Overview of Multiplicity Adjustment Issues in Clinical Trials

Eric Nantz & Jiajun Liu
Eli Lilly and Company
Biopharmaceutical Applied Statistics Symposium
24-OCT-2017

The Big Picture

Late-phase clinical trials include multiple objectives to achieve different goals:

Regulatory Requirement

Greater Differentiation

Better Reimbursement

Meaningful for Patients

Prioritization: Not Always Easy

- Determining highest-priority objective is usually straight-forward (primary endpoint)
- How about remaining objectives?
  - Likelihood of success
  - Dosing strategy
  - “Nice to have” versus meaningful impact to patients and payers

Taking on Multiplicity

- Regulatory agencies require **strong** control of **Familywise Error Rate**

$$FWER \leq \alpha$$

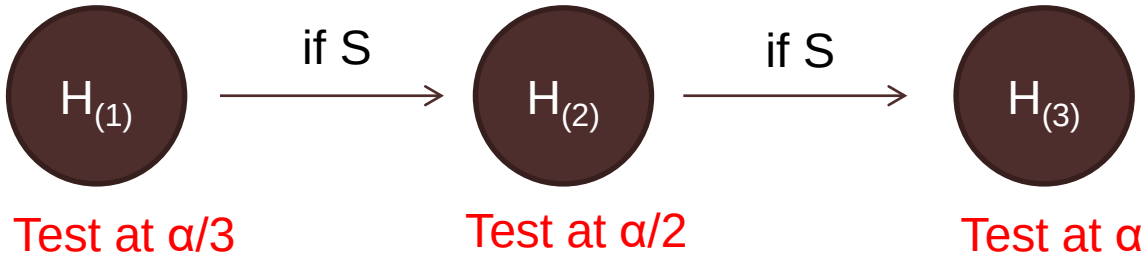
- **FWER Translation**: Probability of at least one false positive (under any configuration of actual true/false null hypotheses)

Common Testing Procedures

- Single-step procedure: Bonferroni
- Stepwise procedures with data-driven hypothesis ordering
  - Holm (Holm, 1979)
  - Hochberg (Hochberg, 1988)
  - Hommel (Hommel, 1988)
- Stepwise procedures with pre-specified hypothesis ordering
  - Fixed sequence procedure
  - Fallback testing procedure (Wiens et al, 2005)
- Procedures of mixed type (pre-specified priority, but actual ordering partially driven by data)
  - Graphical approach (Bretz et al, 2009)

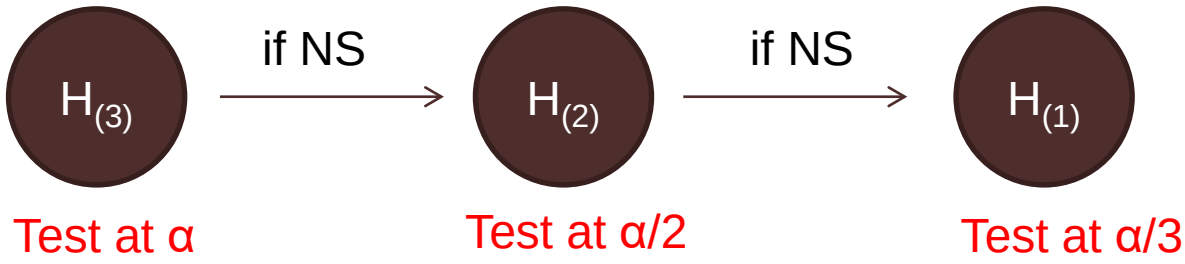
Holm procedure

- 'Step down' algorithm: starting with the hypothesis associated with the most significant p-values: $p_{(1)}, \ldots, p_{(m)}$
  - Step 1: If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$ and go to Step 2 otherwise accept all hypotheses and stop
  - Step i=2,…, m-1: if $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(i)}$ and go to Step i+1, otherwise accept $H_{(i)}, \ldots, H_{(m)}$ and stop
  - Step m: if $p_{(m)} \leq \alpha$, reject $H_{(m)}$, otherwise accept $H_{(m)}$



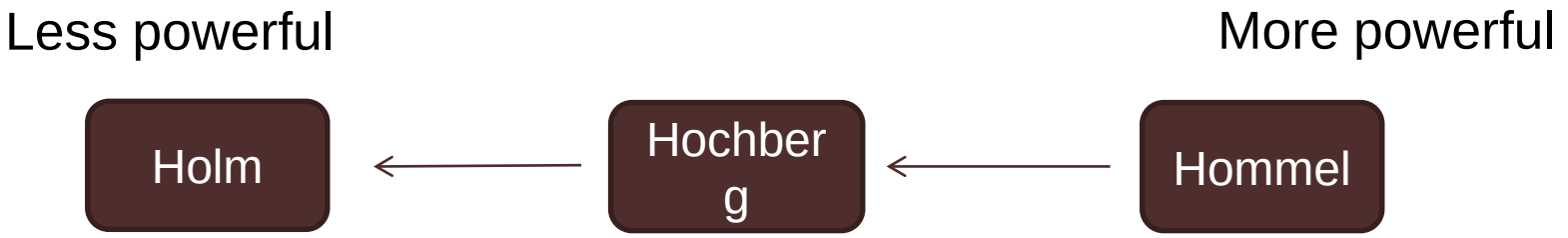
Hochberg procedure

- 'Step up' algorithm: starting with the hypothesis associated with the least significant p-values
  - Step 1: If $p_{(m)} > \alpha$, accept $H_{(m)}$ and go to Step 2 otherwise reject all null hypotheses and stop
  - Step i=2,…, m-1: if $p_{(m-i+1)} > \alpha/i$, accept $H_{(m-i+1)}$ and go to Step i+1, otherwise reject all remaining null hypothesis and stop
  - Step m: if $p_{(1)} > \alpha/m$, accept $H_{(1)}$, otherwise reject $H_{(1)}$

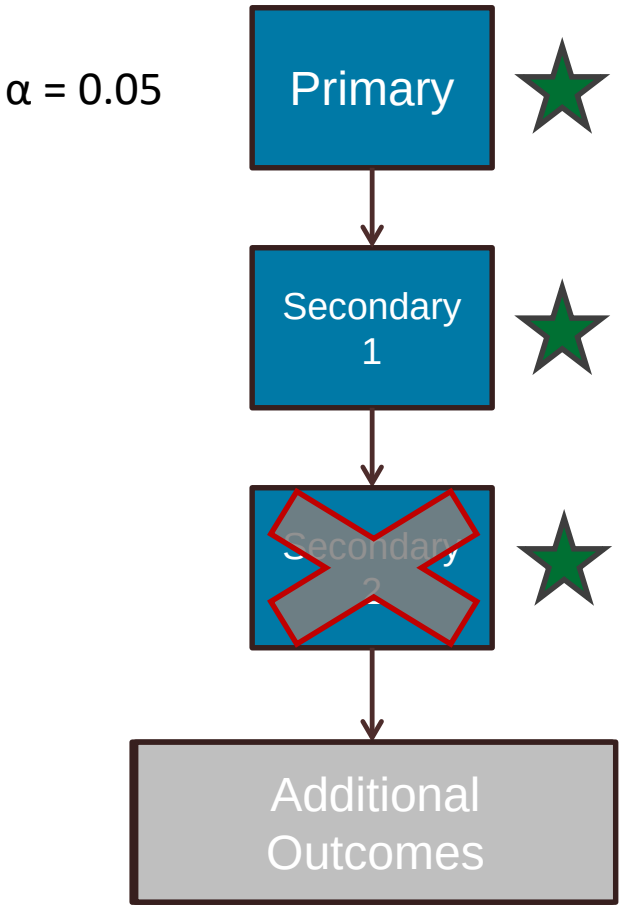


Data-Driven Stepwise Procedures

- Holm strongly controls FWER for any joint distributions of hypothesis tests statistics
- Hochberg and Hommel control FWER when hypothesis test statistics are independent or positively correlated
- Hochberg is the most popular due to its simple algorithm
- Power comparison



Gatekeeping Approach



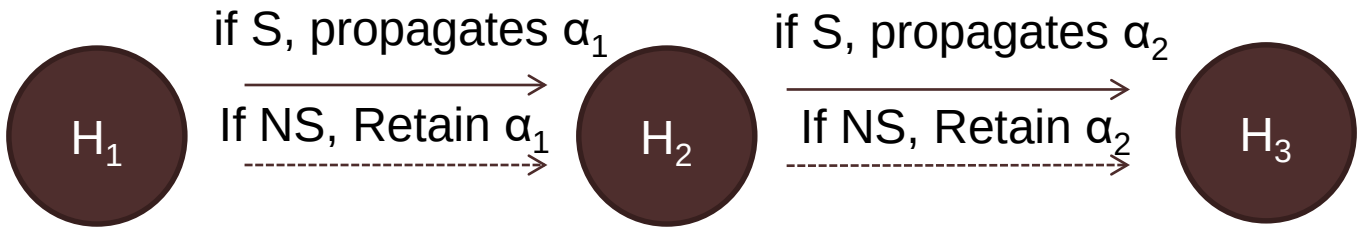
Advantage: Powerful when the pre-specified order matches the ranking of the p-values

Disadvantage: Outcomes do not have the success we anticipated …

- Additional outcomes may have performed well, but will not get a chance if any outcome above fails to pass
- **All or Nothing**: No opportunities to recycle alpha

Fallback Procedure

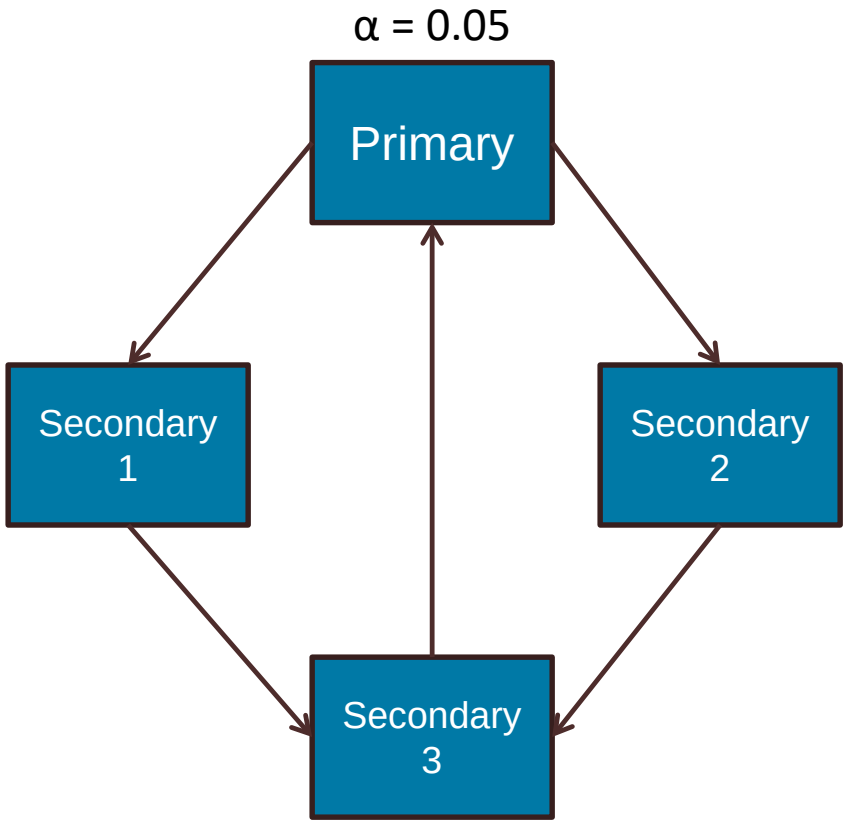
- An attractive alternative to fixed sequence procedure: No single strike rule.
  - Assign overall error rate α according to their weights $\alpha_i=\alpha w_i$, i=1, …, m to ordered hypotheses, H_1, …, H_m.
  - Test H_i at α_i , if H_i is rejected, test H_{i+1} at $\alpha_{i+1}=\alpha_i+\alpha w_{i+1}$, if H_i is accepted, test H_{i+1} at $\alpha_{i+1}=\alpha w_{i+1}$.



Fallback Procedure

- Closed testing procedure: FWER is strongly controlled for any joint distribution of hypothesis test statistics.
- Fixed sequence procedure is a special case of fall back procedure, when w_1 is assigned as 1.
- Tested as weighted Bonferroni procedure if all hypotheses are accepted.
  - Uniformly more powerful than weighted Bonferroni procedure.
- Can only propagate alpha in one direction.
  - Not ‘α-exhaustive.’
  - Can be extended to more powerful alpha-exhaustive approach, i.e. Graph-based multiple test procedures

Graphical Testing Approach



Intuitive procedure using a network graph:

- **Nodes**: Endpoints
- **Edges**: Direction
- **Weights**: How much alpha is propagated from one node to another

Allows for prioritization by adjusting weights accordingly

Spread the wealth: No alpha left on the table!

Key Inputs

- Initial α allocation (typically assigned to primary endpoint)
- General endpoint priority
- Simulated virtual patient data
  - Effect size scenarios covering multiple situations
  - Opportunities to use parametric models, re-sampling of existing patient data, or a combination of both

Scheme Performance Metrics

- **Marginal power**: P(H_X pass) irrespective of other outcomes
- **Mutually-exclusive power**: P(H_X pass ∩ all other $H_{X'}$ not pass)
- These can extend to **combinations** of outcomes meaningful to the analysis

Portfolio Project Workflow

1. Meet with clinical team to construct general testing scheme(s)
2. Generate virtual patient data using effect sizes or other prior information from literature and/or existing data
3. Apply schemes incorporating specific edge weights to obtain pass/fail outcomes for each virtual trial
4. Summarize proportion of “success” for each outcome in all schemes
5. Review results with team
6. Repeat steps above (iterative process)
7. Align on final scheme

Case Study

- Phase III trial with 1 primary endpoint (H1) and 4 key secondary endpoints (H2-H5)
  - **Logical constraint**: H1 must be successful in order to evaluate remaining outcomes
  - **H2** and **H3** important outcomes for a similar clinical domain
    - Larger incremental value for achieving success on at least one than achieving both
  - **H4** less clinically important
  - **H5** considered a “bonus”

Simulation Parameters

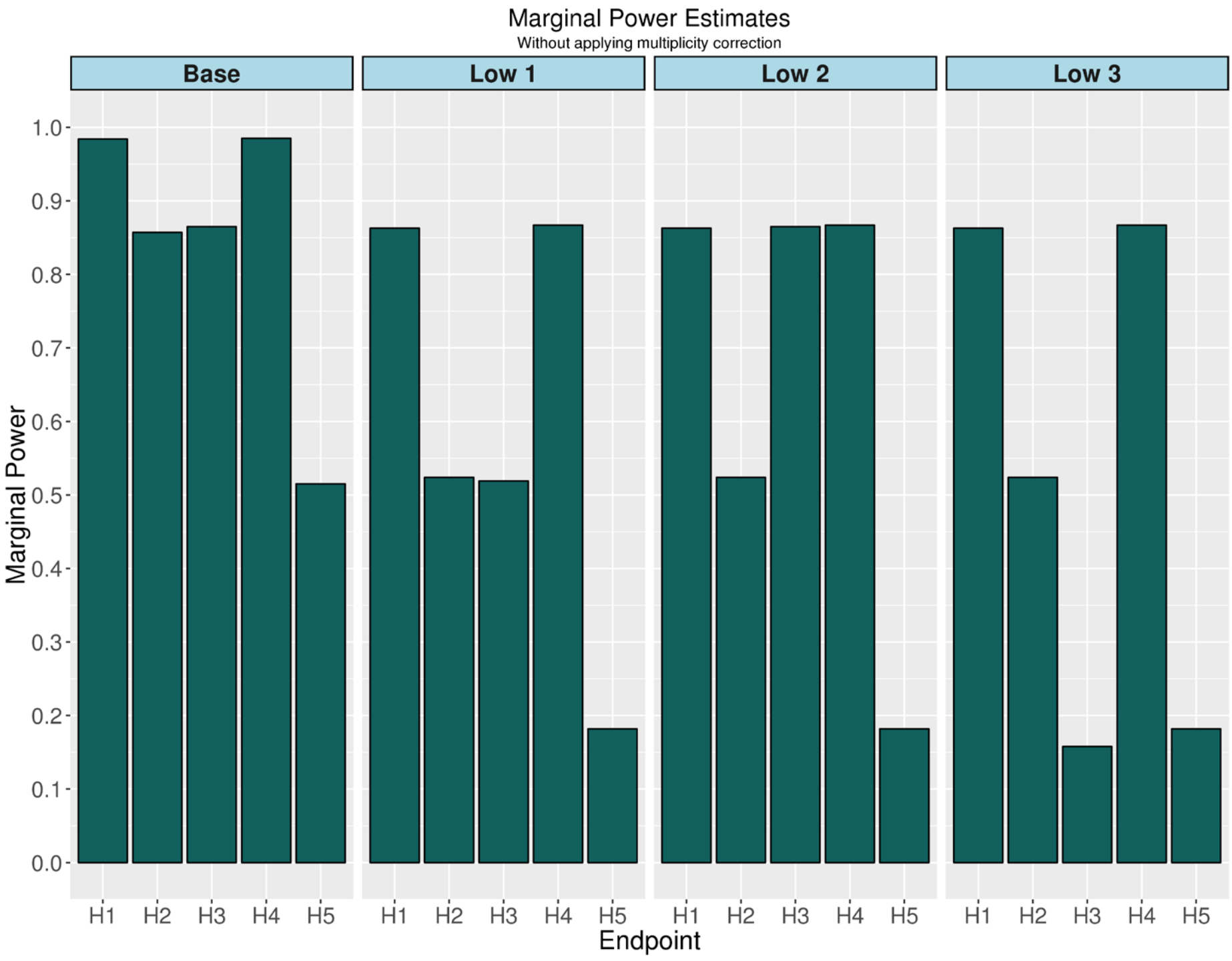
- 1,000 virtual trials with N=800 per trial
- Joint Distribution: $(H_1, H_2, H_3, H_4, H_5) \sim N(\theta, \mathrm{R})$
- Multiple effect size scenarios:

Base Scenario

θ = (4, 3, 3, 4, 2)

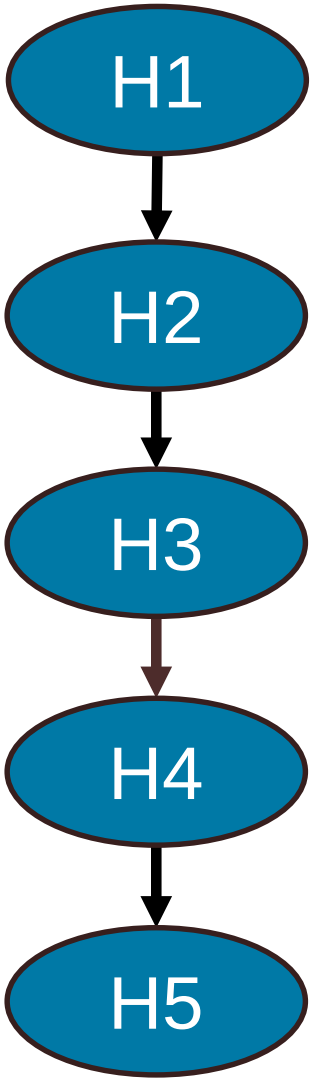
Low Scenario 1	Low Scenario 2	Low Scenario 3
θ = (3, 2, 2, 3, 1)	θ = (3, 2, 3, 3, 1)	θ = (3, 2, 1, 3, 1)

$$\mathbf{R} = \begin{bmatrix} 1 & 0.5 & 0.3 & 0.65 & 0.55 \\ 0.5 & 1 & 0.55 & 0.55 & 0.99 \\ 0.3 & 0.55 & 1 & 0.3 & 0.55 \\ 0.65 & 0.55 & 0.3 & 1 & 0.3 \\ 0.55 & 0.99 & 0.55 & 0.3 & 1 \end{bmatrix}$$

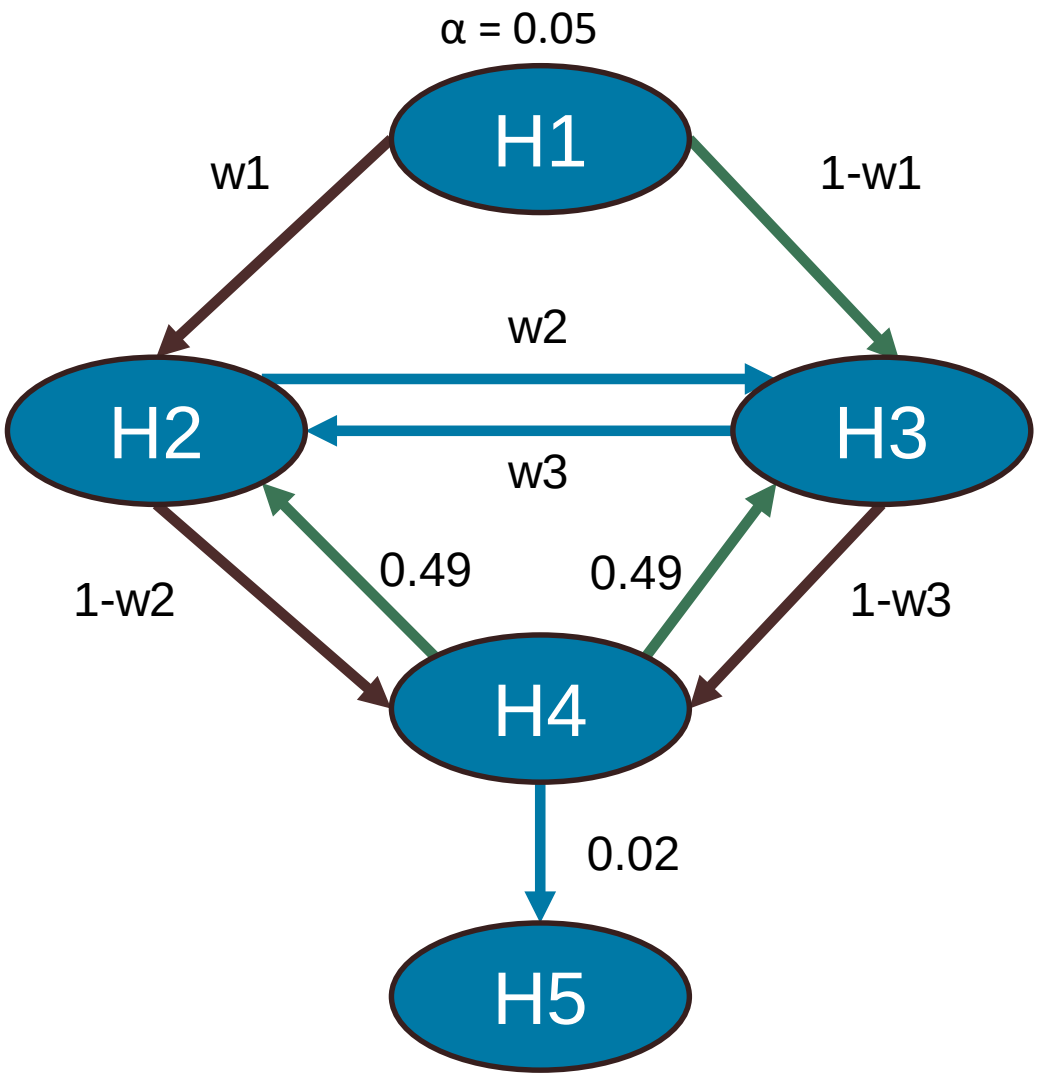

Marginal Power Estimates
Without applying multiplicity correction
Base
Low 1
Low 2
Low 3
Marginal Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
H1
H2
H3
H4
H5
Endpoint

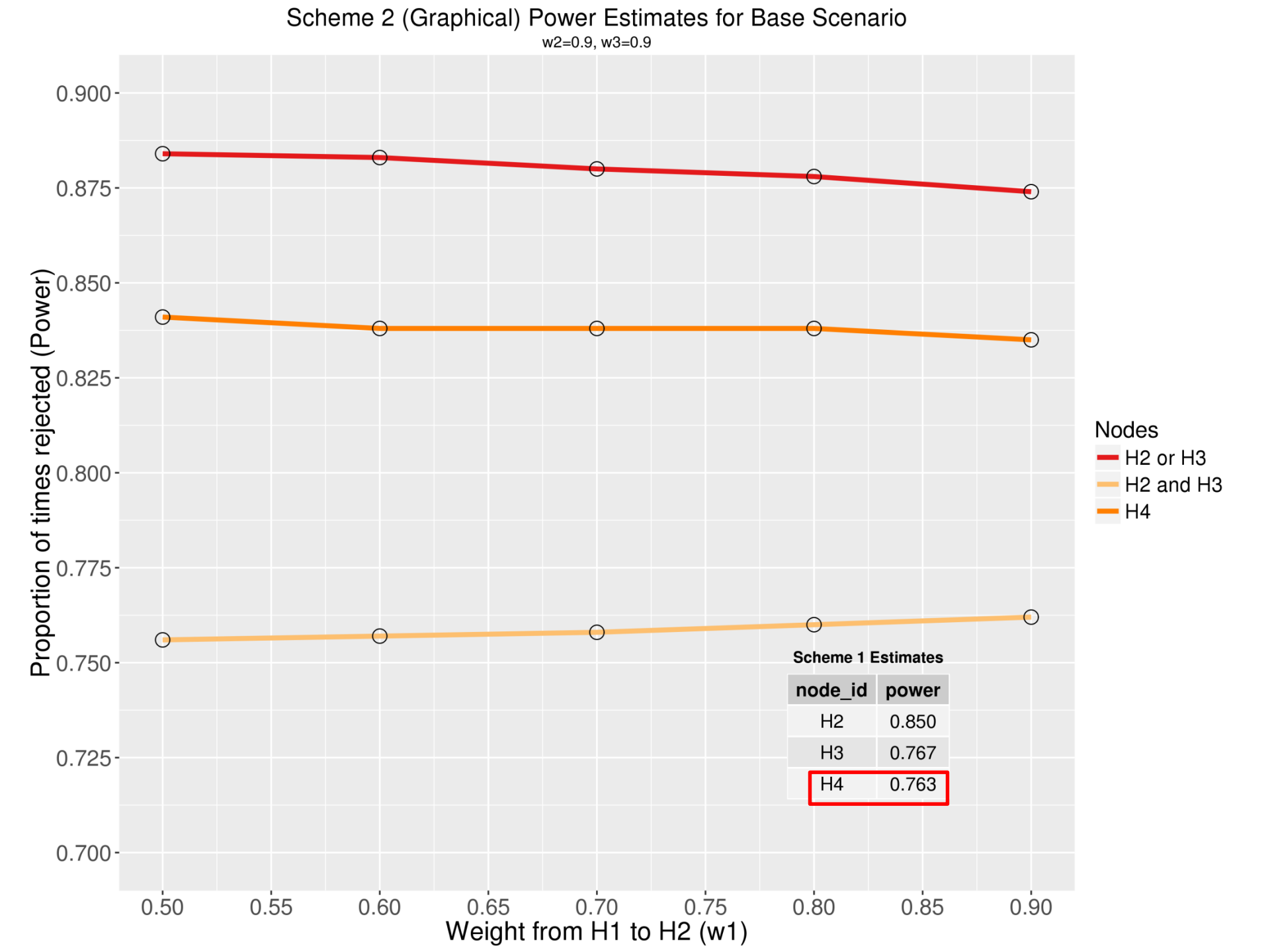
Scheme 1 (Gatekeeping)

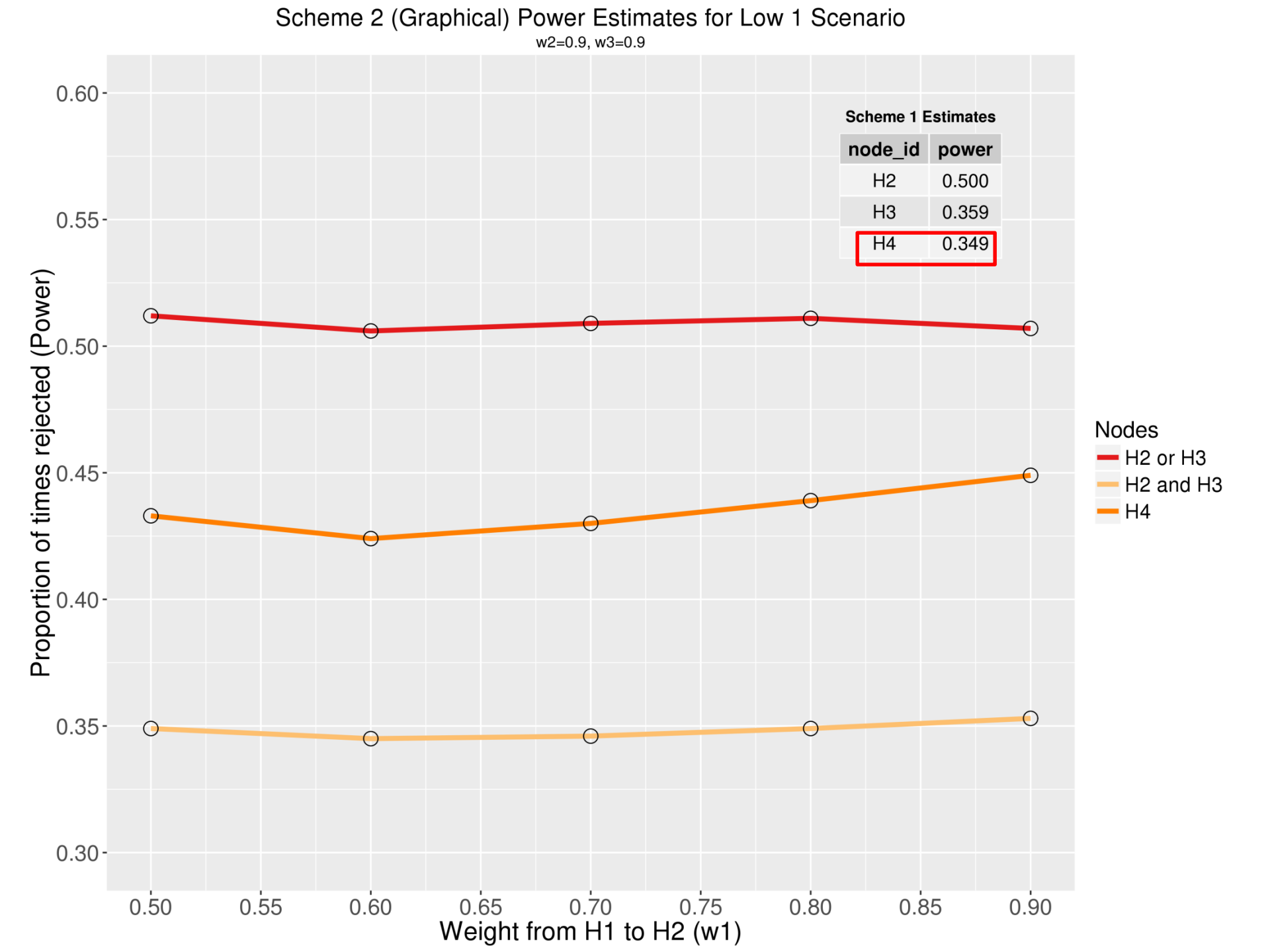
$\alpha = 0.05$

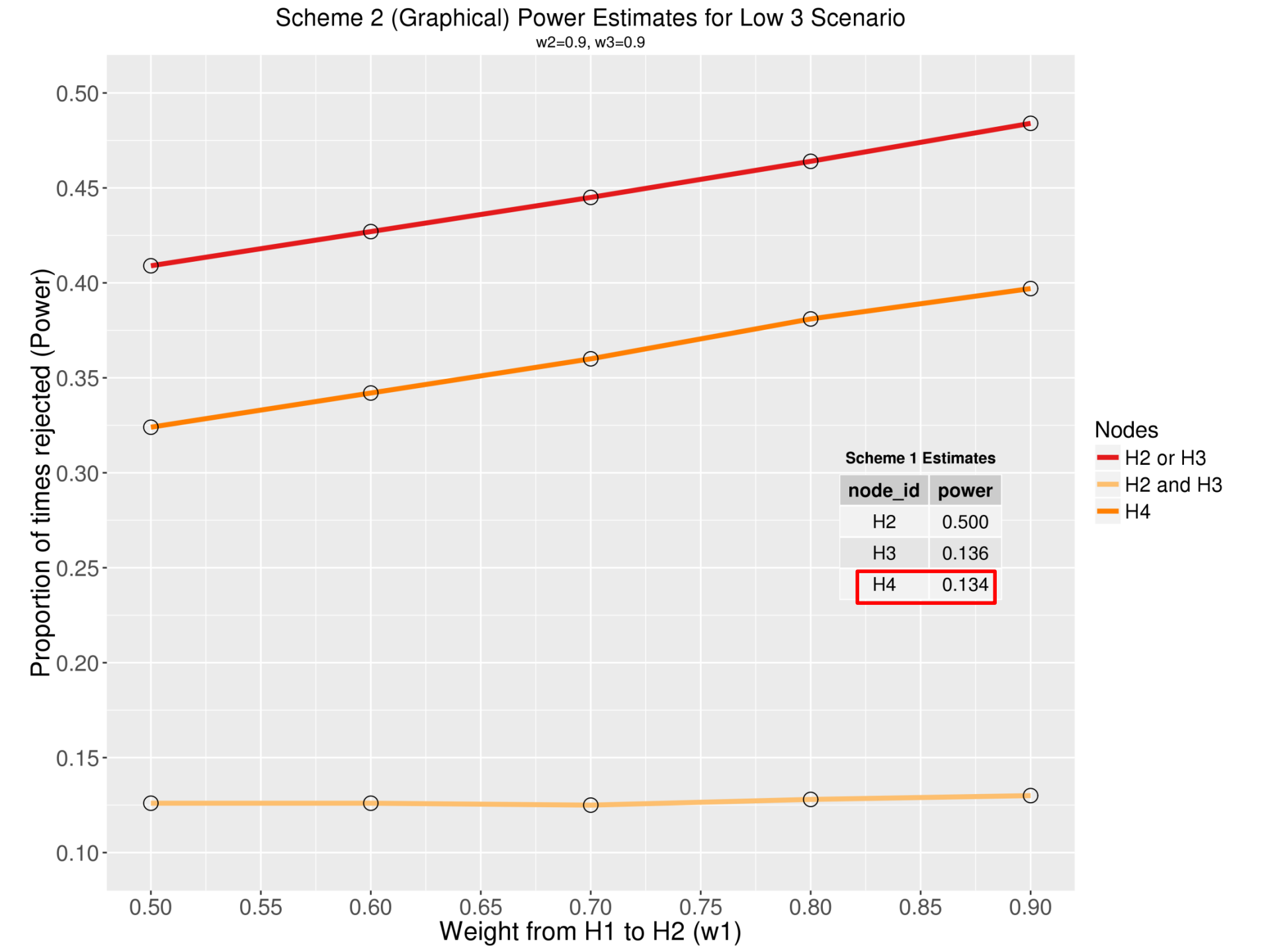


Scheme 2 (Graphical Approach)




Scheme 2 (Graphical) Power Estimates for Base Scenario
w2=0.9, w3=0.9
Proportion of times rejected (Power)
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0.725
0.700
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Weight from H1 to H2 (w1)
Nodes
H2 or H3
H2 and H3
H4
Scheme 1 Estimates
node_id power
H2 0.850
H3 0.767
H4 0.763

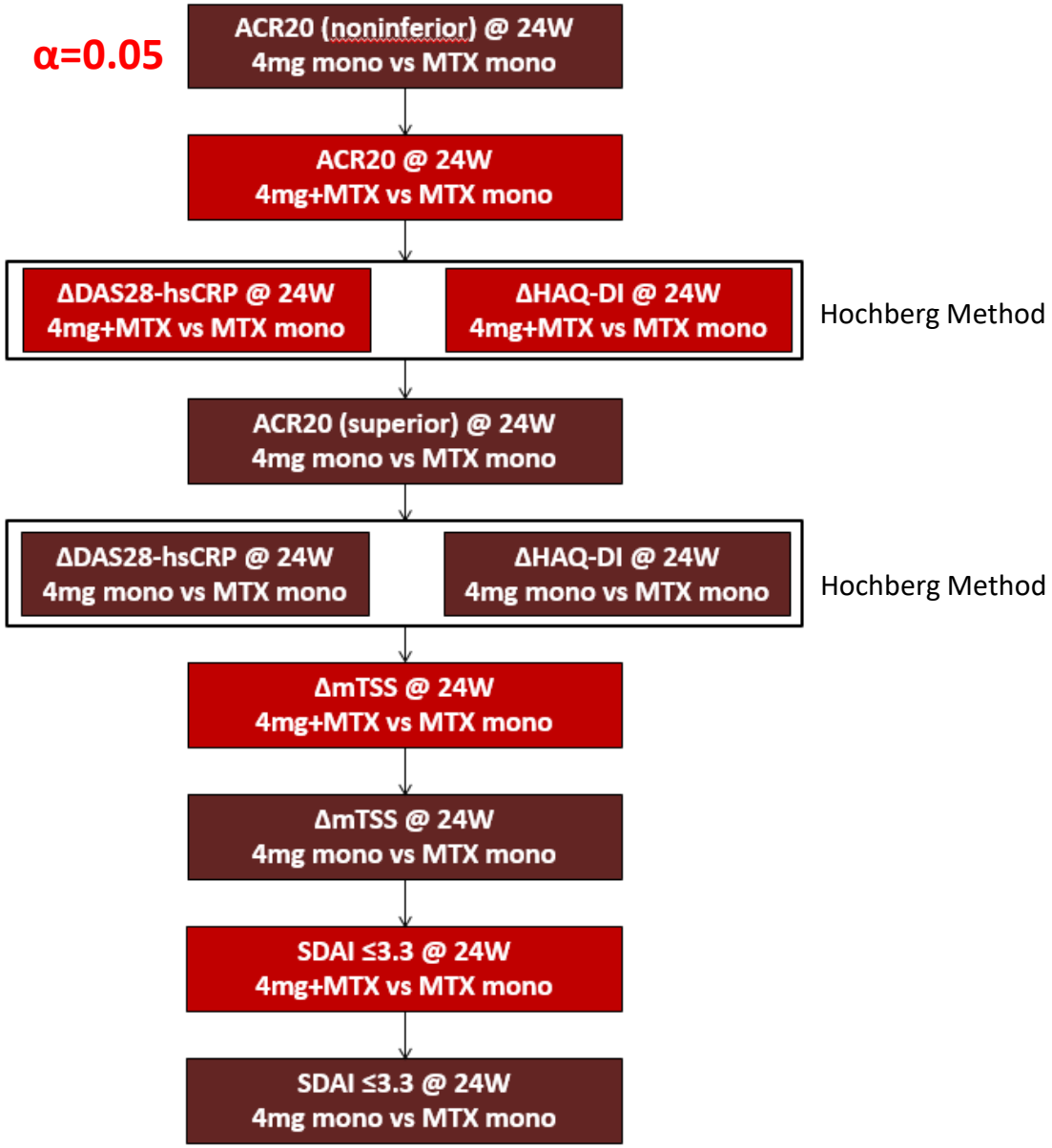

Scheme 2 (Graphical) Power Estimates for Low 1 Scenario
w2=0.9, w3=0.9
Scheme 1 Estimates
node_id power
H2 0.500
H3 0.359
H4 0.349
Proportion of times rejected (Power)
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90
Weight from H1 to H2 (w1)
Nodes
H2 or H3
H2 and H3
H4


Scheme 2 (Graphical) Power Estimates for Low 3 Scenario
w2=0.9, w3=0.9
Proportion of times rejected (Power)
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
Weight from H1 to H2 (w1)
Nodes
H2 or H3
H2 and H3
H4
Scheme 1 Estimates
node_id power
H2 0.500
H3 0.136
H4 0.134

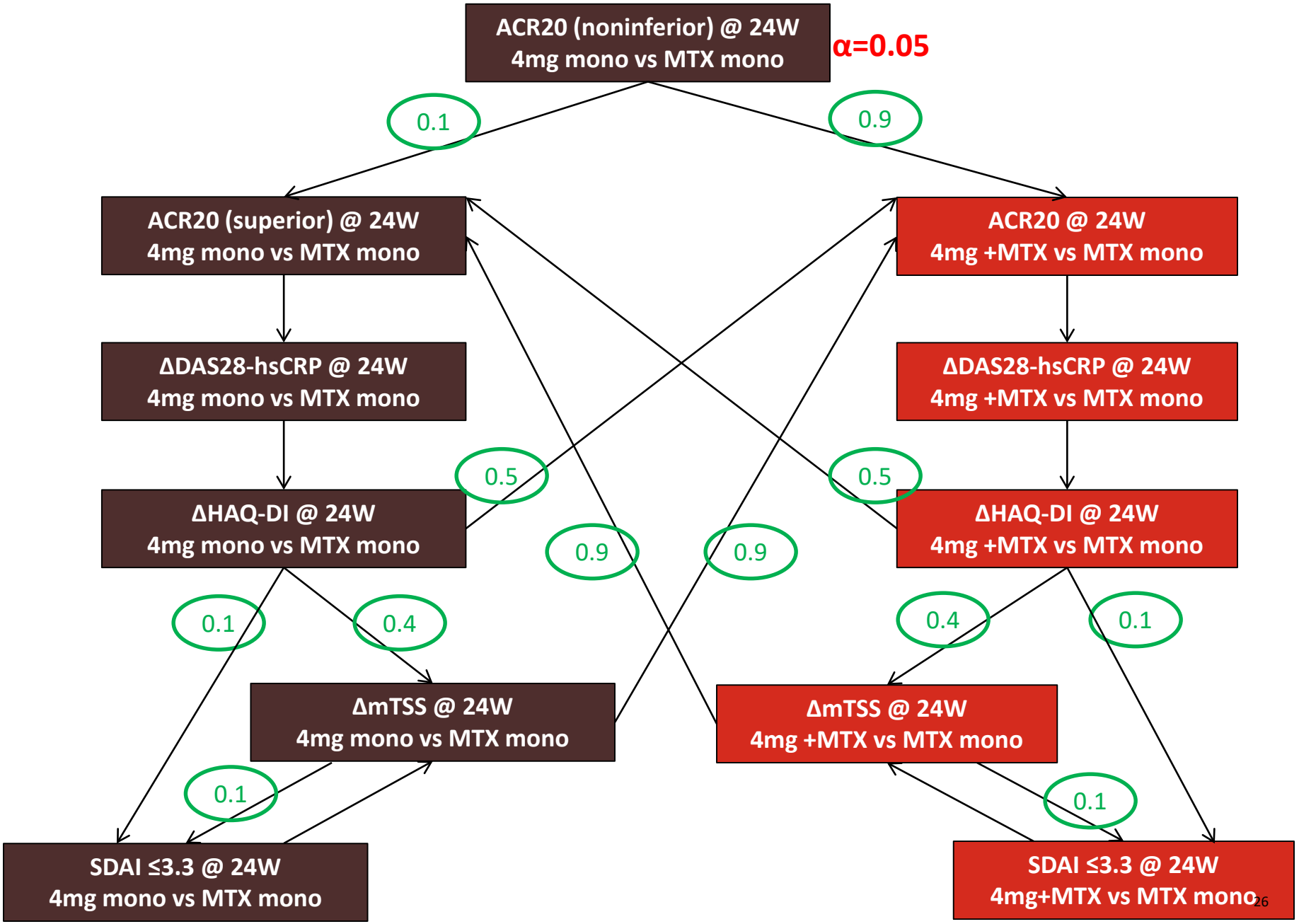
Application to Late Phase Trial

Gatekeeping: Methods Illustration

Original Plan: Stepwise with Some Shared-α Tests

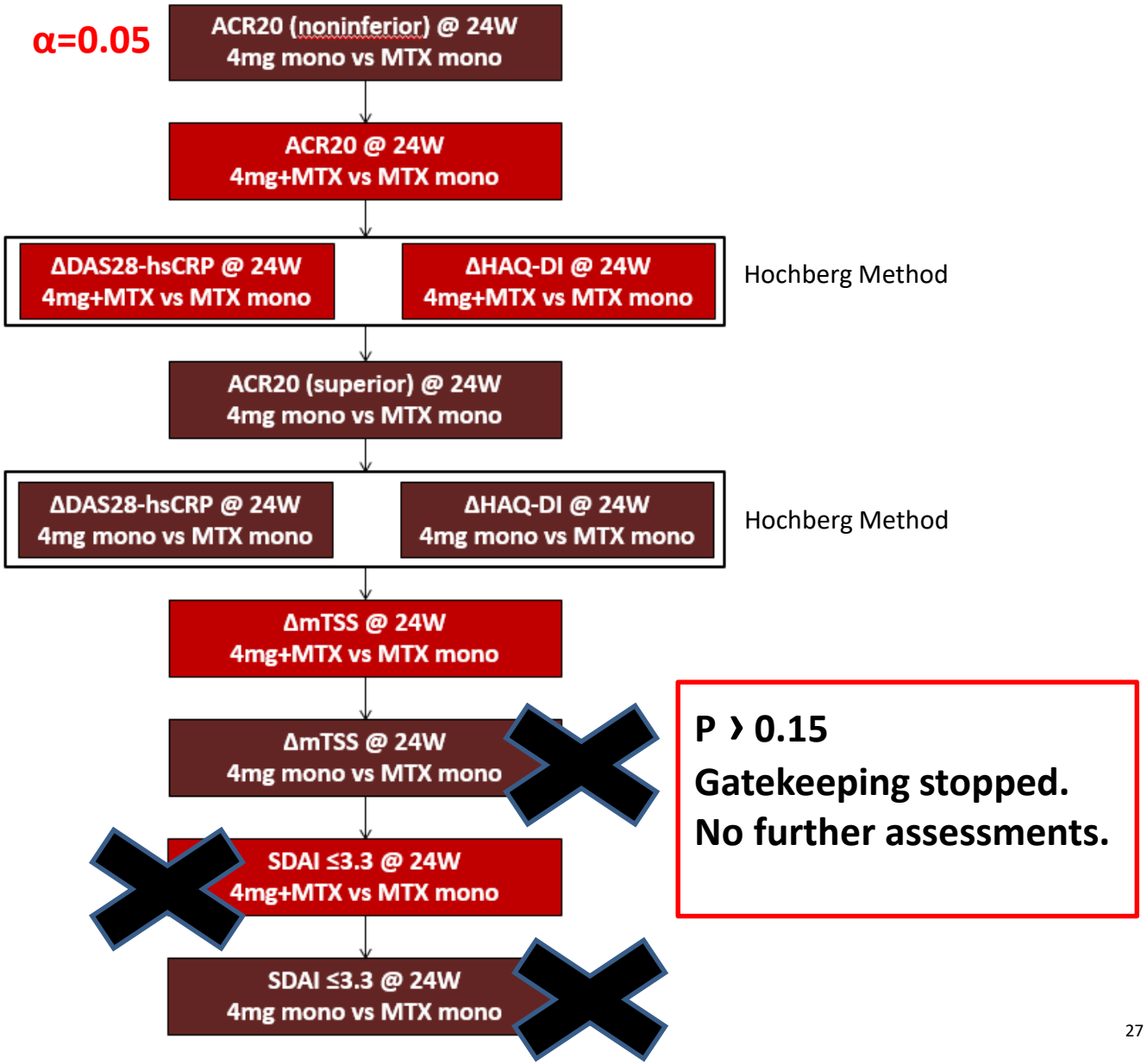


Change to Graphical: Methods Illustration

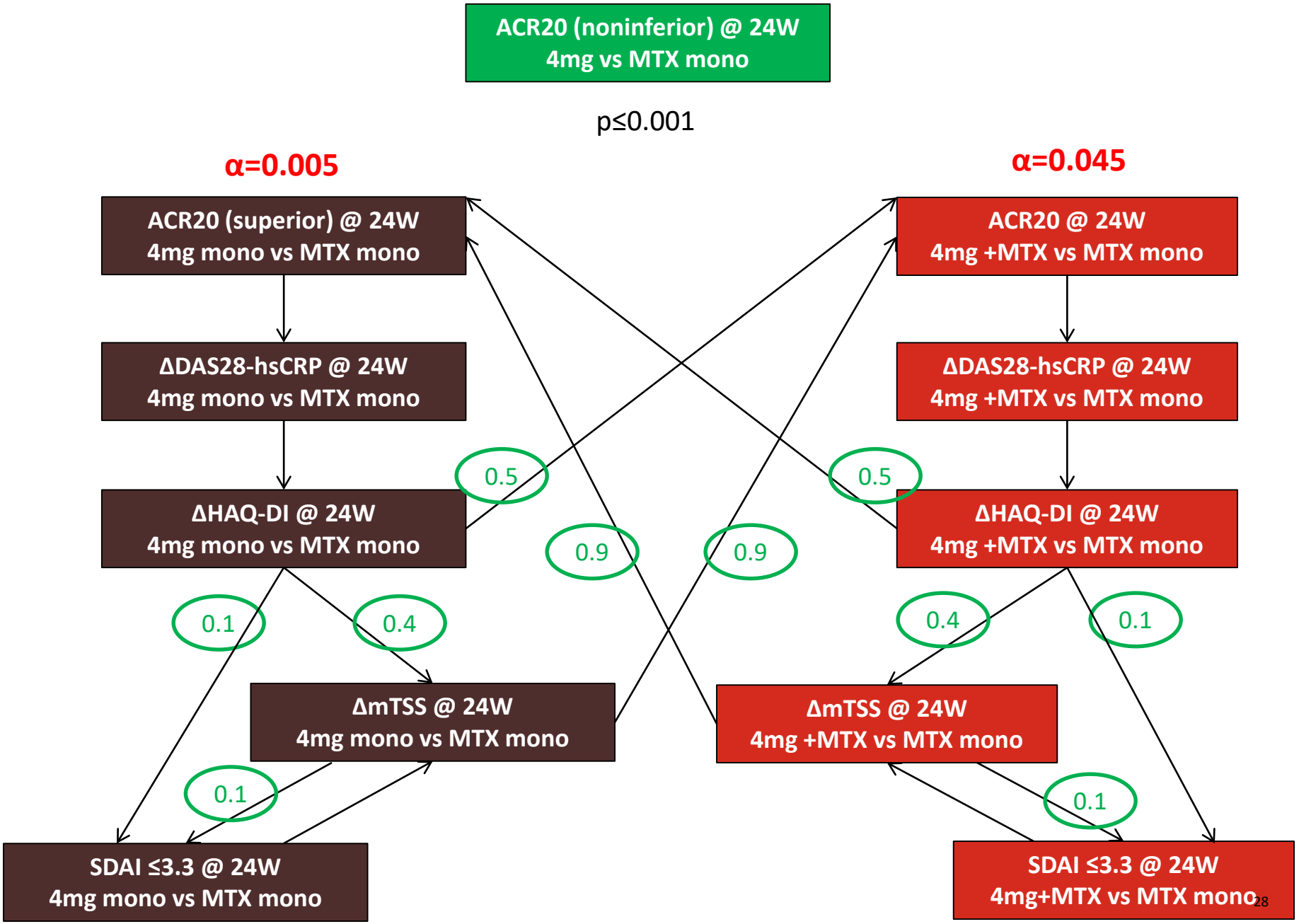


Gatekeeping: Methods Illustration

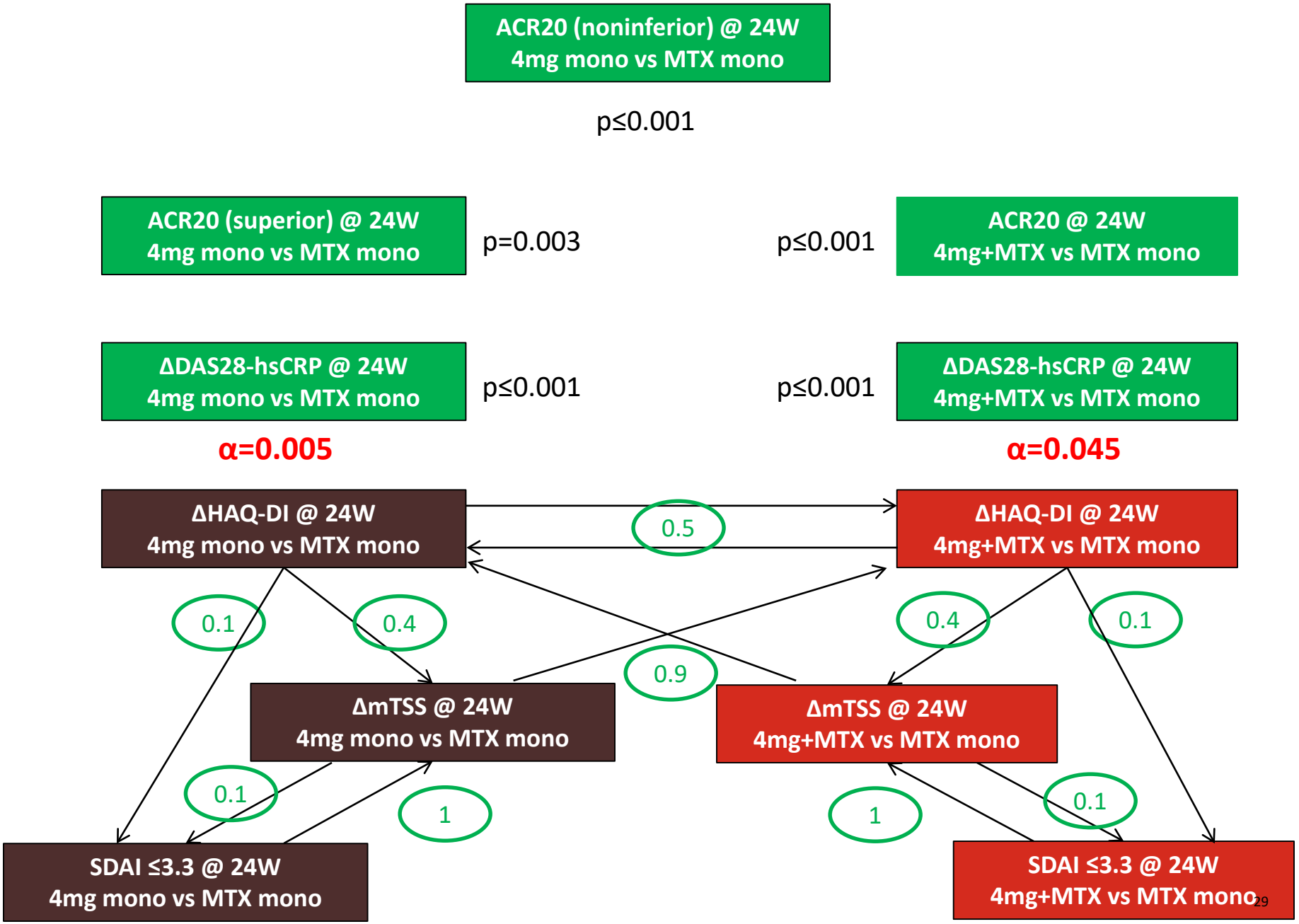
Original Plan: Stepwise with Some Shared-α Tests



Change to Graphical: Methods Illustration



Change to Graphical: Methods Illustration



Change to Graphical: Methods Illustration

ACR20 (noninferior) @ 24W
4mg mono vs MTX mono

p≤0.001

ACR20 (superior) @ 24W
4mg mono vs MTX mono

p=0.003

p≤0.001

ACR20 @ 24W
4mg+MTX vs MTX mono

ΔDAS28-hsCRP @ 24W
4mg mono vs MTX mono

p≤0.001

p≤0.001

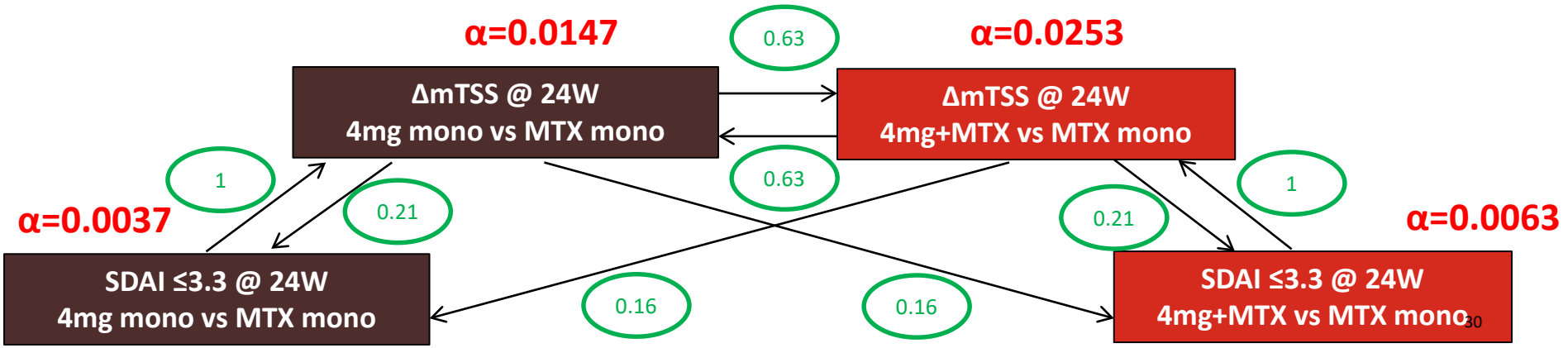
ΔDAS28-hsCRP @ 24W
4mg+MTX vs MTX mono

ΔHAQ-DI @ 24W
4mg mono vs MTX mono

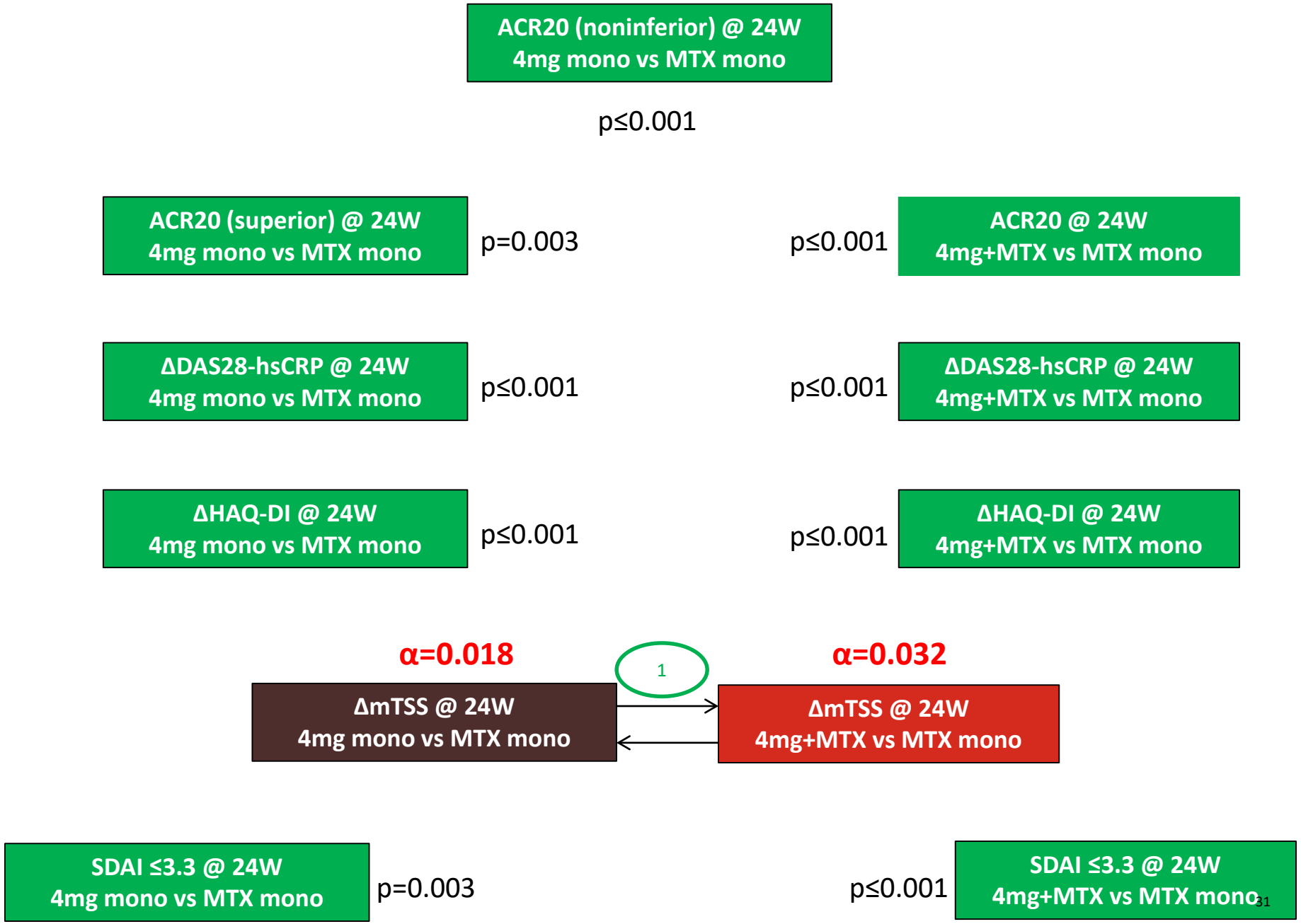
p≤0.001

p≤0.001

ΔHAQ-DI @ 24W
4mg+MTX vs MTX mono



Change to Graphical: Methods Illustration



Change to Graphical: Methods Illustration

ACR20 (noninferior) @ 24W
4mg mono vs MTX mono

p≤0.001

ACR20 (superior) @ 24W
4mg mono vs MTX mono

p=0.003

p≤0.001

ACR20 @ 24W
4mg+MTX vs MTX mono

ΔDAS28-hsCRP @ 24W
4mg mono vs MTX mono

p≤0.001

p≤0.001

ΔDAS28-hsCRP @ 24W
4mg+MTX vs MTX mono

ΔHAQ-DI @ 24W
4mg mono vs MTX mono

p≤0.001

p≤0.001

ΔHAQ-DI @ 24W
4mg+MTX vs MTX mono

α=0.05

p=0.158

ΔmTSS @ 24W
mg mono vs MTX mono

ΔmTSS @ 24W
4mg+MTX vs MTX mono

p=0.026

SDAI ≤3.3 @ 24W
4mg mono vs MTX mono

p=0.003

p≤0.001

SDAI ≤3.3 @ 24W
4mg+MTX vs MTX mono

Summary of Results

Endpoint Category		4 mg vs MTX			4 mg + MTX vs MTX		
	Endpoint	p-value	Gate-keeping	Graph	p-value	Gate-keeping	Graph
Signs and Symptoms							
	ACR20 at Week 24 (N-Inf)	0.001	Sig	Sig	n/a		
	ACR20 at Week 24 (Sup)	0.003	Sig	Sig	0.001	Sig	Sig
	DAS28-hsCRP at Week 24	0.001	Sig	Sig	0.001	Sig	Sig
	HAQ-DI at Week 24	0.001	Sig	Sig	0.001	Sig	Sig
Clinical Remission							
	SDAI ≤ 3.3 at Week 24	0.003	NSig	Sig	0.001	NSig	Sig
Structure							
	mTSS at Week 24	0.158	NSig	NSig	0.026	Sig	Sig

Key Learnings

- Strive to maintain flexibility as well as simplicity
- Awareness of logical constraints early in the process
- Involve cross-functional colleagues (medical, regulator, marketing) throughout all stages
- Isolate key combinations of outcomes to help assess practical advantages and disadvantages

References

- Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Stat Med*. 2014;33(4):693-713.
- Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;28(4):586-604.
- Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biom J*. 2011;53(6):894-913.

References

- Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika.* 1988;75(4):800-802
- Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979;6(2):65–70
- Hommel, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988;75:383-386
- Wiens BL, Dmitrienko A. The Fallback Procedure for Evaluating a Single Family of Hypotheses. *Journal of Biopharmaceutical Statistics* 2005;15(6):929-942
- Dmitrienko. A, Tamhane., A, Liu. L, Wiens. B. A note on tree gatekeeping procedures in clinical trials. *Stat Med. 2008; 06:1-6*