Some key multiplicity questions on primary and secondary endpoints of RCCTs and possible answers

Mohammad F. Huque, Ph.D.
Div of Biometrics IV, Office of Biostatistics
OTS, CDER/FDA

Basel Statistical Society, Basel, Switzerland
March 7, 2011

Disclaimer: This presentation expresses personal views of the presenter and not necessarily of the FDA

Primary and secondary endpoint definitions?

- What should be the regulatory definitions of primary endpoints and secondary endpoints?
- Can a primary endpoint in a trial be called a “key secondary endpoint” or a secondary endpoint because of power considerations?

Primary endpoints (PEs) (…regulatory thoughts)

- These are critical endpoints such that unless there is statistically significant and clinically meaningful evidence of efficacy in one or more of these endpoints for the study treatment, there is (usually) no justification for a claim.
- Regulatory approval of new drugs and biologics rely on statistically significant and clinically meaningful evidence of treatment benefits on one or more primary endpoints of adequate and well-controlled clinical trials.

Secondary endpoints (SEs)

- Not sufficient to support efficacy in the absence of an effect on one or more primary endpoints.
- However, the secondary endpoints can provide additional claims and other important clinical information
- O'Neill, RT (1997): "Secondary endpoint can not be validly analyzed if the primary endpoint does not demonstrate clear statistical significance." *Controlled Clinical Trials* 18: 550-556

"Key" secondary endpoints?

- Primary endpoints (PEs) can form
  - Either a single family
  - Or multiple hierarchical families depending, for example, on their relative importance and power considerations, and the win criteria
- Therefore, there seems no need to call a PE (if more than one) a key secondary

Should there be a separate alpha-control for PEs and separate for SEs?

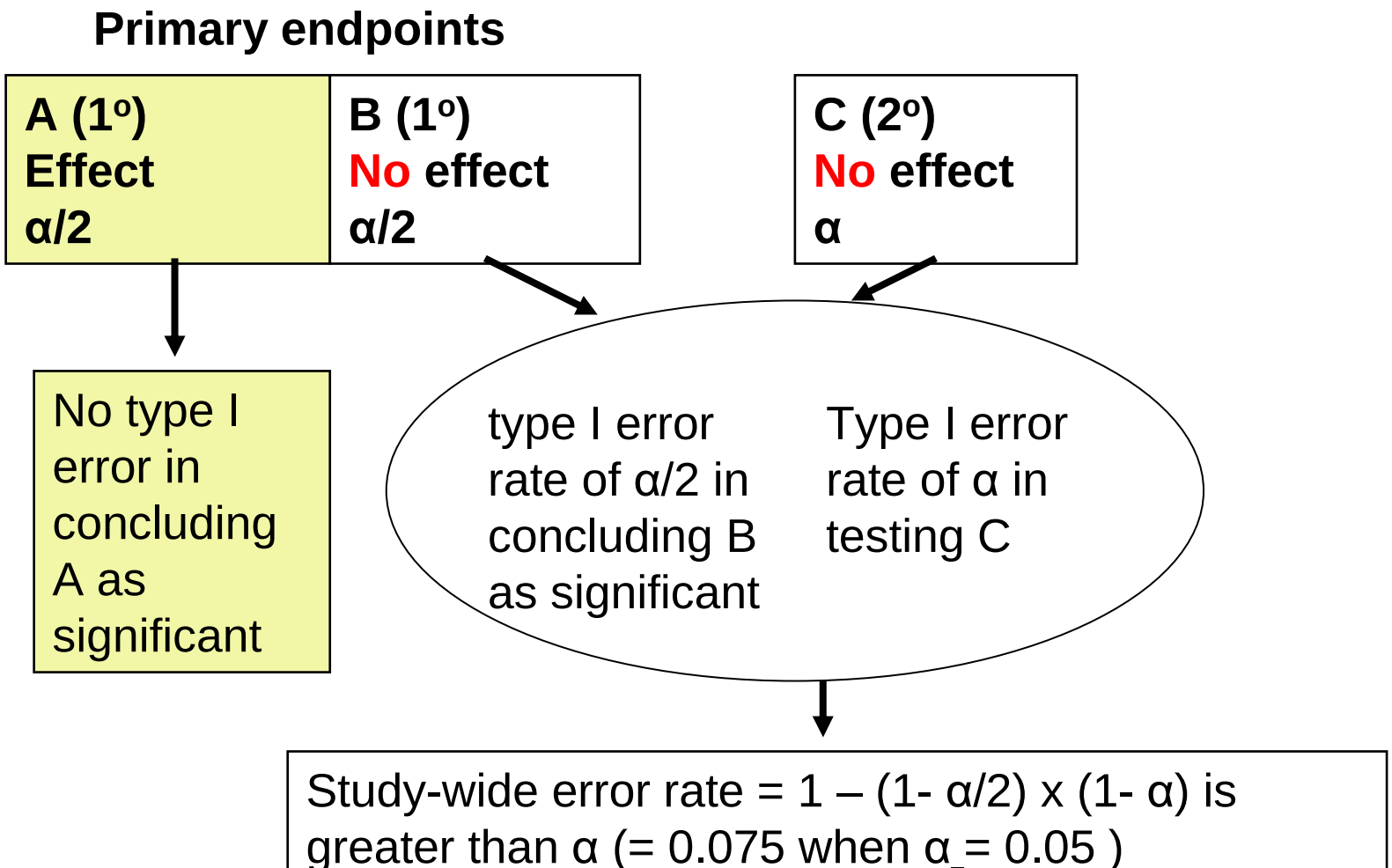
- That is, should there be a separate FWER control for the primary endpoint family (e.g., at the 0.05 level) and a separate FWER control for the secondary endpoint family (also at the same 0.05 level)?

  With the Condition: that the secondary endpoint family is to be tested only after statistically significant and clinically meaningful result in one or more primary endpoints?

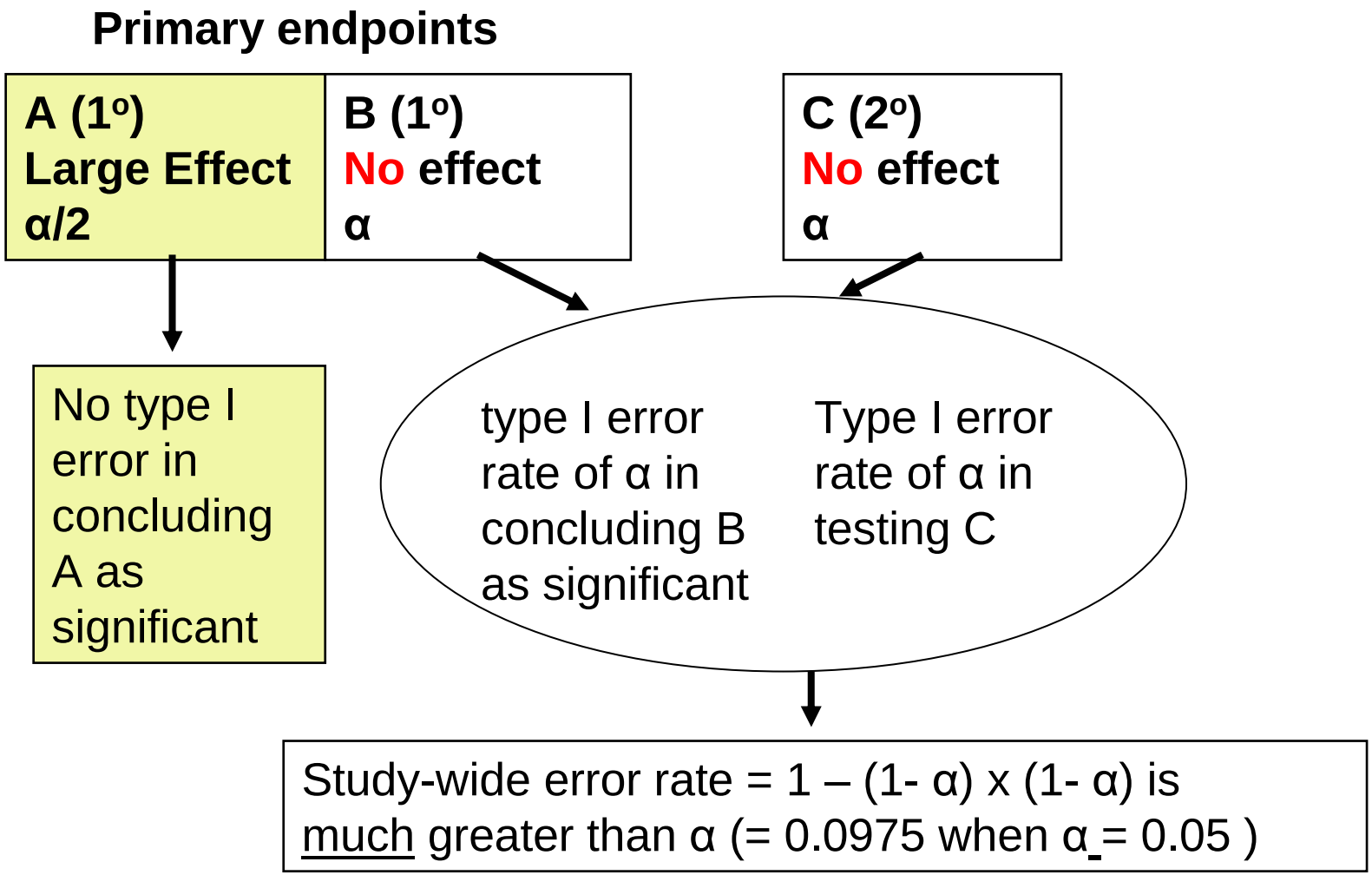
- **Is it a good idea?**

  (Evaluation of the idea in next few slides)

With the Bonferroni test



With the Holm's test



Does the use of α =0.05 for a single study provide convincing evidence of efficacy?

Concerns raised in the literature:

- Berger and Sellke (JASA, 1987)
- Goodman (Ann. Int. Med, 1999)
- Lee and Zelen (Stat. Sc,. 2000)

Yes: with the replication of evidence:

FDA's interpretation of the U.S. Food Drug and Cosmetic Act 1962: At least two "adequate and well-controlled" trials, each convincing on its own, can establish effectiveness

Pr. of finding a statistically significant result at the 0.05 significance level in a second study for various observed *P*-values in the first study

Observed *P*-value (1st study)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
Conditional power method	0.50	0.61	0.73	0.80	0.86	0.91	0.94
Bayesian predictive probability method	0.50	0.58	0.67	0.73	0.77	0.83	0.86

Huque, Alosh and et al., 2009

Bayesian Error Rate* of Falsely Concluding Treatment Efficacy

$$\alpha_B = \Pr(T=-\mid E=+) = \frac{\Pr(T=-)\Pr(E=+\mid T=-)}{\Pr(T=-)\Pr(E=+\mid T=-)+\Pr(T=+)\Pr(E=+\mid T=+)}$$

$$= \frac{\pi_0\alpha}{\pi_0\alpha+(1-\pi_0)(1-\beta)} = \frac{\alpha}{\alpha+(1-\beta)/\psi_0}.$$

Power 1- β = 0.80			
$1-\pi_0$	Value of α_B when α=0.025	Adj. of α when α_B =0.025	Sample size increase by fraction $\hat{f}$
0.30	0.06796	0.00879	1.31766
0.35	0.05485	0.01105	1.24853
0.40	0.04478	0.01368	1.18373
0.45	0.03679	0.01678	1.12147
0.50	0.03030	0.02051	1.06036

* Lee and Zelen (2000)

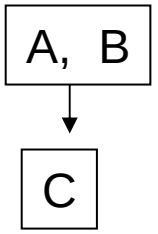
Which analysis methods to use for primary endpoint families?

- Methods should be valid for independent as well as for correlated endpoints and for any joint distribution of test statistics or *p*-values
- Examples:
  - Bonferroni
  - Holm's ?
  - PAAS (for positively correlated endpoints)
  - Sequential testing method
  - Bonferroni based gatekeeping procedures (Dmitrienko et al. and others)
  - (Sequentially rejective) graphical approach (Bretz et al., 2009)
  - Other methods (e.g., truncated Holm's, fallback, etc.)

  Note: Hochberg procedure generally not recommended: Known to fail FWER control in the strong sense for some correlation structures

Improper use of Holm's method can inflate the FWER?

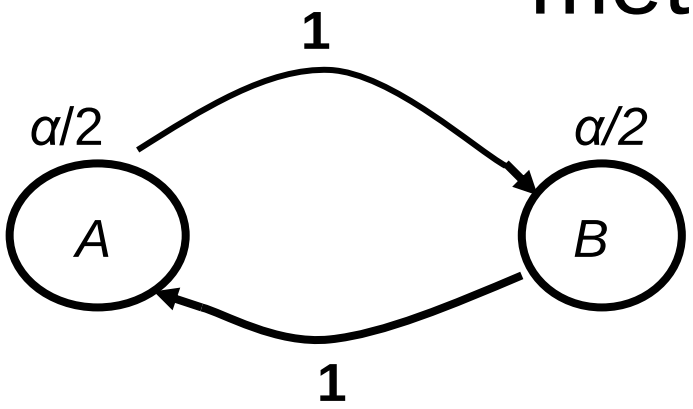
- Test strategy:
  1) Test endpoints in F1 = {A, B } by the Holm's method (i.e. test the smallest of the two p-values p(1) at level α/2 and if successful then test the larger of the two p-values p(2) at level α)
  2) If one of the two endpoints in F1 is successful and the other one is not, then test the endpoint C at level α/2 (This will inflate the FWER I error rate)

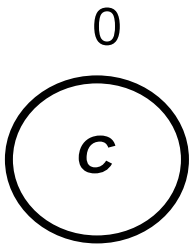


(test A and B by the Holm's method alpha=0.05)

(test C at level 0.025, if A is significant but B is not, or B is significant but A is not)

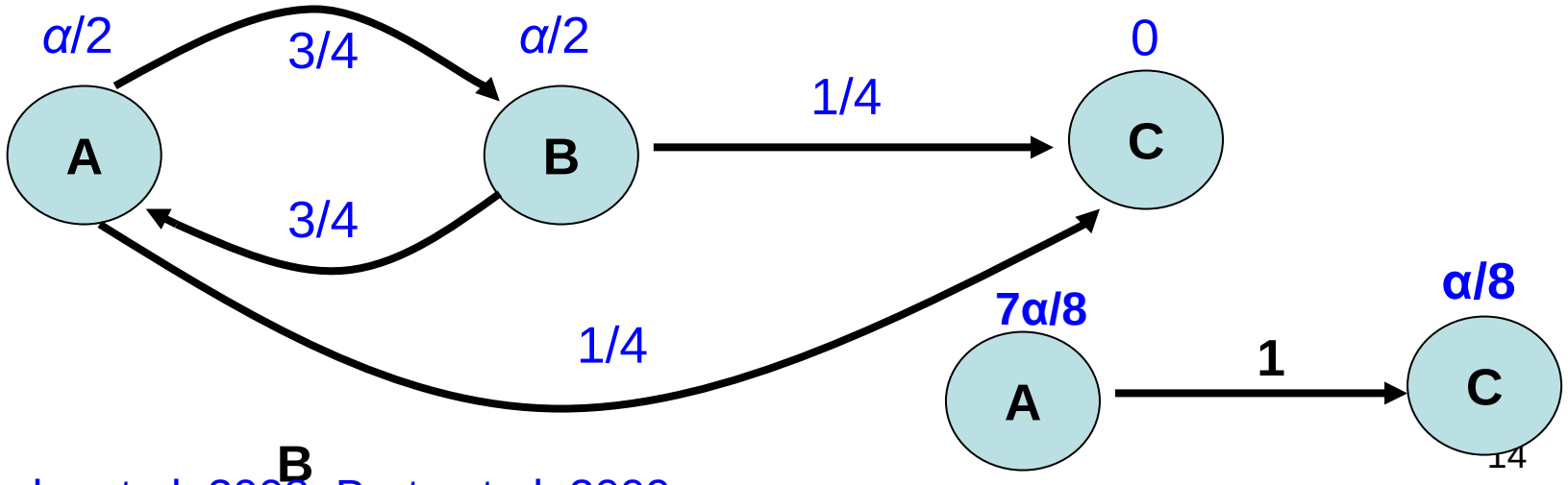
α-exhaustive nature of the Holm's method





C can be tested only when both *A* and *B* are successful

Truncation of the Holm's method



Dmitrienko et al. 2008; Bretz et al. 2009

After B is successful

Can resampling/bootstrap methods be used for PEs?

- A popular a resampling based step-down procedure:

  Step 1: Rejects $H_{(1)}$ associated with $p_{(1)}$ if

  $\Pr\{ \min(P_1, P_2, \ldots, P_m) \leq p_{(1)} \} \leq \alpha$

  Step $j = 2, \ldots, m$: Rejects $H_{(j)}$ associated with $p_{(i)}$ if

  $\Pr\{ \min(P_j, P_{j+1}, \ldots, P_m) \leq p_{(j)} \} \leq \alpha$

  Step m: Rejects $H_{(m)}$ associated with $p_{(m)}$ if

  $\Pr\{ P_m \leq p_{(m)} \} \leq \alpha$

  ✓ Stop further testing when 1st time condition not met

- Above probabilities calculated from the resampling distributions of the minimum *P*-value test statistics

Concerns regarding the previous resampling method for primary comparisons of a confirmatory trials

- Appealing and useful for correlated endpoints
  - However, results approximate: some cases may require simulations to validate the results
- Computation can be difficult (e.g., for time-to-event endpoints)
- FWER (Strong) control:
  - May not be OK in some situations?
  - Westfall & Troendle proof - uses the assumption of subset pivotality condition
- Ref:
  - Westfall and Troendle (2008; Biometrical J.; *multiple testing with minimal assumptions*);
  - Huang et al. (2006; Bioinformatics; *permute or not to permute*)

2.3 Computational issues: the MaxT test and subset pivotality condition

While power is the main concern for choice of a test statistic, expediency becomes important when m is large. There are $O(2^m)$ intersection hypotheses H_I, and if m is large, it is computationally impossible to test every single H_I. However, the computational burden can be eased dramatically if one is willing to

A: test each hypotheses H_I using a "Max" statistic $\max_{i \in I} T_i$, possibly sacrificing power, and

B: assume a model that implies "subset pivotality", which states that the distributions of $\max_{i \in I} T_i \mid H_I$ and $\max_{i \in I} T_i \mid H_{\{1,\ldots,m\}}$ are identical, for all $I \subset \{1, \ldots, m\}$.

If A and B are adopted, one need only test m hypotheses corresponding to the ordered t_i rather than all 2^m intersections; further, resampling can be done simultaneously under a global null $H_{\{1,\ldots,m\}}$, rather than separately for each intersection. Note that "Max" subsumes "Min", where the test statistic is $-T_i$; the Min P test is commonly used (Westfall and Young, 1993).

To illustrate, suppose the observed test statistics are $t_1 \geq \cdots \geq t_m$, corresponding to hypotheses $H_1, \ldots, H_m$ (ordered in this way without loss of generality), and that larger t_i suggest alternative hy-

Westfall and Troendle (2008, Biometrical J.)

Does the FWER (strong) control for the previous resampling method hold for the following scenarios?

1. In a 3-arm trial for as single PE test for:
   - H1: high dose vs. PL
   - H2: low dose vs. PL
   - H3: combined (high + low doses) vs. PL
2. In a 2-arm trial (treatment vs. PL) test for 3 lipid endpoints E1 = reduction in LDL, E2 = in crease in HDL; E3 = increase in the ratio: HDL/LDL
3. Note: In clinical trials hypotheses tested may have dependencies

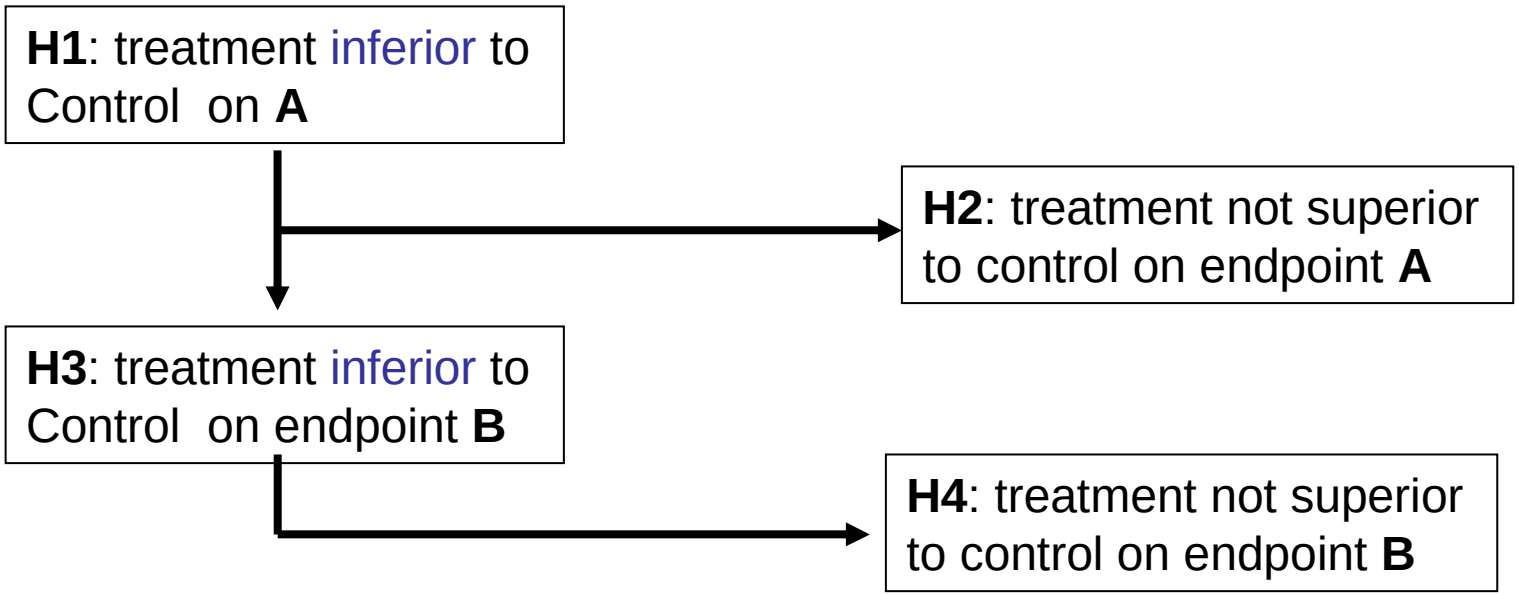
Multidimensional multiplicity problems common in clinical trials

- Dimensions:

  Multiple PEs x multiple doses x NI/Sup tests – leads to many tests

- Problems come with logical constraints: E.g.,
  - No superiority test on a PE unless the endpoint is successful on NI test
  - No low dose test on a secondary endpoint unless test successful on a corresponding PE
- Methods (trying):
  - Tree structured gatekeeping (Dmitrienko et al., 2007)
  - Graphical method (Bretz et al., 2009)
  - Use of Bretz et al. (2009) graphical method with branches
  - Examples to follow:

Consider the following Sup/NI tests on endpoints A and B: Is there a multiplicity issue?



Test strategy (hierarchical):

If H1 is rejected then test for H2 and H3, and if H3 is rejected then test for H4

Will there be FWER control at level 0.05 if each test is at level 0.05?

- Some thinks: Yes
- Reason usually given is:
  - NI tests follow a sequential order and that the test for Sup for each endpoint follows simultaneously after non-inferiority test by the same the same 2-sided 95% confidence interval that establishes NI

A simple proof of inflation if each test at 0.025

- Consider: D1 = treat. diff. (for A), D2 = treat. diff. (for B), and events:

  $A_N = D_1 - 1.96*SE(D_1) > -\delta_1$ (Reject H1)

  $A_s = D_1 - 1.96*SE(D_1) > 0$ (Reject H2)

  $B_N = D_2 - 1.96*SE(D_2) > -\delta_2$ (Reject H3)

  $B_s = D_2 - 1.96*SE(D_2) > 0$ (Reject H4)

- Suppose: Treatment is NI to control on both A and B, but is not superior to control on A and not superior to control on B. Sample size is sufficiently large so that H1 and H3 are both rejected)
- Let: $E1 = A_N A_s B_N (B_s)^c$; $E2 = A_N A_s B_N B_s$; $E3 = A_N (A_s)^c B_N B_s$

  Now, $E2 \cup E3 = (A_N B_N B_s) = B_s$ (because B_s is a subset of B_N which is a subset of A_N

- Therefore: $FWER = Pr(B_s) + Pr(A_s B_N (B_s)^c) = 0.025 + \varepsilon > 0.025$

Solution by the Gatekeeping method

- **Define families of hypotheses:**

  F1 = { *H*1 }, F2 ={ *H*2, *H*3 }, F3 ={ *H*4}.

- **Test strategy:**

  1. Test first *H*1 in F1 at the level α (e.g., $\alpha = 0.05$).
  2. Once the result for *H*1 is significant at level α, testing proceeds to the hypotheses *H*2 and *H*3 in F2 with the alpha that was not lost within the F1 family, which in this case is α

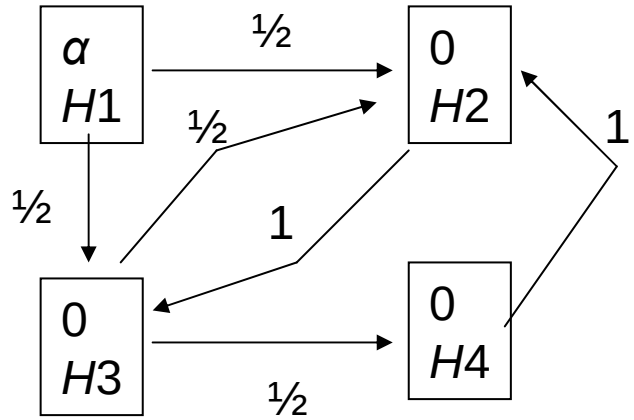
     The test of *H*2 and *H*3 in F2 can be by the Bonferroni test. That is, one would test *H*2 and *H*3, each at level $\alpha/2$.

     If both *H*2 and *H*3 are rejected then a total of $\alpha/2 + \alpha/2 = \alpha$ transfers to F3, and if only *H*3 is rejected then only alpha of $\alpha/2$ transfers to F3.

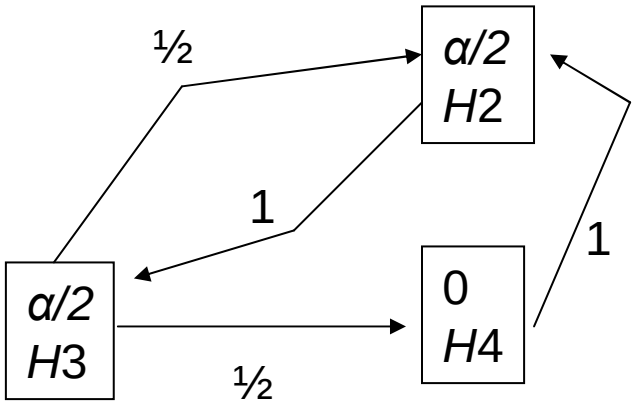
     However, if *H*3 in F2 is not rejected (even if *H*2 is rejected), then there is no passing of alpha from F2 to F3. Consequently, there are no further tests because of the logical restriction.

Solution by the graphical approach

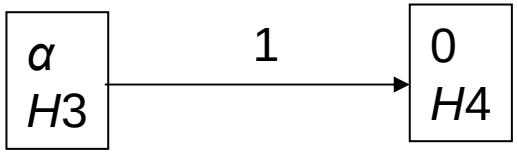
(a) Original graph



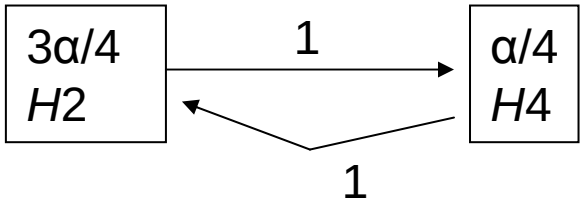
(b) Graph after rejecting *H*1



(c) Graph after rejecting *H*2 in (b)

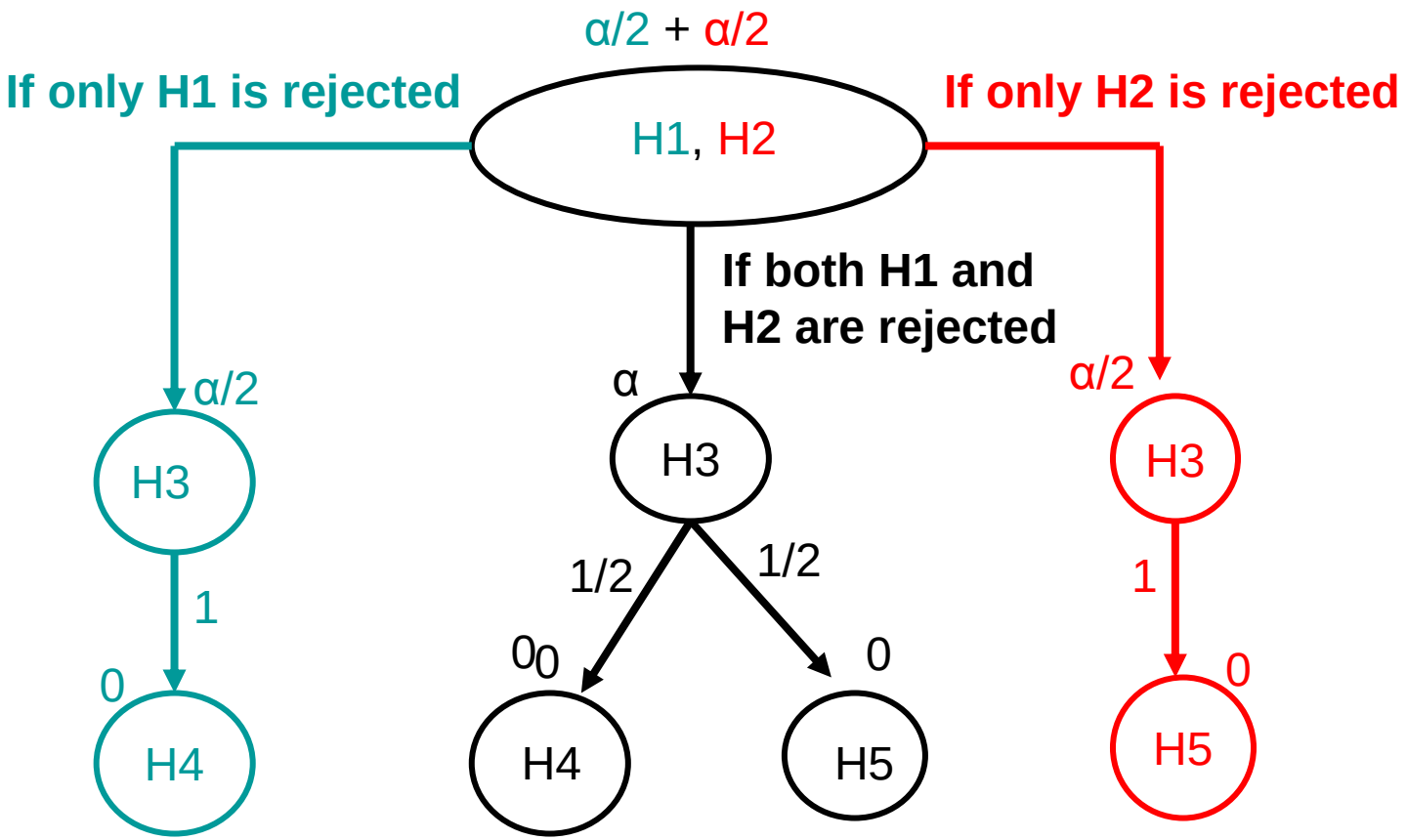


(d) Graph after rejecting *H*3 in (b)



(Ref: Bretz et al., SIM 2009)

Use of graphical method w. branches?



H1: High dose test; H2: Low dose test; H3: combined dose test
H4 and H5 secondary endpoint tests

Questions on composite PE

- **Composite PE**: an endpoint that combines the most relevant clinical endpoints for the drug and the disease under study into a single combined endpoint that is clinically meaningful
- Clinical endpoints combined are called the component endpoints (or simply “components”) and are supposed to be
  - Sensitive to treatment effects, clinically relevant, chosen *a priori*, easy to interpret, and free of errors of ascertainment, etc
  - Endpoint ascertainment methods must capture accurately both the occurrence and non-occurrence of the component events.

Use of composite endpoint as a PE: widespread in clinical trials

- **SCOUT** (NEJM 2010; 363: 905-917): ((nonfatal myocardial infarction, nonfatal stroke, resuscitation after cardiac arrest, or cardiovascular death)
- **ACCORD** (NEJM 2008; 358: 2545-2559): (nonfatal myocardial infarction, nonfatal stroke, or death from cardiovascular causes)
- **ADVANCE** (NEJM 2008; 358: 2560-2572): [composites of major macrovascular events (death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke) and major microvascular events (new or worsening nephropathy or retinopathy)]
- **LIFE** (*Lancet* 2002;359: 995-1003): (death, myocardial infarction, or stroke)
- **TIME** (*Lancet* 2001;358: 951-7): (death, non-fatal myocardial infarction, or hospital admission for acute coronary syndrome)
- **NORDIL** (Lancet 2000; 359-365): (non-fatal stroke, myocardial infarction, or other cardiovascular death)
- **INSIGHT** (*Lancet* 2000;356: 366-372): (cardiovascular death, myocardial infarction, heart failure, or stroke)
- **HOPE** (*Lancet* 2000;355(9200): 253-9): (myocardial infarction, stroke, or cardiovascular death)
- **ACE** (*Lancet* 1999;353: 2179-84): (stroke, MI or death)
- **PRAISE** (NEJM 1996;335: 1107-14): (all cause mortality or hospitalization)
- **CAPRIE** (*Lancet* 1996;348: 1329-39): (ischemic stroke, myocardial infarction, or vascular death)

Composite endpoint topics

Listed by Joachim Röhmel (2004)

- Rationale for composite endpoint
- Types of composite endpoints (Addressed Chi, 2005)
- Analysis approaches
- ✓ Weighing of components
- Power of different procedures
- Influence on the composite of components that are not influenced by the treatment
- ✓ Heterogeneity across components
- Composite endpoint for non-inferiority trials
- ✓ Consistency of the direction of effects

When do multiplicity issues arise in composite endpoint trials?

- **No multiplicity issue**
  - if the trial has a single composite primary endpoint and no intention to claim for treatment efficacy for its components
  - Component outcomes are displayed only in the descriptive sense

  Multiplicity issue
  - Success sought for the total patient population for win either for the composite or for some of its clinically relevant components or for a clinically meaningful sub-composite (multiple ways to win)
  - Success sought for win either for the total patient population or for a targeted subgroup of patients, either for the composite or for some of its clinically relevant components (multiple ways to win)

Issue of heterogeneity across components:

(mortality trending in the wrong direction)

- Example (hypothetical):
  - 2-arm trial: treatment A versus control, composite PE = (death, MI and revascularization)
- Results:
  - Composite endpoint, significant in favor of treatment A: p=0.008
  - Death: in favor of control: p=0.07 (OR =1.80)
  - MI: no difference: p=0.9 (OR = 0.98)
  - Revascularization: highly significant in favor of treatment A: p=0.0001 (OR =0.34)
- Comment:
  - The composite PE seems to give an inflated notion of benefit of treatment A.
  - Clinically relevant component went in the opposite direction.
  - Dilemma: Is this signal of harm by chance or real?

The usual questions:

- How one can design such a trial that would not cause such a dilemma?
- What would be a multiple testing strategy for this new design?

  (Following are some ideas – next 3 slides)

1. Assign clinical utility weights

- E.g., death weighted as 0.7, MI as 0.2 and revascularization as 0.1,
- Accept the composite endpoint result if it is still significant at the 0.05 level with these weights.

Comments

- Idea clinically attractive and simple to apply.
- However, there would in general be disagreement among clinicians about the actual weights.
- This difficulty can possibly be solved through a consensus building conference of disease area experts, or by surveying experts.
- This weighting approach also raises the statistical issue of power when down weighting the most frequent component, e.g., the revascularization component in the above trial?

2. Non-inferiority/superiority approach (Röhmel, 2006).

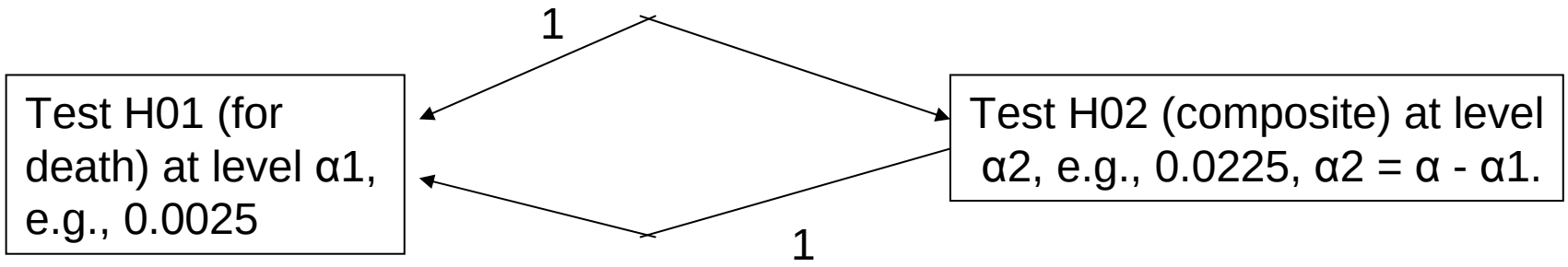
- Set a margin for acceptable inferiority for critical components, e.g., the upper CI for the mortality odds-ratio not to exceed 1.2

Comments

- The trial can be jointly powered with a superiority test for the composite and a non-inferiority test for a critical component such as death.
- The sample size to satisfy the non-inferiority test may not be all that large when the true treatment effect for this test is slightly on the positive side

3. "Save-a-little-alpha" approach (fallback test strategy)

Apply the fallback method with a "loop-back" strategy (Bretz et. al., 2009) with 1-sided tests



1. If H02 is rejected, then test H01 at the full significance level of 0.025 and accept the result for H02 if the 1-sided p for death < α* (e.g., α* = 0.50 or 0.55 to satisfy consistency of direction of effect)
2. If H02 is not rejected then test H01 at level α1
3. One can also start the test on the left side

A heterogeneity situation with un-interpretable results

Table I. Structure of data on death and hospitalization (hypothetical data).

	Treatment	
	Active	Control
Died but never hospitalized during follow-up (%)	15	5
Hospitalized and died during follow-up (%)	5	15
Hospitalized, alive at the end of follow-up (%)	20	20
None of the above (%)	60	60

Source: Lubsen et al. (*Stat in Med* 2002; 21: 2959-2970)

Adjusted $p < 0.03$ (hospitalization endpoint)

Such examples necessitates following the "principle of full disclosure"

- Requires displaying outcomes of the composite endpoint and all its components in a manner that a meaningful interpretation of the results of the composite and its components can be made
- Such displays are done in multiple ways for proper understating of patterns of outcomes and how they are distributed in the treated and the control groups

Issues when the mortality or a sub-composite of "hard" components is of interest

- Consider (a hypothetical) 2-arm trial in type 2 diabetic patients that compares a new treatment to placebo
  - Primary endpoint ***c*** = composite (all cause mortality, non-fatal MI, non-fatal stroke, acute coronary syndrome, endovascular or surgical intervention in the coronary or leg arteries, and amputation of a leg).
- This composite PE contains more than a few components. May have difficulty in showing treatment benefit because of lack of sensitivity to treatment effects in some components.
  - Trial, as a fallback, considers an alterative primary endpoint, a sub-composite ***s*** = (all-cause mortality, non-fatal MI and non-fatal stroke).
  - Note: this sub-composite can be the single mortality component

Results at the completion of the trial

- Results
  - Endpoint ***C***: 2-sided p = 0.085 (favoring treatment)
  - Endpoint ***S***: 2-side p = 0.0195 (favoring treatment)
- Comments:
  - The trial would be considered a failed trial if all alpha of 0.05 was spent on ***C*** and nothing was saved for ***S***.
  - The trial would also be considered as a failed trial if one had designed this trial with the fallback tests with the division of the total alpha as (0.04, 0.01).
  - However, p-value (***S***) = 0.0195 in favor of the treatment can be interpreted as a robust result because there is a trend towards effectiveness on ***C*** with p-value (***C***) = 0.085. (4A method)

The #4A method for such a trial

(adaptive alpha allocation approach)

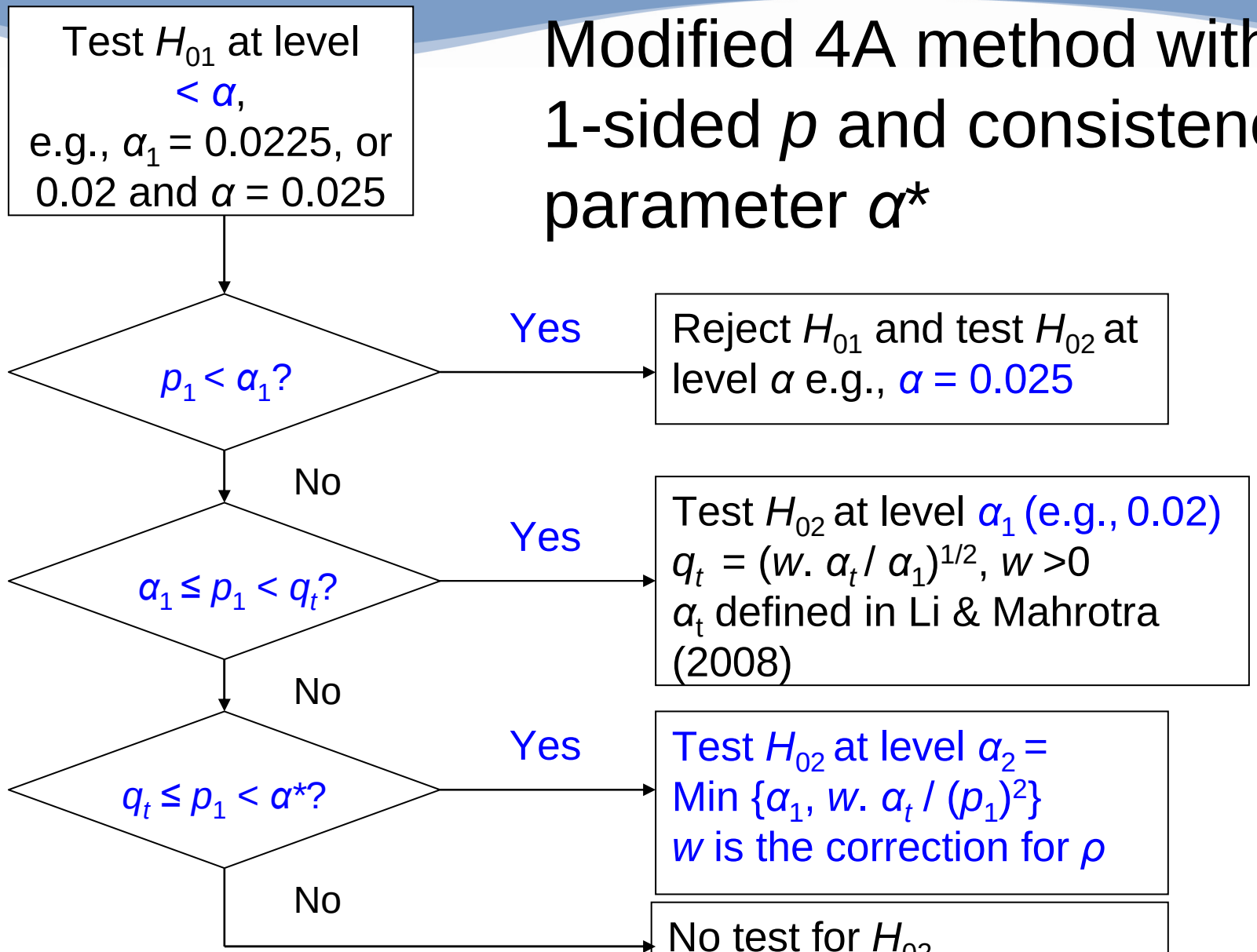
p1 = p-value for endpoint (*C*); p2 = p-value for endpoint (*S*)

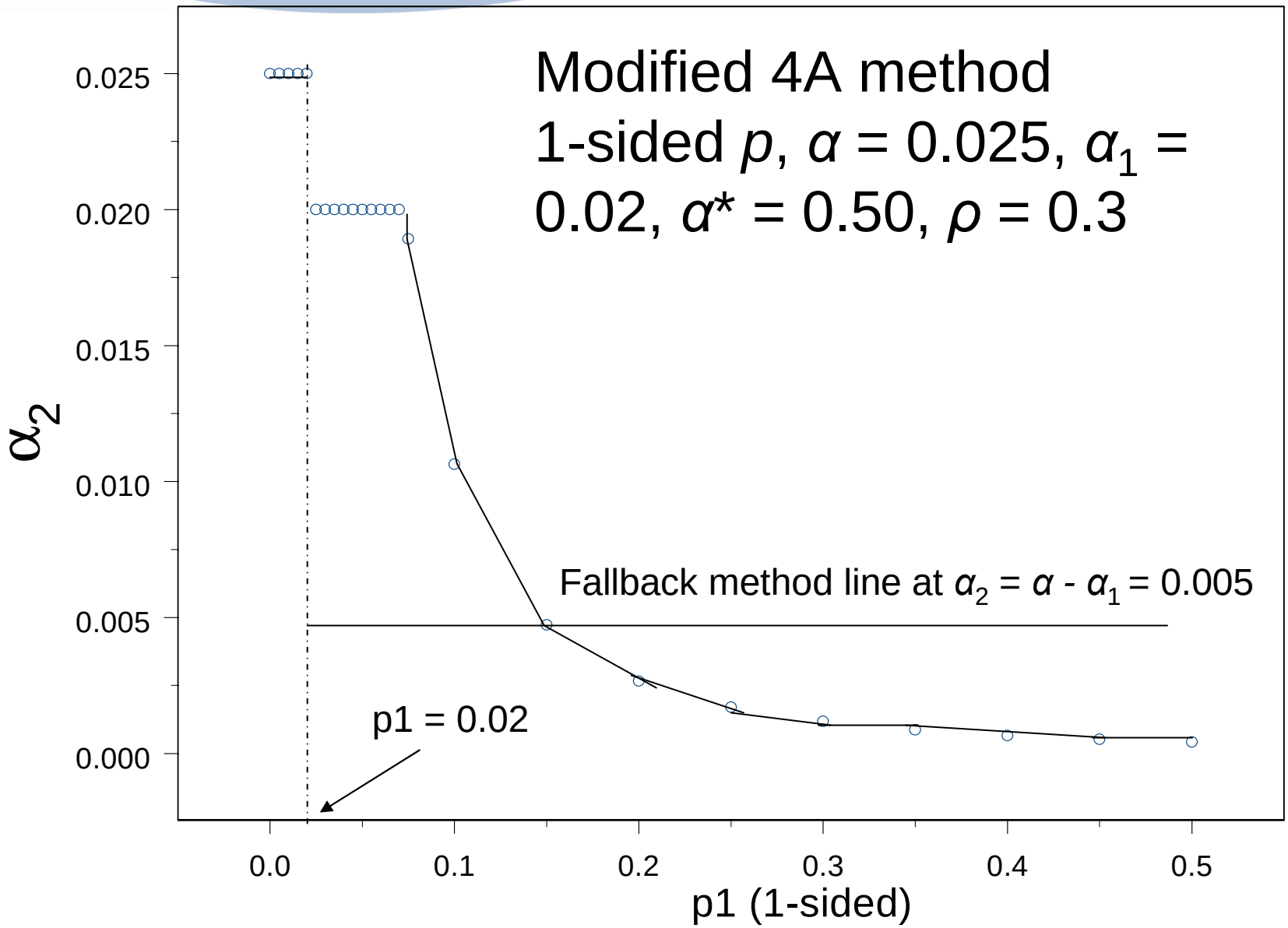
- **The 4A method**
  - Split alpha ($\alpha 1$, $\alpha - \alpha 1$)
  - E.g., (0.04, 0.01)
- If p1 < 0.04, then test p2 at level 0.05
- If p1 ≥ 0.04, then test p2 at level $\alpha 2$ (adaptive):
  - (a) $\alpha 2$ in the interval [0.04, 0.01) for small values of p1 but ≥ 0.04
  - (b) $\alpha 2 \leq 0.01$ for large values of p1

- **The fallback method**
  - Split alpha ($\alpha 1$, $\alpha - \alpha 1$)
  - E.g., (0.04, 0.01)
- If p1 < 0.04, then test p2 at level 0.05
- If p1 ≥ 0.04, then test p2 at level $\alpha 2 = 0.01$

#Reference: Li & Mehrotra, SIM 2008

Modified 4A method with 1-sided *p* and consistency parameter α^*




Modified 4A method
1-sided p, α = 0.025, α1 = 0.02, α* = 0.50, ρ = 0.3
Fallback method line at α2 = α - α1 = 0.005
p1 = 0.02
α2
0.025
0.020
0.015
0.010
0.005
0.000
0.0
0.1
0.2
0.3
0.4
0.5
p1 (1-sided)

The method of #CAS (consistency assured strategy)

- If $p_1 < \alpha_1$ (where $\alpha_1 < \alpha$), then consider the first endpoint as successful and test the second endpoint at the full significance level of α.
- If p_1 falls in the interval $\alpha_1 \leq p_1 < \alpha$ and at the same time $p_2 < \alpha$, then consider both endpoints as successful.
- But, if $\alpha \leq p_1 < \alpha^*$, then test the second endpoint at level γ_2, where $\gamma_2 \leq \alpha$.
- Finally, if $p_1 \geq \alpha^*$ then there is no test for the second endpoint.

(# Huque & Alosh (2010, JBS: to appear)

4A, modified 4A, CAS other similar methods - caveats

- The adaptive significance levels for the second endpoint for application purposes at present are for the cases
  - Endpoints are either statistically independent or the test statistics of the endpoints jointly follow a normal probability model.

    Therefore, the trial needs to be sufficiently large so that that the joint normal probability model can be assumed for the test statistics.
  - The significance level of the second endpoint test, besides depending on the assigned alpha of the first endpoint test and its observed *p*-value, also depends on the correlation between the test statistics.

    Therefore, as the trial may not have an accurate knowledge about the value of this correlation for the patient population of the trial, this significance level should be chosen for the most conservative value of correlation.
- The robustness properties of these methods for situations when the joint probability model of the test statistics deviates from the joint normal probability model has not yet been studied

Conclusions for the trial results

p1 (composite) = 0.085; p2 (sub-composite) = 0.0195

Method	Conclusion
Fixed Sequence	p2 not significant if all alpha = 0.05 spent on the composite endpoint test
Fallback with alpha split (0.04, 0.01)	p2 not significant p2 > 0.01
4A and Modified 4A	p2 < α_2 (significant) Significant
CAS	Significant

When multiplicity adjustments are not necessary in a trial?

1. When the trial specifies a single primary or single composite endpoint for a claim of treatment efficacy
2. All specified primary endpoints need to show clinically relevant treatment benefits.
    - No type I error rate inflation concern, but concern about the type II error rate.
3. Primary endpoints are hierarchically ordered and are tested in a fixed-sequence with each test at the same significance level of α (e.g., $\alpha = 0.05$)
    - If the earlier endpoints in the sequence are under powered, the procedure is likely to stop early and miss the opportunity to evaluate treatment effects for latter potentially useful endpoints.

Multiple analyses for the ITT data set (for the same endpoint and the method)

- Irregularities are common in the intention-to-treat (ITT) data set because of:
  - Some patients may drop early
  - Some may fail protocol criteria
  - Some may not take medications as prescribed
  - Some may take concomitant medications
- Usual Dilemma: How to deal with these irregularities?
- As the true endpoint measurements for these cases are unknown, there is usually concern about bias in the result. Therefore, multiple analyses are done for same endpoint on varying the assumptions about these unknown measurements
- As the purpose of these analyses is to investigate the extent of bias, there is no multiplicity adjustment.

Analyses of the same endpoint data by alternative methods

- Analysis of the same endpoint by alternative methods, in addition to the analysis by the pre-specified method, e.g.,
  - analysis of the same time-to-event endpoint by log-rank test and by the generalized Wilcoxon test
  - analysis of variance on excluding/including certain design factors.
  - analysis by the parametric and non-parametric methods
- Technically, one can adjust for these multiple analyses if they were pre-specified.

  However, this is rarely done, as the purpose of these analyses is usually to demonstrate that the results found are robust and hold regardless of different methods applied

Other situations

- Correction for bias: imbalance in certain key risk factors (pre-specification needed)
- Performing a less conservative after a conservative analysis (e.g., ITT analysis) is significant:
  - for better estimate of the size of the treatment effect and the statistical strength
- Descriptive analyses: E.g., for interpreting the result of an analysis of a primary or a secondary endpoint.
  - E.g., After the result for a continuous endpoint is significant showing the results by response categories
  - E.g., Forest plot for a visual demonstration of consistency of results by baseline risk factor or by center and region (caution: some results may go in wrong direction by chance)

Concluding Remarks

- PEs vs. SEs differ in concept and purpose
  - ✓ Efficacy of a treatment is derived on demonstrating clinically meaningful and statistically significant results in one or more primary endpoints that satisfies a pre-defined clinical win scenario.
  - ✓ Secondary endpoints alone are not suitable for this special purpose.
- Multiplicity in efficacy analyses arises <u>when multiple ways to win for efficacy</u>
  - ✓ Causes inflation of the type I error rate requiring statistical adjustments for its control
  - ✓ Many useful statistical approaches to handle this

Concluding Remarks

- Multiplicity adjustment approaches:
  - Necessary to use methods that control FWER control in "strong" sense for making "specific" claims of treatment benefits.
  - Is the strategy of separate FWER control for the family of secondary endpoints reasonable after clinically meaningful and statistically significant treatment efficacy already concluded based on primary endpoints? It has issues such as inflation of the study-wise error rate
  - For primary endpoint families: use methods that are valid for independent as well as for correlated endpoints and for any joint distribution of test statistics
  - Resampling based methods may not be used for primary endpoints – reasons addressed
  - Bonferroni or Bonferroni-based gatekeeping methods have advantages

Concluding Remarks

- Multiplicity adjustment approaches (cont'd)
  - ✓ Graphical methods useful
  - ✓ Truncated Holm's method – for more power for the 1st primary family
  - ✓ Gatekeeping method w. truncated Holm's tests provides some flexibility
- Some situations - when multiplicity adjustments may not be necessary in a trial.

Concluding Remarks

- There is a widespread interest in using a composite endpoint as a primary endpoint
  - interest in reducing multiplicity and the sample size of the trial.
  - considerations for composite endpoint trials
- Multiplicity problems arise
  - when, in addition, to the composite endpoint, individual components of a composite are intended as possible claims.
- Special interest in the mortality component
  - there are new methods for addressing issues (e.g., 4A, CAS, etc.)
- Interpretation can be challenging in the presence of heterogeneity
  - but meaningful tests still possible on sub-composites satisfying at least directional consistency of effects

Ref: Useful references on multiplicity