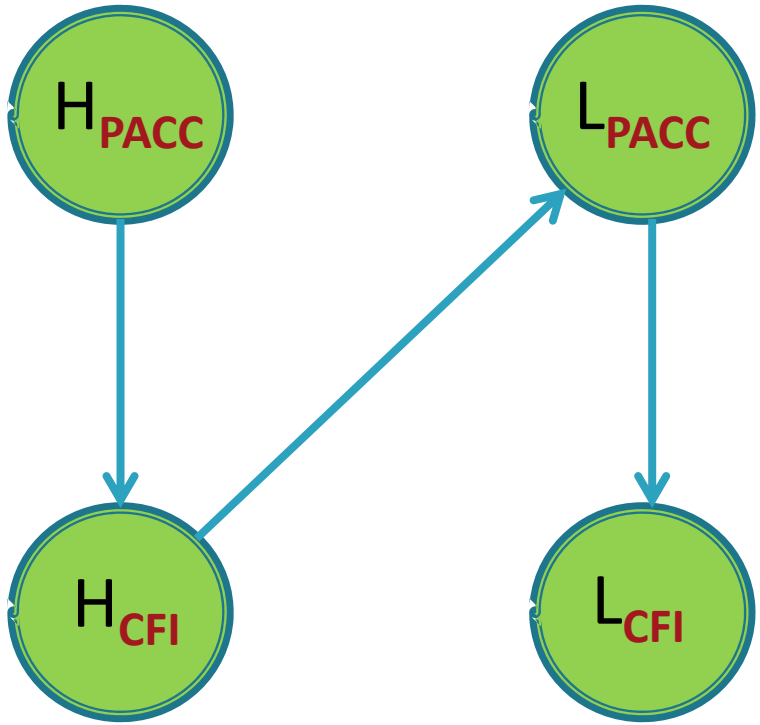
- An alternative approach: Partitioning Decision Path (Liu and Hsu [JASA 2009])

Case Study #2: Neurodegenerative

- **Primary Hypothesis**: Compound X slows cognitive decline vs placebo as measured by PACC
- **Key Secondary Hypothesis**: Compound X improves cognitive function and performance vs placebo based on CFI
- **3 arms**: placebo, low dose and high dose of new treatment

Fixed Sequence Testing

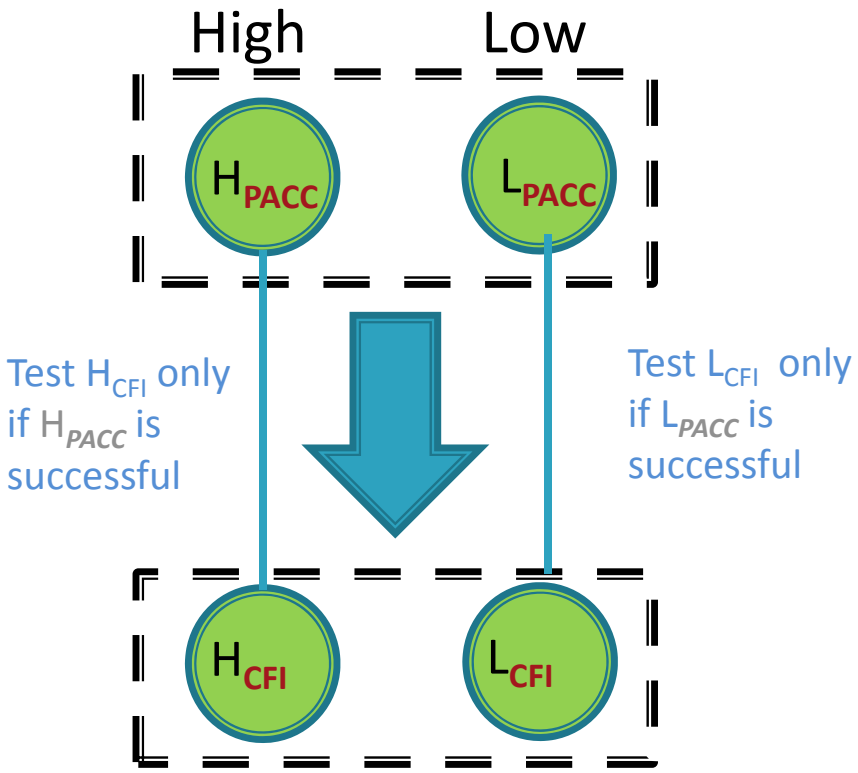
α=5% used in testing sequence until fail a hypothesis



"Sequential by Dose"

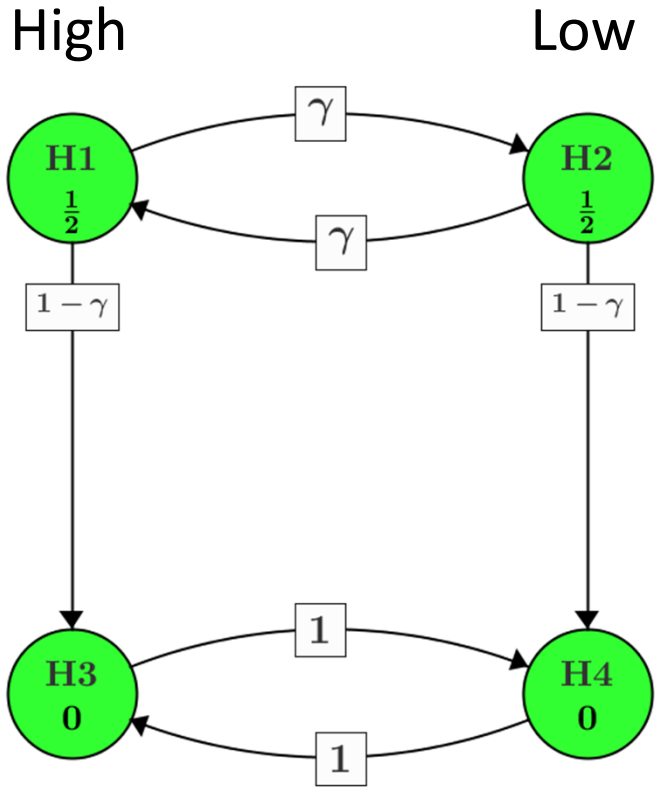
Parallel Gatekeeping

Truncated Hochberg



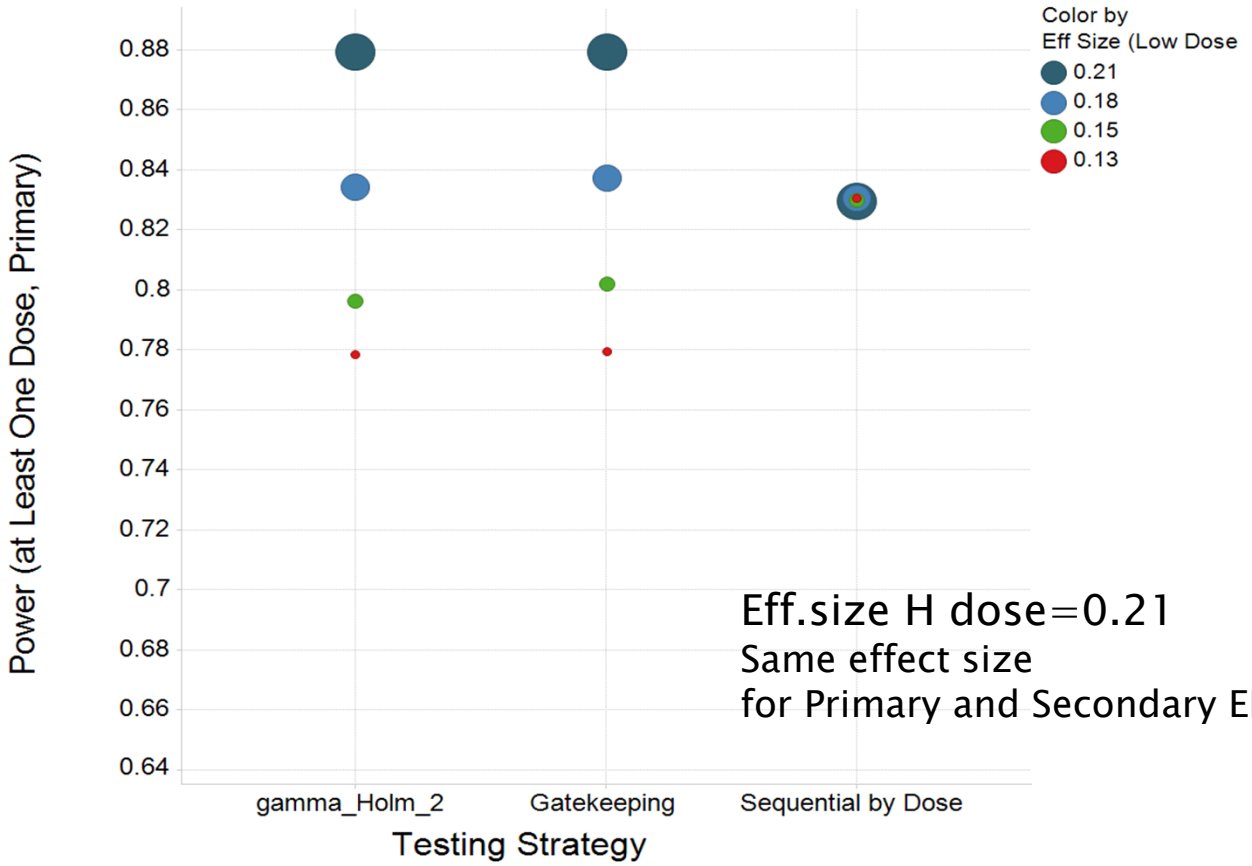
Hochberg test for CFI family

Truncated Holm's



Truncation parameter (γ): ensure that some fraction of α is left for CFI even if only one dose is significant for PACC

Power, at Least One Dose superior to PLB (Primary EP)



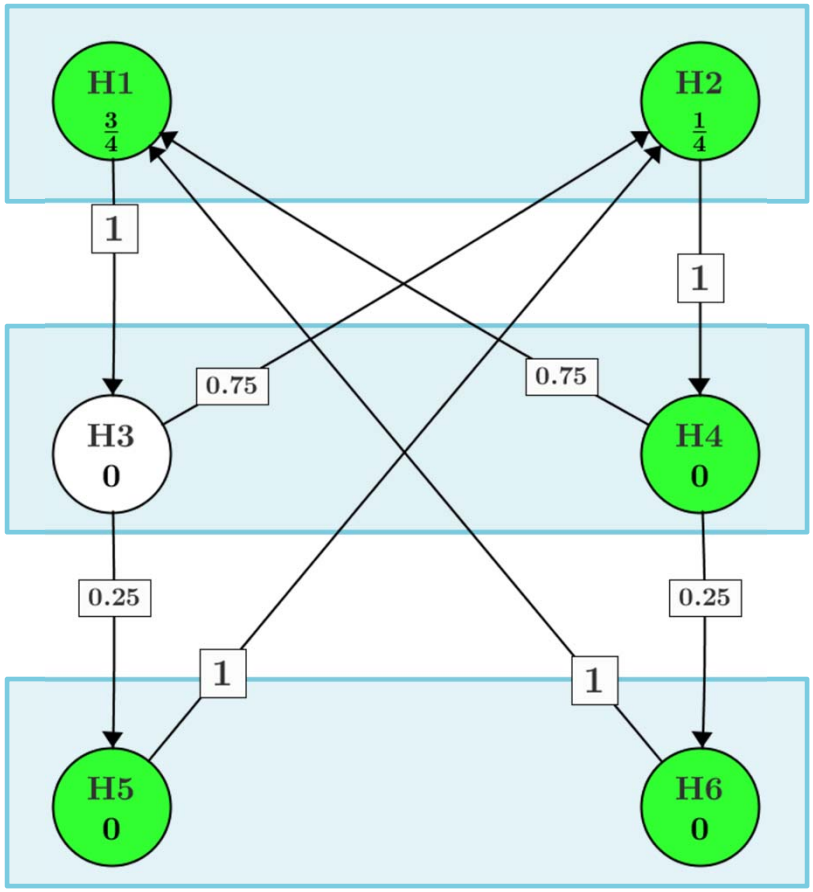
Eff.size H dose=0.21
Same effect size
for Primary and Secondary EP

- No uniformly most powerful procedure is available. Power results depends on the underlying true scenario
  - Need to decide which scenario are more likely and optimize MTP accordingly

Case Study #3: indication not revealed, non-inferiority and superiority + multiple endpoints

- **Primary objective**: establish **non-inferiority** on primary endpoint, of two doses of Compound X versus active control
- 1st **Key secondary objective**: compare the effects of two doses of Compound X versus active control on cognitive function
- 2nd **Key secondary objective**: establish **superiority** on primary endpoint, of two doses of Compound X versus active control

Recommended MTP



Primary, **non-inferiority on primary endpoint**, with NI margin 5

Key Secondary, **superiority on cognitive function**

Superiority on primary endpoint

γ -- controls α transferred to primary endpoint superiority

Power by initial weight allocated to primary endpoint NI high dose

Initial Weight for NI on high dose	NI margin/SD	PE true delta high dose/SD	PE true delta low dose/SD	Delta Cognitive high dose, effect size	delta Cognitive low dose, effect size	Power PE NI high dose	Power PE NI low dose	Power Cognitive high dose	Power Cognitive low dose	Power PE superiority high dose	Power PE superiority low dose
0.5	0.1	0.2	0.1	0.3	0.4	0.864	0.623	0.748	0.609	0.335	0.227
0.75	0.1	0.2	0.1	0.3	0.4	0.884	0.607	0.782	0.592	0.338	0.228
1	0.1	0.2	0.1	0.3	0.4	0.903	0.546	0.815	0.538	0.343	0.230
0.5	0.1	0.2	0.2	0.3	0.4	0.887	0.878	0.788	0.859	0.487	0.489
0.75	0.1	0.2	0.2	0.3	0.4	0.893	0.863	0.798	0.841	0.487	0.489
1	0.1	0.2	0.2	0.3	0.4	0.903	0.736	0.815	0.725	0.493	0.491

PE: primary endpoint

Team's Inputs: Significant Impact on MTP design

- From commercial and regulatory point of view, do we have to have more than two endpoints? And if we do, how do we order their importance?
  - – impact on sample size
- Is there a clear dose response to order the arms? How confident are we in terms of betting on a strict order, or a particular arm?
  - – impact on simulation and which MTP to choose and also the sample size
- What is the reasonable effect size for each arm and each endpoint?
  - – impact on simulation results and hence the decision of which MTP to choose and also the sample size
- How do we define "win": win on at least one dose, or win on at least two doses?
  - – impact on power simulation and hence sample size
- Is there preference between winning on a particular regimen vs. another?
  - – impact on allocating weights and redistribution weights, and hence sample size

Objectives, priorities and assumptions imply preferred MT strategies

- Needs to be supported by **power simulation** under different scenarios

Team Interaction

- Train other functions
  - e.g., Multiple testing workshop – introducing the methods and impact, with real case studies
- Understand Target Product Profile (e.g., certain advantage vs. competitor drug) and get involved in strategic level discussion
- Meta analyses (internal + external vendor) for design assumptions
- Comprehensive simulations with easy-to-understand data display

In Summary

- We encounter many multiple testing problems
- Important to partner with other functions to come up with an efficient MTP, which also aligns with the development strategy (priorities)
- Methods evolving over time:
  - Case 1: Dunnett/Dunnett-Bonferroni gatekeeping: "left alpha on the table"
  - Case 2: Alpha exhaustive, but did not "recycle" alpha back to higher-level families
  - Case 3: Alpha exhaustive, and "recycle" alpha back to higher-level families
- New challenges:
  - Subgroup in confirmatory setting (how to deal with joint distribution for non-continuous outcome)
    - Ding et al (2016); Lin et al (2018), to appear
  - Multiplicity adjustment while searching for subgroups
  - Move towards confidence intervals

References

Bretz F, et al. *A graphical approach to sequentially rejective multiple test procedures*. Stat in Med, 2009; 28:586–604

Bretz F, et al. *Graphical approaches for multiple comparision procedures using weighted Bonferroni, Simes, or parametric tests*. Biometrical Journal, 2011; 53: 894–913

Bretz F, et al. *Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures*. Stat in Med, 2011; 30: 1489–1501

Burman, Sonesson, Guilbaud (2009) *A recycling framework for the construction of Bonferroni based multiple tests. Statistics in Medicine* 28: 739–761.

Ding, Y., Lin, H., and Hsu, J.C. (2016). Statistics in Medicine 35, 1580–1594.

Dmitrienko A, et al. *Gatekeeping strategies for clinical trials that do not require all primary effects to be significant*. Stat in Med, 2003; 22:2387–2400

Dmitrienko A, Offen W, Wang O, Xiao D. Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics* 2006; **5**: 19–28.

Dmitrienko A, Wiens BL, Tamhane AC, Wang X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* 2007; **26**: 2465–2478.

Hochberg (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802.

Holm S. *A simple sequentally rejective multiple test procedure*. Scandinavian Journal of Statistics 1979; 6:65–70.

Lin, H., Xu, H., Ding, Y. and Hsu, J.C. (2018). Biometrical Journal. To appear.

References

Hommel G, et al. *Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies*. Stat in Med, 2007; 26:4063-4073

Liu Y, Hsu J. Testing for Efficacy in Primary and Secondary Endpoints by Partitioning Decision Paths. *Journal of American Statistical Association* 2009: 1661-1670.

Maurer, Hothorn, Lehmacher (1995) Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie* 6:3-18.

Maurer, Bretz (2013) *Memory and other properties of multiple test procedures generated by entangledgraphs*. Statistics in Medicine 32: 1739-1753.

Maurer, Bretz (2013) *Multiple testing in group sequential trials using graphical approaches.* Statistics in Biopharmaceutical Research 5: 4, 311-320.

Wiens BL. *A fixed sequence Bonferroni procedure for testing multiple endpoints*. Journal of Biopharmaceutical Statistics 2003; 2:211-215.

Wiens, Dmitrienko (2005) The fallback procedure for evaluating a single family of hypotheses. Journal of *Biopharmaceutical Statistics* 15: 929-942.

Rohmeyer, Klinglmueller (2011) *gMCP: Graphical approach to multiple comparison procedures.*

Wassmer, Dragalin (2015) *Designing issues in confirmatory adaptive population enrichment trials*. JBS 25: 651-669

Xi and Tamhane (2015) *Allocating recycled significance levels in group sequential procedures for multiple endpoints.* Biometrical Journal 57: 90-107.

Xu, Nuamah, Liu, Lim, Sampson (2009), A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints. Pharmaceut. Statist.; 8: 301-316