Multiplicity Problems in Clinical Trials – A Regulatory Perspective

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

Kathleen Fritsch, Ph.D., Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA

BASS Conference, Rockville, Maryland
November 05, 2014

Disclaimer

- This presentation reflects the views of the presenters and should not be construed to represent FDA’s views or policies.

Outline (Part I)

- Motivational examples (Huque)
- Modern confirmatory controlled clinical trials (Huque)
  - What is different about these trials?
  - Types of multiplicity problems one generally encounters in these trials
- The FDA draft guidance on multiple endpoints (Kathy Fritsch)
  - Key concepts and principles
- Concluding Remarks (Kathy Fritsch)

Outline Part II (Huque)

- Statistical methods (addressed in the draft)
  - Traditional methods
  - Methods based on the concepts of alpha lost and alpha saved
- Additional topics
  - Design and analysis issues for composite endpoint trials
  - Sample size issues for co-primary endpoint trials
  - Subgroup analyses issues for confirmatory trials
  - Closed testing and partitioning methods for solving multiplicity issues of clinical trials
- Final Remarks

Some recent works on the topic

- ✓ **FDA draft guidance on "multiple endpoints in clinical trials," 2014 (to be released soon for public comments)**
- Huque MF, Dmitrienko A, and D'Agostino RB. Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research* 2013 (November)
- Alosh, M; Bretz, F; Huque, MF**.** Recent advances in addressing multiplicity issues in clinical trials. *Statistics in Medicine 2013*
- Dmitrienko A, D'Agostino RB, and Huque MF. Key multiplicity issues in clinical drug development, *Statistics in Medicine* 2013; 32: 1079 –1111
- Dmitrienko A, D'Agostino, RB. Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 2013

Two books and an European points to consider document

- *Multiple Testing Problems in Pharmaceutical Statistics -* 2009

  Editors: A. Dmitrienko, A. C. Tamhane, and F. Bretz. Published by Chapman, and Hall/CRC Press, New York

  Chapter 1: Multiplicity Problems in Clinical Trials. A Regulatory Perspective (by Huque MF, and Röhmel J)

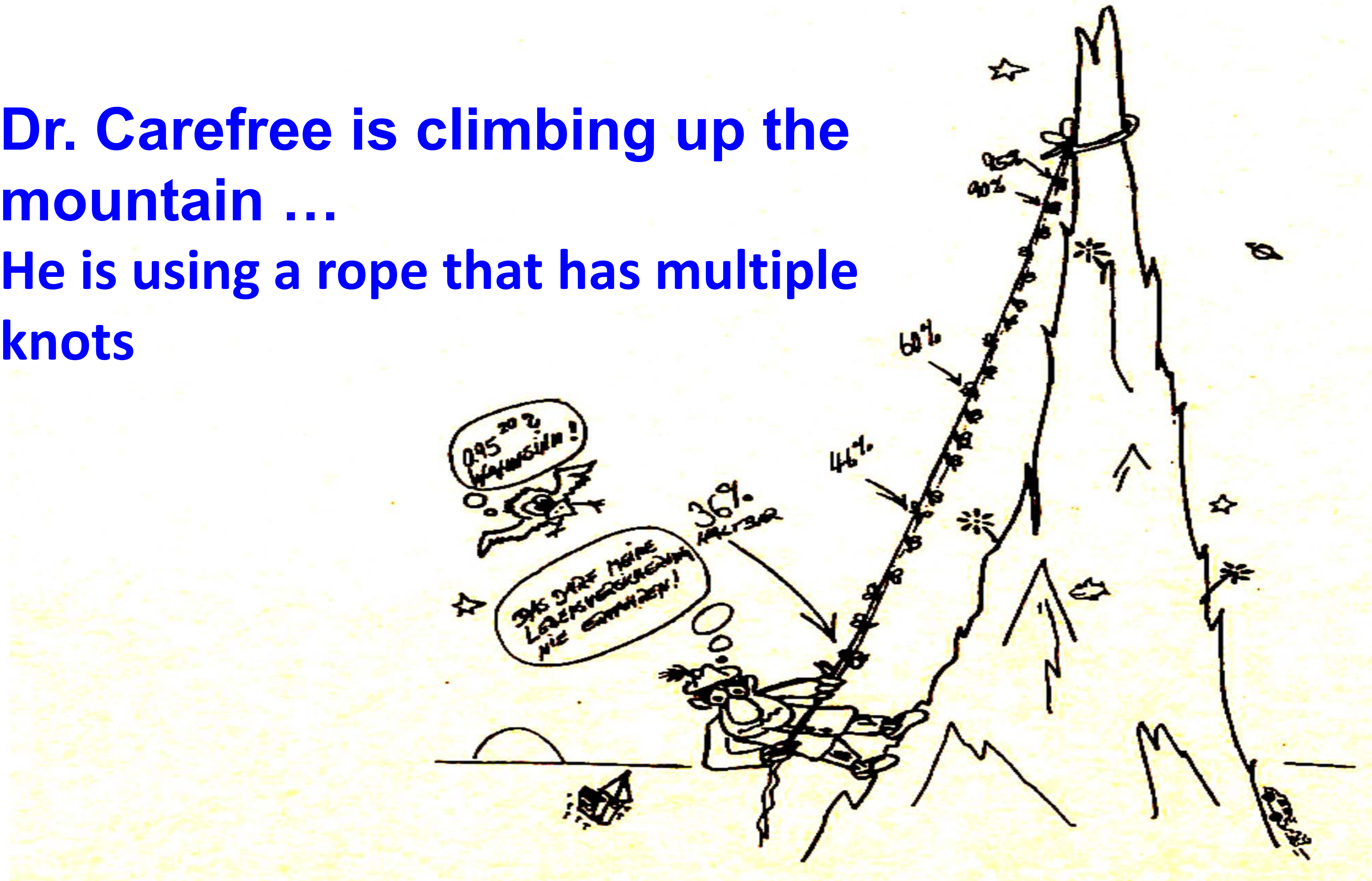
- *Multiple Comparison Using R* - 2010

  by Bretz, F., Hothorn, T., and Westfall, P; Published by CRC Press, New York

- CPMP/EWP/908/99. “Points to Consider on Multiplicity Issues in Clinical Trials,”
  - Available at *http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf*

Dr. Carefree is climbing up the mountain …

He is using a rope that has multiple knots



(aus Beck-Bornholdt und Dubben, 1999)

Picture from a presentation by Franz Koenig (DIA/EMA Conference 2011, London)

Problem:

Each knot can break with a probability of 5%. Guess the probability of falling down the mountain!
Is it 0% or 5% or is it more?

Calculations by a statistician:

Knots	1	2	3	4	5	50
Prob.	0.05	0.10	0.14	0.19	0.23	0.92

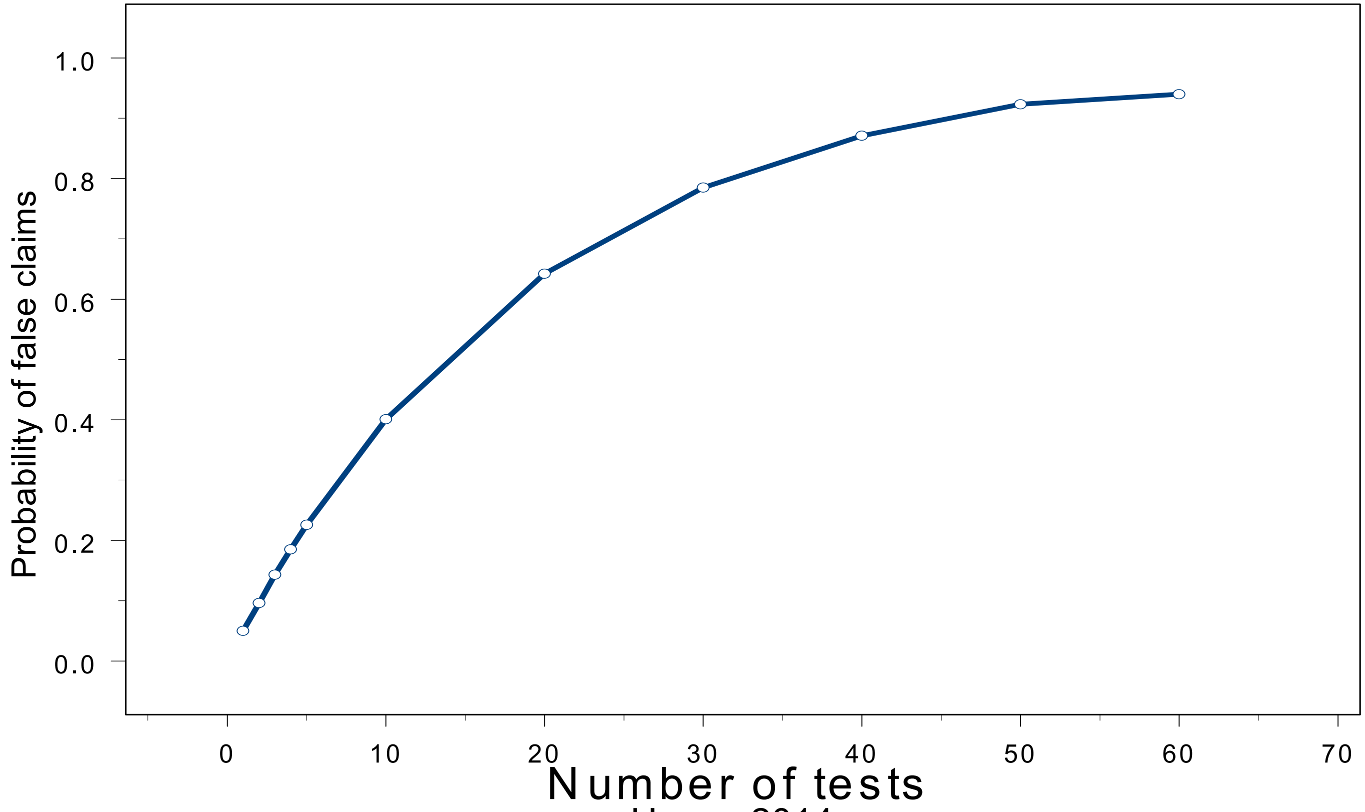
Multiplicity!
Similar problems and challenges arise when testing multiple endpoints (or multiple hypotheses)!!!

Consider a **simulation experiment**

- Simulate on the computer **two-endpoint trials** that compare a treatment to a control, with no treatment effects in any of the two endpoints. Simulate one million times.
  - One would find that about one-hundred-thousand trials (10%) conclude treatment effects on observing *p*-values < 0.05 for at least one of the two endpoints
- These are false positives (or Type I errors) that occur by the play of chance alone in the absence of any treatment effect.
  - The proportion increases with the number of endpoint (or hypotheses) tested
- This phenomenon in testing multiple endpoints (or multiple hypotheses) is known as the inflation of the false +ve error rate or the Type I error rate

Probability of false significant treatment effect findings (Type I error) in a trial can be very high

- When analyzing many endpoints and subgroups each at significance level of 0.05



Assumption: independent tests

Multiplicity and Dr. Carefree!

- Remember Dr. Carefree!
- Using the language of hypothesis testing, the problem is when carrying out more and more (=multiple) tests, the probability of making at least one type I error increases.
- This probability of at least one type I error in testing a family of null hypotheses is sometimes referred as the family-wise error rate (FWER).

Consider two different clinical trial situations

- Situation A: Looking for a significant p-value (P < .05) for a pre-specified single primary endpoint out of say 10 proposed multiple endpoints
- Situation B: Looking for a significant p-value for any of the 10 endpoints
  - Probability of finding a significant p-value (P <0.05) in this case by chance is much greater than that in B

Sometimes one sees the following

- The trial has a single primary enpoint, but has many secondary endpoints – often as large as 10
- All alpha (e.g., 0.05) is spent on the test for the primary objective.
  - If win, then test each secondary endpoint at alpha of 0.05 for significance – statistically problematic
  - If failed, then still try to make the case for treatment effect for a secondary endpoint if that endpoint appears clinically meaningful with p-value < 0.05 or smaller – statistically problematic

Carvedilol example in CHF patients

- Pivotal trials failed on the PE (improvement in ability to exercise).
  - All alpha was lost on the PE, but the drug was approved for the mortality benefit after two AC meetings
  - Mortality endpoint was not the specified PE in the confirmatory trials evaluated
- Two articles with opposite views:
  - Fisher LD. Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections upon hypothesis testing. *Cont. Clin. Trials* 1999; 20:16–39
  - Moye´ L. Endpoint interpretation in clinical trials: the case for discipline. *Cont. Clin Trials* 1999; 20:40–49

Why Problematic?

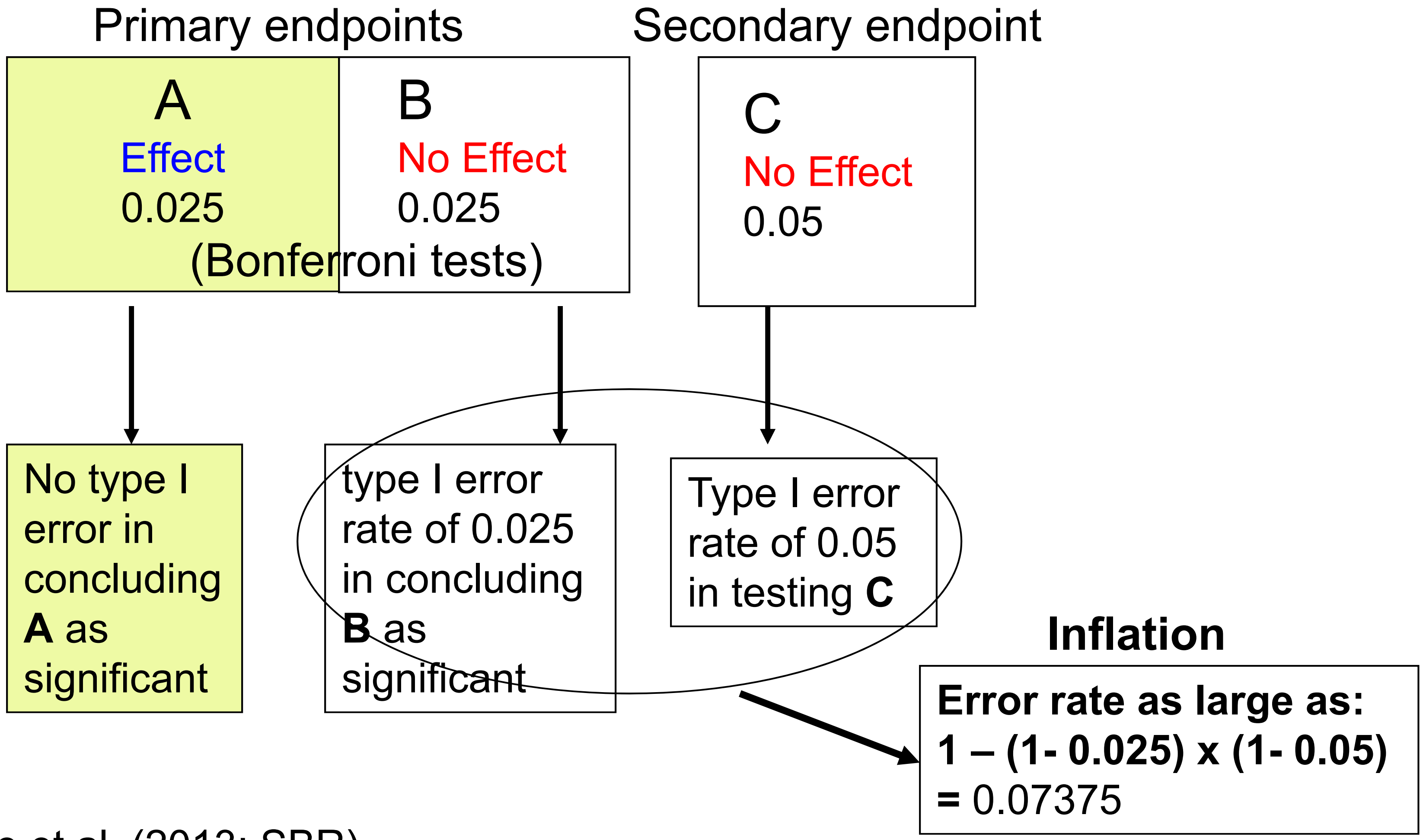
Example 1 (Dmitrienko, D'Agostino, and Huque 2013)

- Consider treatment-to-control comparisons on 3 endpoints:
  - A is primary and B and C are secondary
  - Test strategy: (1) test A at level 0.05; (2) if the test for A is significant, then test B and C each at level 0.05
- Under the global null hypothesis of no treatment effects in any endpoint:
  - The probability of erroneously concluding treatment effect in any endpoint = 0.05. Why?
  - Endpoints B and C are tested only if endpoint A is significant at level 0.05 which renders the size of error rate for secondary endpoints not to exceed 0.05
- Why is it then a problem?

Example 1 (cont'd)

- The previous calculation focused only on one null hypothesis configuration of true and false null hypotheses
- **Doing this can lead to a substantial underreporting of true error rate!!!**
- For example, consider the configuration:
  - The null hypothesis for A is false but those for B and C are true
  - Then the error rate can be as high as $1 - (1 - 0.05)^2 = 0.097$ (assuming tests are independent)

Ex2: Test PEs A and B, each at level 0.025, if win in one of them, then tests the secondary endpoint C at level 0.05



Primary endpoints

Secondary endpoint

A
Effect
0.025

B
No Effect
0.025

(Bonferroni tests)

C
No Effect
0.05

No type I error in concluding **A** as significant

type I error rate of 0.025 in concluding **B** as significant

Type I error rate of 0.05 in testing **C**

Inflation

Error rate as large as:
1 – (1- 0.025) x (1- 0.05)
= 0.07375

Huque et al. (2013; SBR)

Modern (confirmatory) clinical trials

- Include multiple objectives:
  - One primary objective and multiple secondary objectives.
  - Multiple primary objectives and multiple secondary objectives.
- Provide opportunities for winning for multiple treatment benefit claims in the same trial.
- Use novel statistical concepts and methods that save some or all of trial alpha (α) once the trial wins on the primary objective(s).
  - This saved alpha is then used for secondary objectives.

Modern trials face “multiplicity” issues

- Comparing treatments for more than one endpoint
- Comparing several doses of a drug to a control
- Comparing a treatment to control for non-inferiority and for superiority on each of several endpoints and at several doses
- Comparing treatments on multiple primary and secondary endpoints

(Cont’d)

Modern trials face “multiplicity” issues

- Analyzing a composite primary endpoint for claiming treatment benefits for the composite as well as for one or more of its components
- Analyzing for a win either for the total population or for a targeted subgroup
- Conducting Interim analysis
- Making design modifications
- Etc.
- **A complex trial design may combine some or more of the above posing a complex multiplicity problem**

Multiplicity problems encountered in these trials are generally of two types

- Unidimensional multiplicity problems
  - Multiple objectives considered in a clinical trial can be placed in a single family; in other words, they represent the same source of multiplicity
- Multidimensional multiplicity problems
  - Advanced multiplicity problems: or problems with several sources of multiplicity

Unidimensional multiplicity problems

- **Case example 1**:
  - The efficacy profile of a single dose (in comparison to placebo) of a new treatment is evaluated on two (2) endpoints
- **Case example 2**:
  - Three (3) doses of a new treatment are tested versus a common control (e.g., placebo) on a single endpoint
- **Case example 3**:
  - A single-primary endpoint trial compares a single dose of the treatment to control for the overall patient populations as well as for a prospectively defined subpopulation

Multidimensional multiplicity problems

- **Case example 4:**
  - The efficacy profile of new treatment versus a control is to be evaluated at two different dose levels on two primary endpoints and on two secondary endpoints. (w. logical restrictions)

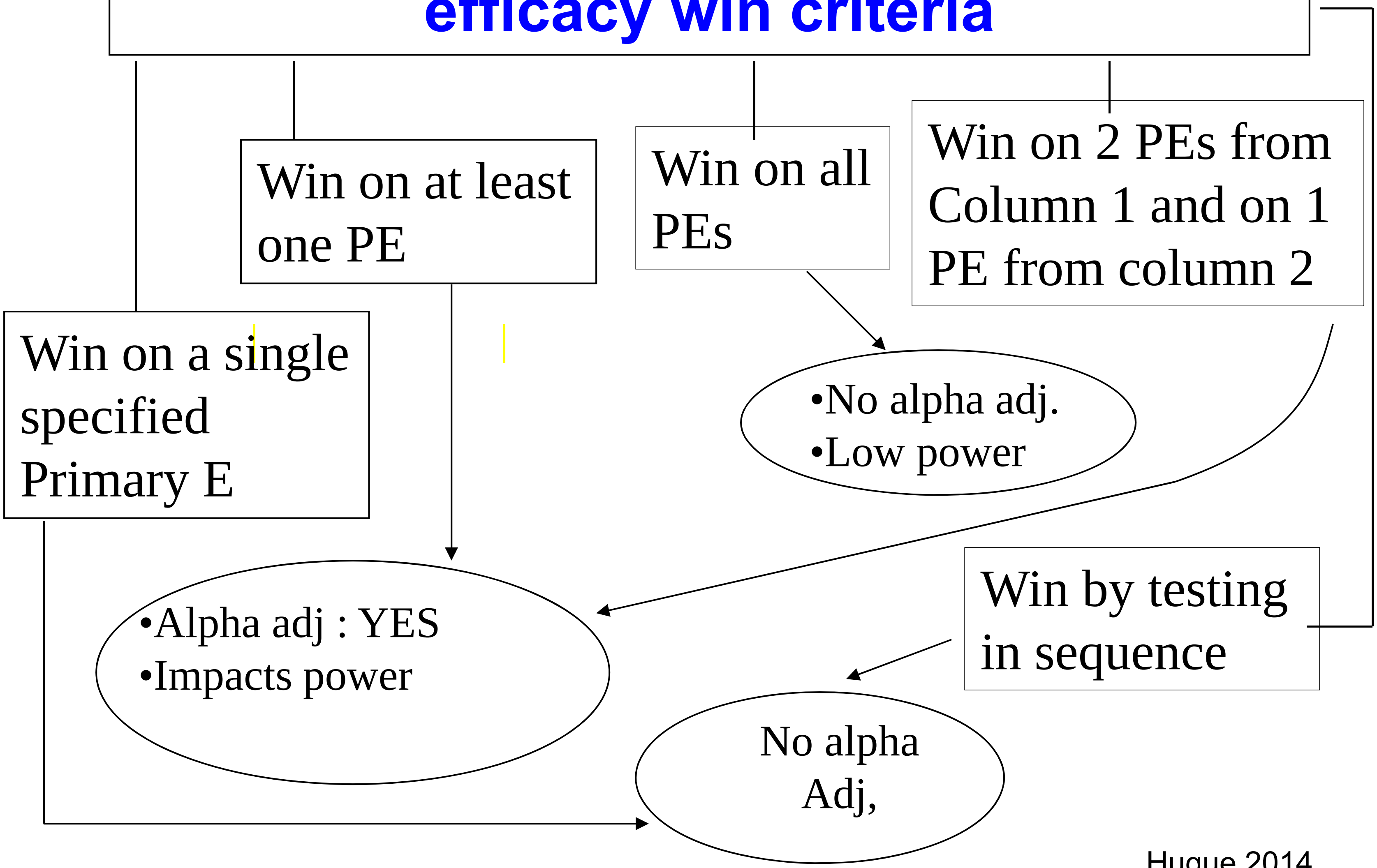
    Examples of logical restrictions:

    1) If a dose is found ineffective for any of the primary endpoints then it can not be tested for a secondary endpoint.
    2) Considered that a primary endpoint is paired to a secondary endpoint as (PE1, SE1). If a dose is found ineffective for PE1 then it can not be tested for SE1.

Multidimensional multiplicity problems (cont'd)

- **Case example 5:**
  - o A trial compares a treatment to control on two primary endpoints (E1 and E2) to determine first that the treatment is non-inferior (NI) to control on endpoint E1. The analytic plan is as follows:
    1) Test E2 only after NI for E1 is first established
    2) Test for superiority on an endpoint only after NI for that endpoint is first established
  - o Dimensionality of the problem increases if the trial is a multi-dose trial.

Trial designs also come with different efficacy win criteria



Win on a single specified Primary E

Win on at least one PE

Win on all PEs

Win on 2 PEs from Column 1 and on 1 PE from column 2

- No alpha adj.
- Low power

- Alpha adj : YES
- Impacts power

Win by testing in sequence

No alpha Adj,

Huque 2014

Good News – Statistical approaches are available for addressing multiplicity

- Last decade has witnessed a surge of research in the development of new methods for addressing multiplicity issues of clinical trials.
- There has been remarkable innovations in statistical methodology in dealing with all sorts of multiplicity problems of clinical trials
- There are now many statistical approaches for addressing different aspects of multiplicity for improving scientific credibility and success of clinical trials.

Last few years have witnessed new useful statistical methods on:

- Methods that allow recycling of alpha from one family to the next
- Gatekeeping and tree-structured methods
- Graphical methods
- Hybrid methods (e.g., combining the Bonferroni and Holm’s critical values)
- Computation of adjusted p-values for any complex hierarchical testing method, e.g., for gatekeeping testing schemes
- The fallback and adaptive alpha allocation approaches (the 4A)
- “Partitioning principle” based test strategies
- Methods for planned subgroup analysis
- Consistency ensured (adaptive) methods
- Others (e.g., related to interim analyses and adaptive designs)

Good News

- The FDA on recognizing the importance of this topic has written draft guidance so that these methods can be applied for regulatory decision making

About this FDA draft guidance

- It is a unique document ever tried at the FDA, written by an FDA committee of statistical and clinical experts.
- It is written in a non-technical language in order to reach a broad audience.
- It includes concepts and methods that were written after much discussions and deliberations for bringing clarity – that is why, it has taken some time to finish it.
  - Achievement: Key statistical methods are described in a way with illustrations that could also be easily understood by clinicians

This draft has 5 sections

I. Introduction

II. Introductory concepts and principles

III. Multiple endpoints: general principles

IV. Statistical concepts, methods and principles

V. Supportive descriptive statistics and graphs

References

Appendix

Scope

- Multiplicity topics addressed are mostly related to adequate and well-controlled studies.
- Some multiplicity topics are beyond the scope of this Guidance. For example, following topics are not addressed.
  - Safety
  - Subgroup analyses
  - Sequential/adaptive designs

Illustrative examples

- Includes illustrative examples related to methods that apply to multiple endpoints.
- Emphasizes that these methods also apply to other situations, such as to different doses, time points, and study population subsets

Includes a number of stat methods w. illustrative examples

- Bonferroni Method
- Holm Procedure
- Hochberg Procedure
- Fixed Sequence Method
- Modified Fixed Sequence method
- Gatekeeping Testing Strategies
- Truncated Holm Procedure for Parallel Gatekeeping
- Multi-branched (Tree-structured) Gatekeeping Procedures
- Resampling Based Multiple Testing Procedures
- Graphical method

Addresses a number of multiplicity topics and issues including

- Primary and secondary endpoints
- Multiplicity and its extent
- False positive error rate and its control
- Prospectively planned and post-hoc analyses
- Co-primary endpoints and issues
- Composite and multi-component endpoints and issues
- Descriptive statistics and graphs for labeling

Defines and explains endpoint families

- Primary endpoints
  - Endpoint(s) necessary and/or sufficient to establish efficacy (define a successful trial)
- Secondary endpoints
  - Not sufficient to establish efficacy in the absence of an effect on the primary endpoints; not required for establishing efficacy
  - Potentially could lead to additional labeling claims
- Exploratory endpoints
  - Hypothesis generating (clinical utility unknown)

Defines and explains “multiplicity”

- Multiplicity refers to situations in a trial in which multiple statistical tests or analyses create multiple ways to “win” for treatment efficacy or safety.
  - This can cause the **false positive error rate** **(Type I error rate)** to inflate beyond the desired level, e.g., 0.05, if each test is performed, for example, at the same alpha level of 0.05.
- This inflation in a trial can be substantial and problematic, but

  it can be controlled to a desired level by an appropriate, prospectively planned statistical strategy using the statistical framework of testing multiple hypotheses.

Huque 2014

Explains when is it necessary to adjust for multiplicity?

- When there are one or more claims of treatment benefits based on primary and secondary endpoints.
- When the win criteria are such that one can win in multiple ways, i.e., there are multiple pathways for winning.

✓ Such situations require multiplicity adjustments when they cause inflation of the Type I error rate.

Explains what is not multiplicity

- Often there are multiple analyses for the intention-to-treat (ITT) data set for the same PE and by the same method
  - These multiple analyses are done for the same endpoint on varying the assumptions about some data points because of missing data, protocol violations, use of concomitant medications, etc.
- As these analyses are sensitivity analyses for assessing the primary analyses results, there is no multiplicity adjustment for them

What is not multiplicity (cont'd)

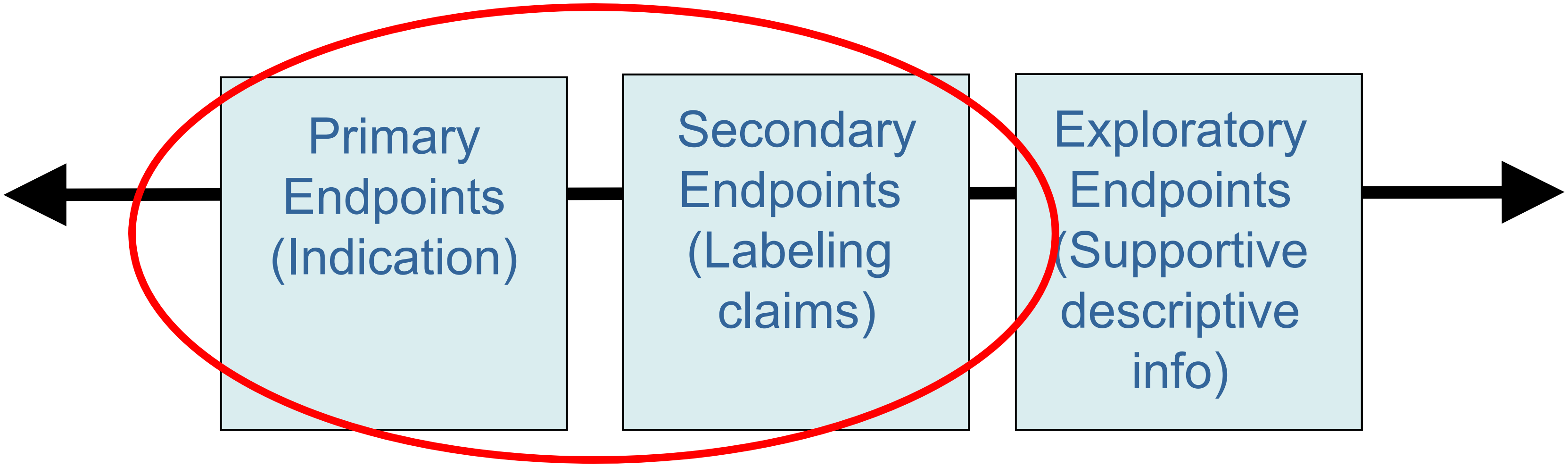
- Often there are analyses of the same endpoint data by alternative methods, e.g.,
  - analysis of the same time-to-event endpoint by log-rank test and by the generalized Wilcoxon test
  - analysis by the parametric and non-parmaetric methods.
- Technically, one can adjust for these multiple analyses 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.

What is not multiplicity (cont’d)

- The Draft Guidance considers descriptive analyses and graphs that go into the labeling as being “not multiplicity” – Section V of the document is devoted to this topic.
- These analyses are supposed to be further elaborations of effecs that has been established in a statistically rigorous way.
- **Caution: These analyses should be recognized as insufficient to justify additional drug efficacy claims beyond those supported by the prospective analyses**.

Considers error rate control for the primary and secondary families of hypotheses

Continuum for Type I Error Control


Primary Endpoints (Indication)
Secondary Endpoints (Labeling claims)
Exploratory Endpoints (Supportive descriptive info)

→To all primary and secondary endpoints

→Overall error rate should not exceed a pre-specified α

Huque 2014

Recommendations on stat methods for controlling the Type I error rate

- Methods generally used for the primary and secondary endpoints should be those that allow finding of significant treatment effects at the individual endpoint level, without inflating the Type I error rate
- These methods permit an individual conclusion about efficacy with respect to each endpoint tested in the primary and the secondary family
- **Some methods (often called global procedures) allow a conclusion of treatment efficacy in the global sense. Such methods generally inflate the Type I error rate for making conclusions on the individual endpoints.**

Emphasizes prospective planning as a key to addressing multiplicity

- An important component in controlling for multiple comparisons is to specify in the protocol all planned study endpoints, time points, subgroups, and analyses in advance.
- Changes in the analysis plan to perform non-prospectively stated analyses can reintroduce a multiplicity problem

Explains pitfalls of post-hoc analyses

- Although post-hoc analyses of trials that fail on their specified endpoints may be useful for generating hypotheses for future testing, they do not yield definitive results.
- The results of such analyses can be biased, as the choice of analyses surely can be influenced by a desire for success.
- It is difficult to confirm how many different analyses were performed; in this situation, there is no credible way to correct for the multiplicity of multiple analyses and control theType I error rate.
- **Consequently, post hoc analyses generally do not provide evidence of effectiveness.**

Explains when in clinical trials co-primary endpoints are used

- Situation 1: When there are two or more critically important different features of a disorder
  - These features are so critically important to the disease that a drug will not be considered effective without demonstration of a treatment effect on all these disease features.
- Example:
  - Migraine headaches are characterized by the presence of pain, photophobia, phonophobia, and nausea.
  - A treatment is considered effective for migraines if all four aspects of the disorder are shown to be improved by the drug treatment.

Co-primary endpoints (cont'd)

- Situation 2:
  - When there is a single identified critical feature of the disorder, but there is no single patient evaluation that is both specific for the disease feature and is clinically interpretable.
  - In these cases, two endpoints are often used.
- Example:
  - Alzheimer's disease trial with endpoints: ADAS-Cog and a global measure of function (e.g., global assessment)
  - One endpoint assures that the effects occurs on the core disease feature, and the other that the effect is clinically meaningful.

Statistical considerations for co-primary endpoints

- **When using co-primary endpoints, testing each individual endpoint at the 0.05 level does not cause inflation of the Type I error rate,**
  - rather the impact of co-primary endpoint testing is on the Type II error rate."
- In general, unless clinically very important, the use of more than two co-primary endpoints should be carefully considered because of the loss of power.
- **Relaxation of alpha is not generally acceptable because doing so will undermine the unequivocal demonstration of an effect on each disease aspect considered essential for showing that the drug has the desired effect.**

Huque 2014

Gives an idea of creating a single endpoint from multiple co-primary endpoints

- **Idea:** A successfully treated patient will be that who improves on all the identified necessary endpoints.
- For this, each of the endpoints can be made dichotomous by applying the specified threshold for improvement.
- This can allow classifying patients as responders versus non-responders, and a primary endpoint might be formulated to compare the proportion of responders in each group.

Addresses composite and multi-component endpoint issues in detail (some key points:)

- A common approach in practice has been to combine multiple endpoints (called components) to a single composite (or a single multi-component endpoint) when
  - components individually are expected to yield small treatment effects, but collectively they can show a clinically meaningful benefit.
- Such an approach can effectively reduce the size of the trial if components contribute to the total treatment effect in a meaningful way.
- If individual components were tested simultaneously (e.g., by the Bonferroni test), when expecting only small treatment effects in each, then such an approach would not be practical.

Interpretation of the composite endpoint findings

- The treatment effect on the composite describes the overall clinical effect of the treatment when
  - components all are of reasonably similar clinical importance, and
  - components exhibit some consistency of treatment effects.
- Interpretation difficulties arise when
  - the clinical importance of different components is substantially different, and
  - the treatment effect is mainly on the least important component.

Interpretation of the composite endpoint findings (cont'd)

- If a critical component (e.g., mortality) is adversely affected by the treatment, even if one or more components of less importance are favorably affected, so giving an overall favorable statistical result.
- Then, in that case, while the overall analysis indicates that the treatment is successful, careful examination of the data may call this conclusion into question.

✓ **A key recommendation: For interpretation purposes, component endpoint data are to be fully displayed and carefully examined.**

(Draft Guidance addresses this issue in detail)

Multiplicity issues in composite endpoint trials

- There is no multiplicity issue if the trial has a single composite endpoint as the sole primary endpoint, and there no claim of treatment benefit for its specific components.
  - Component outcomes are analyzed and displayed only in the descriptive sense as an aid to interpreting the result of the composite endpoint.
- Multiplicity issues arise when, for example,
  - claims of treatment benefit are sought for the composite endpoint, as well as for its sub-composites or for its individual components.
- Most of these multiplicity issues can be address by a variety of multiple testing methods (e.g., by gatekeeping and graphical methods)

Section V: Descriptive statistics and graphs for labeling

- These are to substantiate further the results of the primary and secondary objectives already established by rigorous statistical methods.
- Used, for example, for:
  - showing treatment effects using histogram and cumulative distribution plots;
  - checking consistency of results across patient subgroups by the use of forest plots.
- Cautions regarding their improper uses.

Outline Part II (Huque)

- Statistical methods (addressed in the draft)
  - Traditional methods
  - Methods based on the concepts of alpha lost and alpha saved
- Additional topics
  - Design and analysis issues for trials with event type composite endpoints
  - Sample size issues for co-primary endpoint trials
  - Subgroup analyses issues for confirmatory trials
  - Closed testing and partitioning methods for solving multiplicity issues of clinical trials
- Final Remarks

Two types of multiple hypotheses testing problems

- (1) Union-intersection testing problems

Given k multiple null hypotheses H_1, H_2, …, H_k to be tested against their corresponding alternative hypotheses K_1, K_2, …, K_k

$H_I = \cap H_j$ is tested against $K_U = \cup K_j$; j =1, …, k

Comment:

Carrying out of the individual null hypotheses tests at an unadjusted nominal α level leads to an inflated probability of rejecting H_I and can compromise the validity of the statistical inference.

Two types of multiple hypotheses testing problems (cont'd)

- (2) Intersection-union testing problems

$H_U = \mathrm{U}H_j$ is tested against $K_\mathbf{I} = \cap K_j$; j =1, …, k

Comments:

–No multiplicity adjustment is necessary for controlling the overall Type I error rate, but individual hypothesis can not be tested at levels higher than the nominal significance level of α.

–Application: testing of co-primary endpoints in clinical trials; e.g., Alzheimer's trials require to show treatment effects on both the cognition and the global clinical endpoint

Hybrid problems

- Example: Consider an epilepsy trial with primary endpoints: PE1 = seizure rate, PE2 = drop attack rate, and PE3 = seizure severity. A win criterion may be the following:
  - Show a beneficial effect either on PE1 or on both PE2 and PE3.
  - Thus, PE2 and PE3 act as co-primary where PE1 is not co-primary.

    $H_h = H_1 \cap (H_2 \cup H_3)$ versus $K_h = K_1 \cup (K_2 \cap K_3)$

  Question: Will the Type I error rate be controlled at level α, if one tests H_1 at level α/2, and tests each H_2 and H_3 at level $\alpha/2$?

Classification of multiple testing procedures into single and multi-step procedures

- *Single-step* procedures (e.g., the Bonferroni method,):
  - Provide simultaneous testing and simultaneous (adjusted) confidence intervals for assessing the magnitude of the treatment effects. They tend to cause loss of study power.
  - Are characterized by the fact that rejection or non-rejection of a hypothesis does no depend on the decisions on the other hypotheses tested
  - The order in which the hypotheses are tested is not important.

Multi-step Procedures

- Generally more efficient, better preserving the power, but do not readily provide adjusted confidence intervals. (How about 1-sided confidence limits?)
- Rejection or non-rejection of a null hypothesis may depend on the decisions on other hypotheses (example: Holm test)

- There are several kinds of multistep procedures, for example
  - ***step-down***
  - ***step-up***, and
  - **sequential procedures**: Order of hypotheses to be tested are pre-specified as compared to determined by the data

Step-down and step-up procedures

- Step-down procedures (e.g., Holm procedure):
  - One calculates the p-values from all tests to be considered at one time and starts with the smallest *p*-value and then steps-down to the next smallest *p*-value and so on.
- Step-up procedures (e.g., Hochberg procedure):
  - One proceeds in the reverse direction. That is, one starts with the largest *p*-value and steps-up to the second largest *p*-value, finally reaching the smallest *p*-value.

Bonferroni method

- The draft guidance introduces the un-weighted and weighted Bonferroni methods with examples
- Recognizes the following: :
  - The Bonferroni method tends to be conservative in controlling for the study overall Type I error rate if the number of endpoints (or hypotheses) tested is large or endpoints are strongly positively correlated.
  - Consider a case of three endpoints: All of three endpoints give nominal *p*-values between 0.025 and 0.05, i.e., all 'significant' at the 0.05 level. Such an outcome seems intuitively to show effectiveness on all three endpoints.
  - However, the Bonferroni method will fail to declare any of these p-values as significant.

Bonferroni method (cont'd)

- When there are more than two endpoints (e.g., 5 endpoints) with substantial correlation between them the true family-wide type I error rate may decrease from 0.05 to approximately 0.04 to 0.03 (when the correlation is 0.6 to 0.8)
- **The Bonferroni method is assumption free - ideal for testing primary hypotheses when they are very few**

  Can be applied without being concerned about the endpoint types, their (joint) statistical distributions, and the type of correlation structure

Holm and Hochberg tests

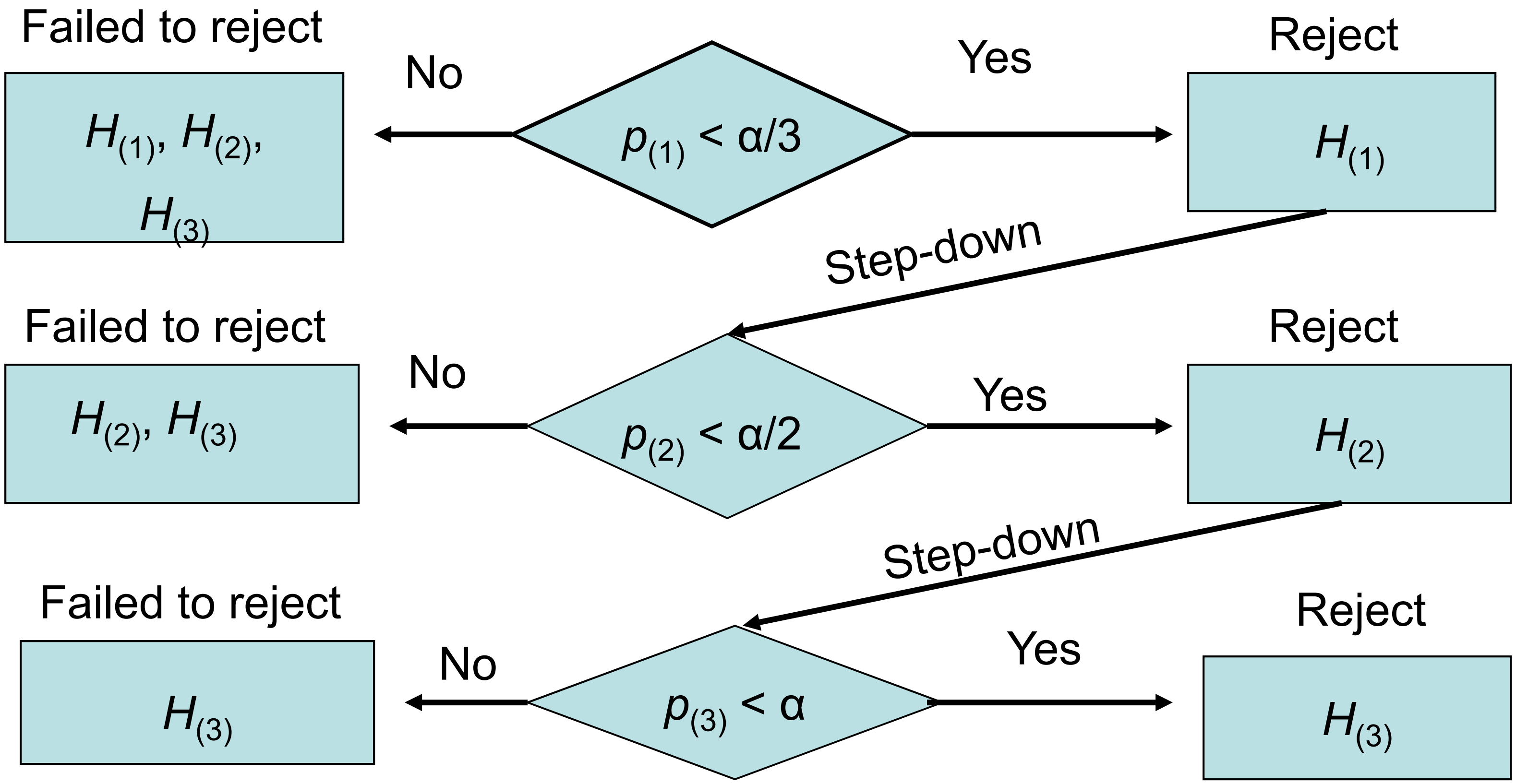
- Both uses the same alpha critical values but in different ways
- Consider *k* null hypotheses $H_1, H_2, . . . , H_K$ in a family with associated *p*-values of $p_1, p_2, . . . , p_K$, and suppose that $H_{(1)}, H_{(2)}, . . . , H_{(K)}$ are the ordered null hypotheses corresponding to the ordered *p*-values of $p_{(1)} \leq p_{(2)} \leq . . . \leq p_{(K)}$.
- In both the Holm and Hochberg tests, alpha critical values in testing null hypotheses are

$$c_{(i)} = \alpha/(K - i + 1) \text{ for } i = 1, . . . , K.$$

e.g., for $K = 3$, $c_{(1)} = \alpha/3$, $c_{(2)} = \alpha/2$, and $c_{(3)} = \alpha$

Holm test with K =3

Ordered p-values of $p_{(1)} \leq p_{(2)} \leq p_{(3)}$
Associated hypotheses $H_{(1)}$, $H_{(2)}$, $H_{(3)}$
Start from the top with the smallest p-value $p(1)$ then step-down



Failed to reject

$H_{(1)}$, $H_{(2)}$, $H_{(3)}$

No

$p_{(1)} < \alpha/3$

Yes

Reject

$H_{(1)}$

Step-down

Failed to reject

$H_{(2)}$, $H_{(3)}$

No

$p_{(2)} < \alpha/2$

Yes

Reject

$H_{(2)}$

Step-down

Failed to reject

$H_{(3)}$

No

$p_{(3)} < \alpha$

Yes

Reject

$H_{(3)}$

Holm test

- Assumption free - similar to Bonferroni test
  - But, ordering of hypotheses is data dependent.
  - Uniformly, more powerful then the Bonferroni test
- Can reject more hypotheses than the Bonferroni test, e.g.,

  K=3; p_1 = 0.01, p_2 = 0.024, p_3 = 0.04

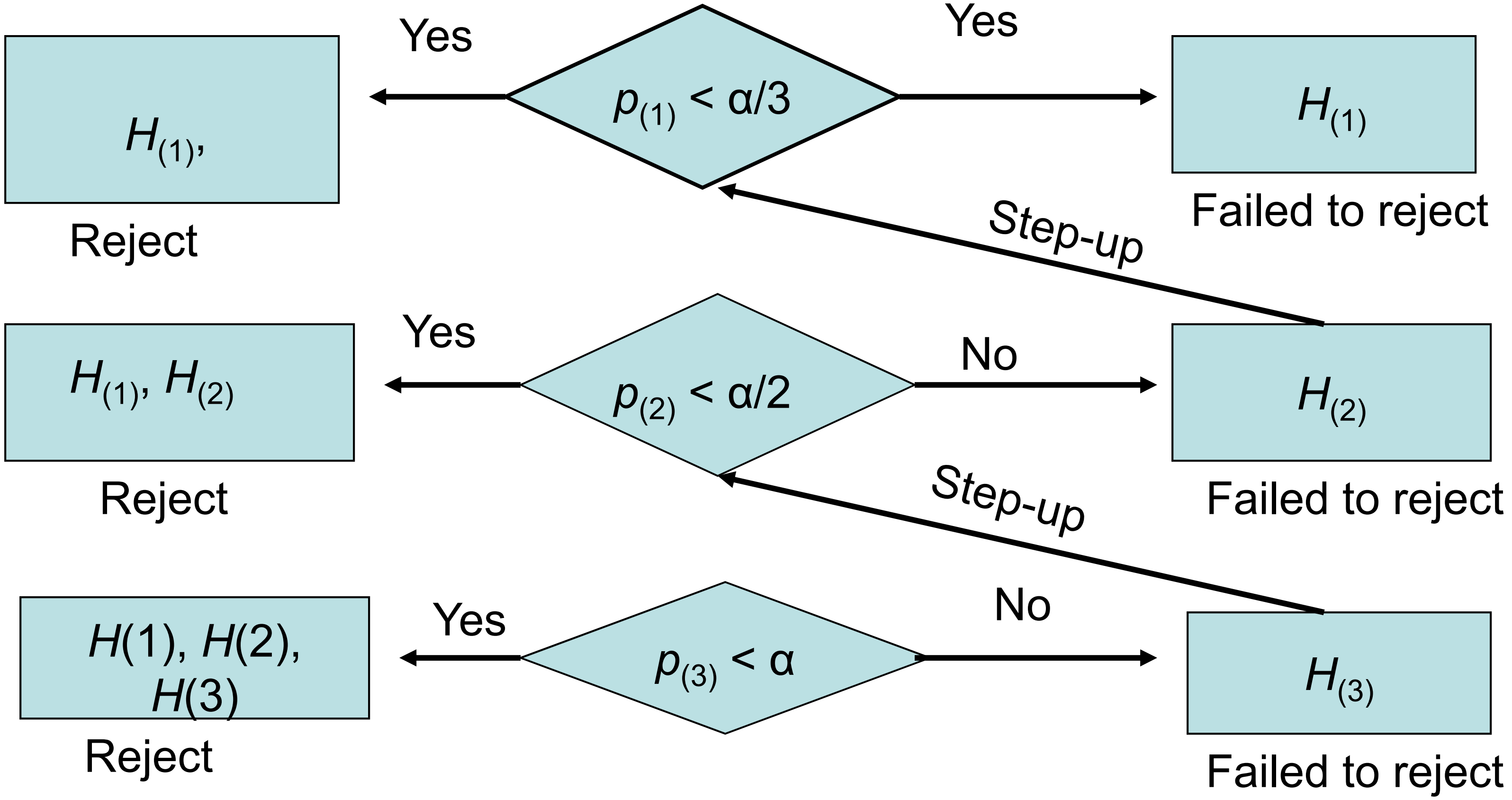
  Only, one p-value of p_1 = 0.01 significant at level 0.05 by the Bonferroni test, but all significant by the Holm test
- Still conservative:
  - E.g., none is significant at level 0.05 if K=3 and 0.025 < (p_1, p_2, p_3) < 0.05

Hochberg test with *K* =3

Start from the bottom with the largest p-value *p*(3) then step-up

Ordered *p*-values of $p_{(1)} \leq p_{(2)} \leq p_{(3)}$

Associated hypotheses $H_{(1)}$, $H_{(2)}$, $H_{(3)}$



$H_{(1)}$,

Reject

Yes

$p_{(1)} < \alpha/3$

Yes

$H_{(1)}$

Failed to reject

Step-up

$H_{(1)}$, $H_{(2)}$

Reject

Yes

$p_{(2)} < \alpha/2$

No

$H_{(2)}$

Failed to reject

Step-up

H(1), *H*(2), *H*(3)

Reject

Yes

$p_{(3)} < \alpha$

No

$H_{(3)}$

Failed to reject

Hochberg procedure

- **Not assumption-free like the Bonferroni and Holm tests.**
  - It is known to provide adequate overall α-control for independent and for certain types of positively correlated tests (Sarkar and Chang, 1997; Sarkar, 1998), but its properties for other types of dependent endpoints are not fully known.
  - For 2-endpoint tests (i.e., testing for two null hypotheses): OK if test statistics follow bivariate normal density with positive correlation. This result follows from Example 1 given in Sarkar & Chang (1997) and also from the work of Samuel-Cahn (*Biometrika* 1996).
- **Therefore, unless the Sarkar et al. conditions hold or the Type I error control is clearly demonstrated, the Hochberg procedure is generally not recommended for confirmatory trials.**

Comments: PAAS

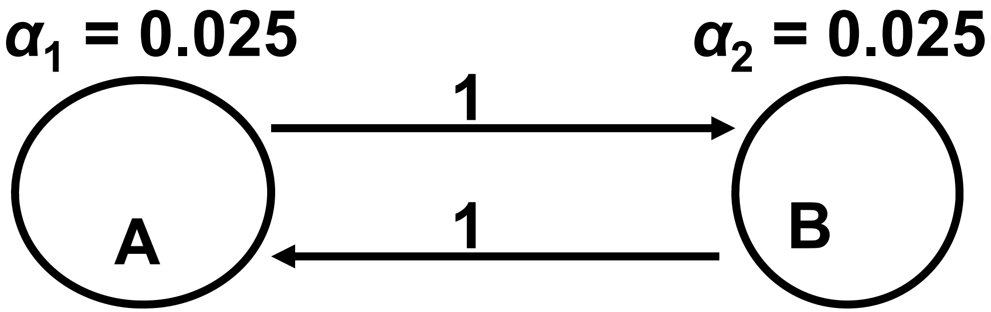
- The PAAS (Moyé, 2000) is a single-step method
  - Has a slight advantage in power over the Bonferroni
  - Allows equal or unequal allocations, but, as with the Bonferroni, each specific endpoint must receive a prospective allocation of a specific amount of the overall alpha.
  - Alpha allocations are required to satisfy the equation:

$$(1 - \alpha_1)(1 - \alpha_2) \ldots (1 - \alpha_k) \ \ldots \ (1 - \alpha_m) = (1 - \alpha)$$

- Caveat: The PAAS assures strong FWER control for all comparisons that are independent or positively correlated.

Stat test strategies commonly use the concepts of “alpha saved” and “alpha lost”

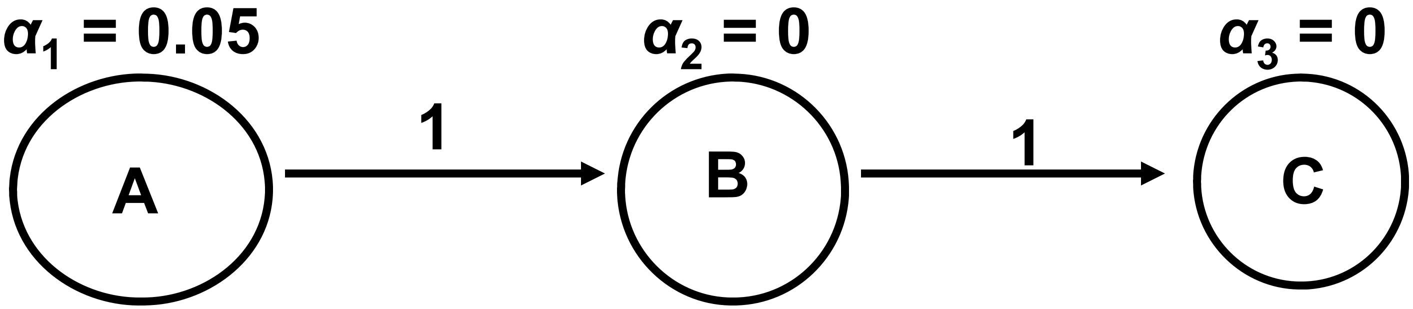
- If an endpoint (or hypothesis) is tested at a level alpha (e.g., alpha = 0.025) and the *p*-value is significant at that level then that alpha of 0.025 is “saved” and can be accumulated to test a second prospectively specified endpoint (or hypothesis)



Thus, if A is successful, then alpha at B is 0.025 +1*0.025 = 0.05

This is basically the graphical representation of the Holm’s test for testing two endpoints

A test strategy that does not require alpha-adjustments



If A is successful, alpha for B becomes 0 +1*0.05 = 0.05. Then, if B is successful alpha for C is 0.05. But, if anytime, a test is not significant there is no further test

This test strategy is known as the fixed sequence test method

Fixed sequence test strategy

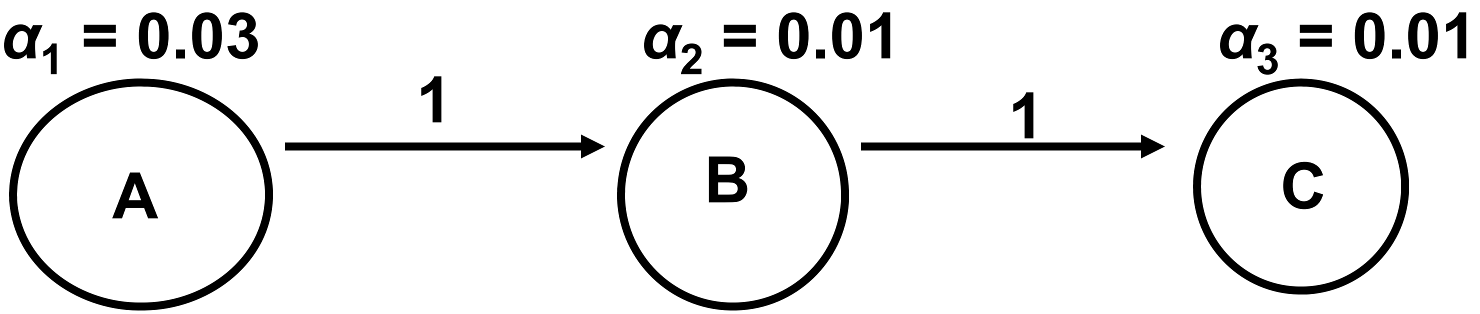
- A fixed-sequence statistical strategy tests endpoints in a pre-defined fixed sequential order, all at the same significance level α (e.g., $\alpha = 0.05$), moving to a second endpoint <u>only</u> after a success on the previous endpoint.
- Such a test procedure does not inflate the type I error rate so long as there is

1) pre-specification of the testing sequence, and
2) there is no further testing once the sequence breaks, i.e., the result is not significant

Comments – fixed sequence method

- **Drawback**: If a hypothesis in the sequence is not rejected then a statistical conclusion cannot be made about the endpoints planned for the subsequent hypotheses, even if they have extremely small *p*-values.
  - Suppose, for example, that in a study the *p*-value for the first endpoint test in the sequence is $p = 0.250$, and the *p*-value for the second endpoint is $p = 0.00001$.
  - Despite the apparent "strong" finding for the second endpoint, no formal favorable statistical conclusion can be reached for this endpoint.

A fix for this drawback: Save a little alpha on A and distribute the remainder to others



If A is successful, alpha for B becomes 0.01 +1*0.03 = 0.04, and if B is also successful, then test for C is at level 0.05 (This test strategy is known as the fallback method)

Comments –fallback method

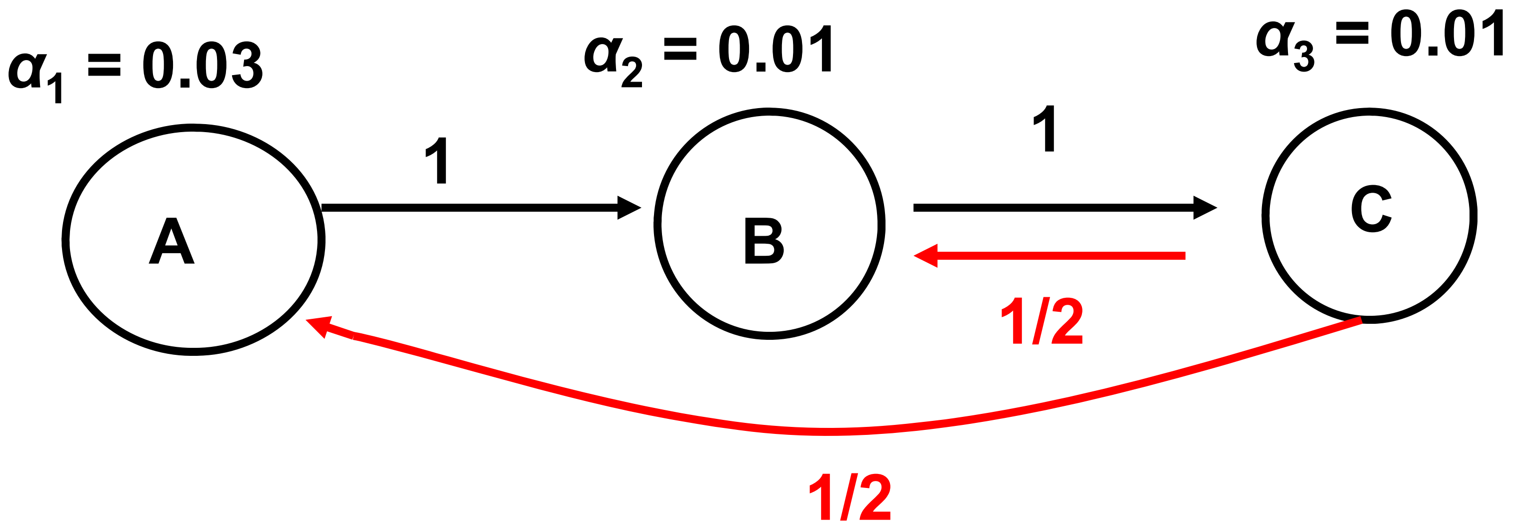
- One usually assigns the majority of the alpha to the first endpoint and the remainder to the second endpoint, although other distributions are also valid.
- It is often used when there is another endpoint thought less likely to be statistically significant, and thus is not designated as the first endpoint.
  - However, this second endpoint is of such clinical importance that it would be valuable to have an unexpectedly robust finding for this endpoint that would be statistically interpretable without inflation of the Type I error.

Comments –fallback method (cont'd)

- The statistical power of the fallback method primarily depends on the magnitude of the effect on each of the ordered endpoints and alpha assigned to them
- As with the simple fixed sequence method, the overall power of the fallback method exceeds than that of the Bonferroni test
  - because when the earlier endpoints show significant results, the method uses larger alpha levels for later endpoints than is possible under the Bonferroni method.

Extension of the fallback method

Consider the situation: **A and B both fail but C is successful**



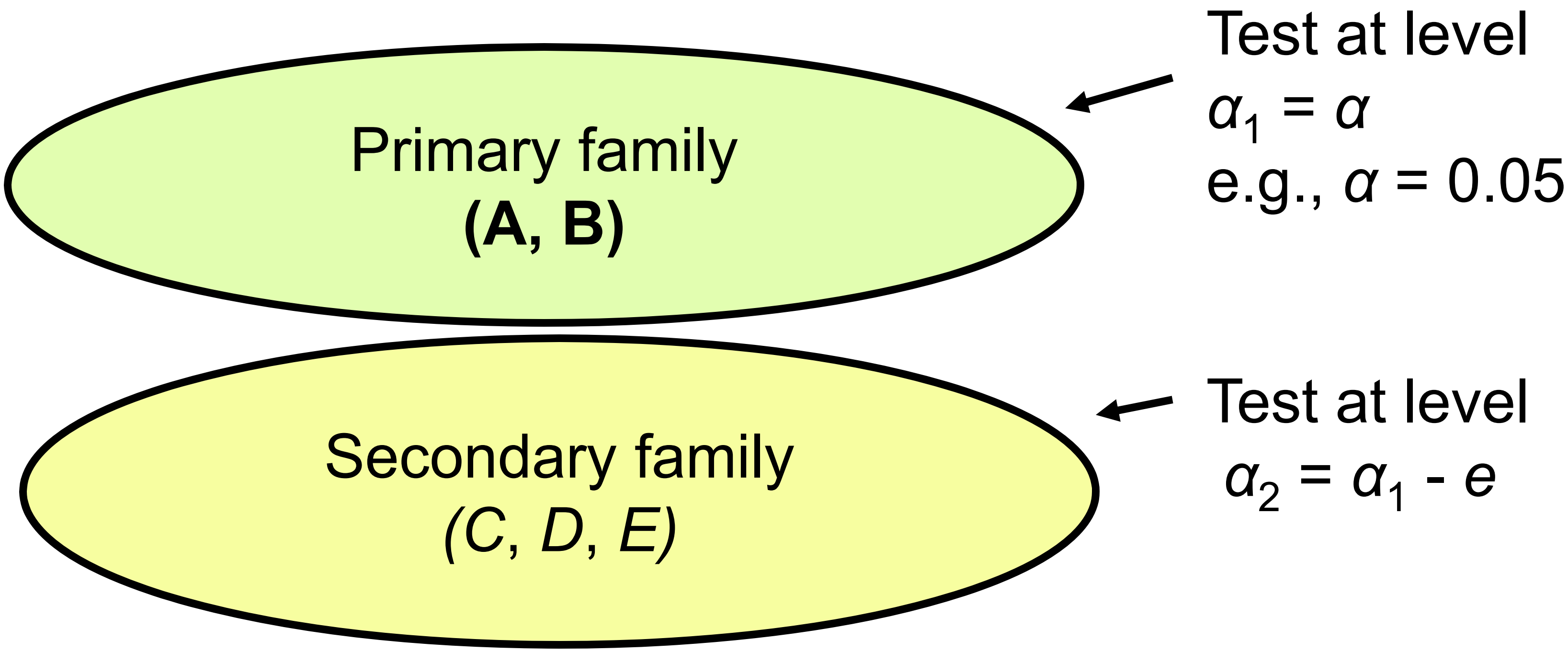
Then A and B can be retested at slightly higher levels

Gatekeeping test strategy

- Useful for testing the Primary and Secondary families of endpoints
- The usual strategy is to test all endpoints in the primary family by a method such as Bonferroni and proceed to the secondary family of endpoints only if there has been statistical success in the primary family.
- This allows all of the trial alpha to be used for the primary family. Thus, maximizing the study power for those critical endpoints.

Gatekeeping test strategy

- Consider two families of endpoints (or hypotheses), one primary and the other secondary



Primary family
(A, B)

Test at level
$\alpha_1 = \alpha$
e.g., $\alpha = 0.05$

Secondary family
(C, D, E)

Test at level
$\alpha_2 = \alpha_1 - e$

“e” depends on how many endpoints in the primary family are successful. If all endpoints are successful in this family are successful then e = 0.

The Gatekeeping testing (cont'd)

- Two common gatekeeping test strategies are: serial and parallel.
- Serial strategy is applied when the endpoints of the primary family are tested as co-primary endpoints.
  - If all endpoints in the primary family are statistically significant at the same level α (e.g., α = 0.05), the endpoints in the second family are examined by any one of several possible methods (e.g., Holm procedure).
  - Else, the secondary endpoint family is not tested.

The Gatekeeping testing (cont'd)

- Parallel gatekeeping strategy (PGS) is applied when the endpoints in the primary family are not all co-primary endpoints, and a testing method (e.g., Bonferroni or Truncated Holm method) that allows the pass-along of alpha from one family to the next is specified.
  - In this strategy the second endpoint family is examined when at least one of the endpoints in the first family has shown statistical significance.

An example of PGS with Bonferroni tests

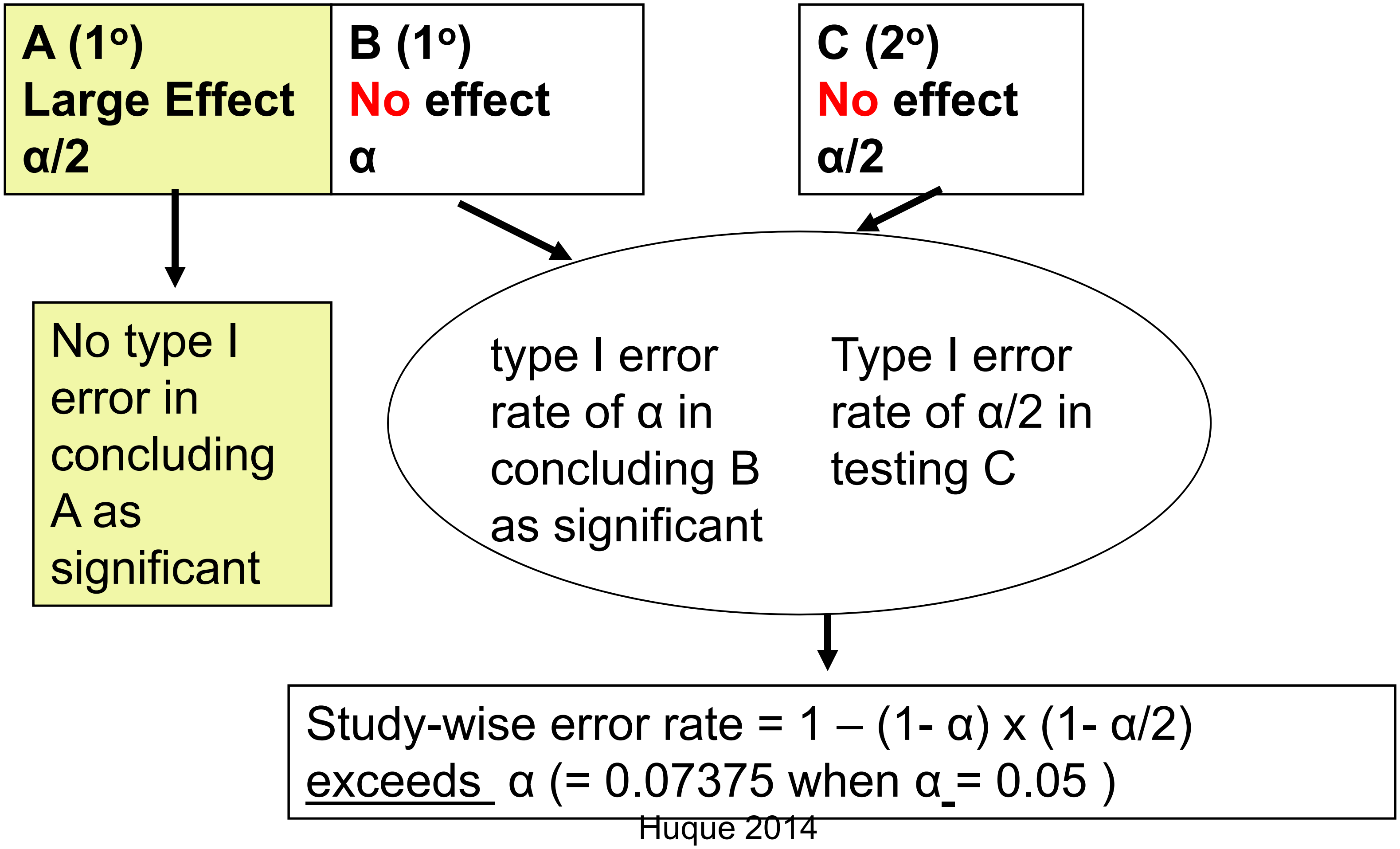
- Consider a trial whose primary objective is to test for superiority of a treatment to placebo for five endpoints: A, B, C, D and E.
- Primary family F1 = {A, B} and secondary family F2 = {C, D, and E}.
- Statistical plan:
  - Test endpoints A and B in F1 by the Bonferroni method at endpoint-specific alpha levels of 0.04 and 0.01, respectively, with the total initial α = 0.05 assigned to F1.
  - Test the second family by the Holm procedure with whatever amount of alpha is passed along to it.

An example of PGS with Bonferroni tests (cont'd)

- Suppose that for the family F1, the *p*-values for the endpoints A and B are 0.035 and 0.055, respectively.
  - Then the result for endpoint A is significant, but the result for endpoint B is not, leaving alpha of 0.04 as unused and alpha of 0.01 as used.
- Therefore, the total alpha available for testing the endpoints in F2 is 0.04 and not 0.05.

Regular Holm (also Hochberg) tests can’t be used for the primary family

Primary endpoints



A (1^{o})
Large Effect
$\alpha/2$

B (1^{o})
No effect
α

C (2^{o})
No effect
$\alpha/2$

No type I error in concluding A as significant

type I error rate of α in concluding B as significant

Type I error rate of $\alpha/2$ in testing C

Study-wise error rate = $1 – (1- \alpha) \times (1- \alpha/2)$
exceeds α (= 0.07375 when $\alpha = 0.05$)

Problem with the regular Holm and Hochberg tests in the PGS

- Bonferroni method (or any other separable method; see, Dmitrienko et al., 2008) has an important property of preserving some alpha for testing the secondary endpoint family when the result for at least one of the endpoints in the primary family is statistically significant.
  - The endpoint-specific alpha from each test that successfully rejects the null hypothesis is summed and becomes the alpha available to the secondary endpoint family.
- The conventional Holm and Hochberg tests does not have this property because they are alpha-exhaustive; they lead to non-separable tests.

Use of the truncated Holm test in the PGS (Dmitrienko et al., 2008)

- The truncated Holm test allows passing of alpha, but the calculation of un-used alpha is different than that by a Bonferroni based method
- In the truncated Holm, the critical values for tests are convex combinations of the critical values of the original Holm test and that of the Bonferroni test

  $c_\mathrm{i} = \gamma(\alpha/(k - i + 1)) + (1\text{-}\ \gamma)(\alpha/k)$,

  where, $0 \leq \gamma < 1$ is the truncation fraction.

  - At $\gamma = 0$, this construct gives the Bonferroni alpha-critical value of α/k.
  - The actual procedure for the truncated Holm remain the same, except that the above new critical values c_i are used

PGS with the truncated Holm test, K =3 (primary family)

- Family 1 test:
  1) Reject $H_{(1)}$ if $p_{(1)} < c_1 = \alpha/3$, else stop testing
  2) Reject $H_{(2)}$ if $p_{(2)} < c_2 = (\gamma + 2)\alpha/6$ after rejecting $H_{(1)}$, else stop testing, and
  3) Reject $H_{(3)}$ if $p_{(3)} < c_3 = (2\gamma + 1)\alpha/3$ after rejecting $H_{(1)}$ and $H_{(2)}$.
- Alpha saved for Family 2 is:
  a) All α when in Family 1 all null hypotheses are rejected
  b) $\alpha - 2c_2 = (1 - \gamma)\alpha/3$ when in Family 1 $H_{(1)}$ is rejected but $H_{(2)}$ and $H_{(3)}$ are retained
  c) $\alpha - c_3 = 2(1 - \gamma)\alpha/3$ when in Family 1 both $H_{(1)}$ and $H_{(2)}$ are rejected but $H_{(3)}$ is retained

An illustrative example

- Consider treatment-to control comparisons on three endpoints in the primary family with the control of alpha at the 0.05 level.
  - Test critical values for the conventional Holm are: 0.05/3, 0.05/2, and 0.05 , and those for the equally weighted Bonferroni method are 0.05/3, same for each comparison
  - The endpoint-specific alpha levels for the truncated Holm with a "truncation fraction" of f =1/2 are*:*

    $\alpha_1 = (0.05/3)f + (0.05/3)(1\text{-}f) = 0.0167$

    $\alpha_2 = (0.05/2)f + (0.05/3)(1\text{-}f) = 0.0208$

    $\alpha_3 = (0.05)f + (0.05/3)(1\text{-}f) = 0.0333$

An illustrative example (cont'd)

- The unused alphas for passing to secondary family are:

(i) 0.05 if all three tests are successful

(ii) $(0.05 - \alpha_3) = 0.05 – 0.0333 = 0.0167$, if the 1st two tests are successful but the last one is not

(iii) $(0.05 - 2\ \alpha_2) = 0.05 – 2(0.0208) = 0.0084$, if the 1st test is sucessful but the other two are not.

Three gatekeeping approaches
(Dmitrienko et al., 2013)

Family 1

H1, H2

Family 2

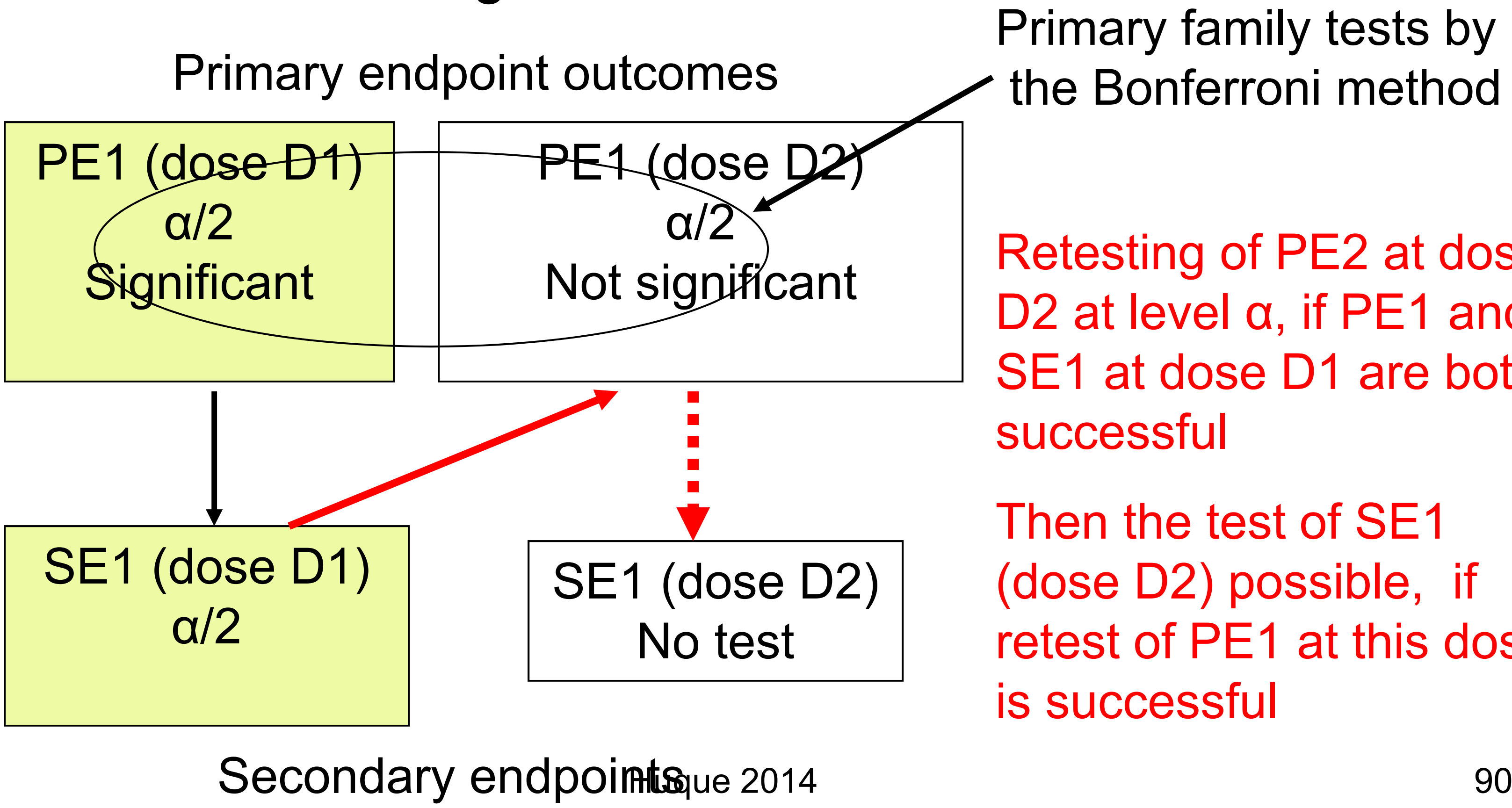
H3, H4

- Regular gatekeeping strategy:
  - Family 1 tests are “independent” of the of the tests in Family 2.
- Gatekeeping with re-testing
  - Hypotheses in Family 1 can be retested after positive outcomes of tests in Family 2
- Gatekeeping with simultaneous testing

Gatekeeping tests w. re-testing

Consider 2 endpoints (PE1, SE1) and two dose levels

Logical restriction: SE1 at dose D2 can not be tested if PE1 at this dose is not significant



Primary endpoint outcomes

Primary family tests by the Bonferroni method

PE1 (dose D1)
α/2
Significant

PE1 (dose D2)
α/2
Not significant

SE1 (dose D1)
α/2

SE1 (dose D2)
No test

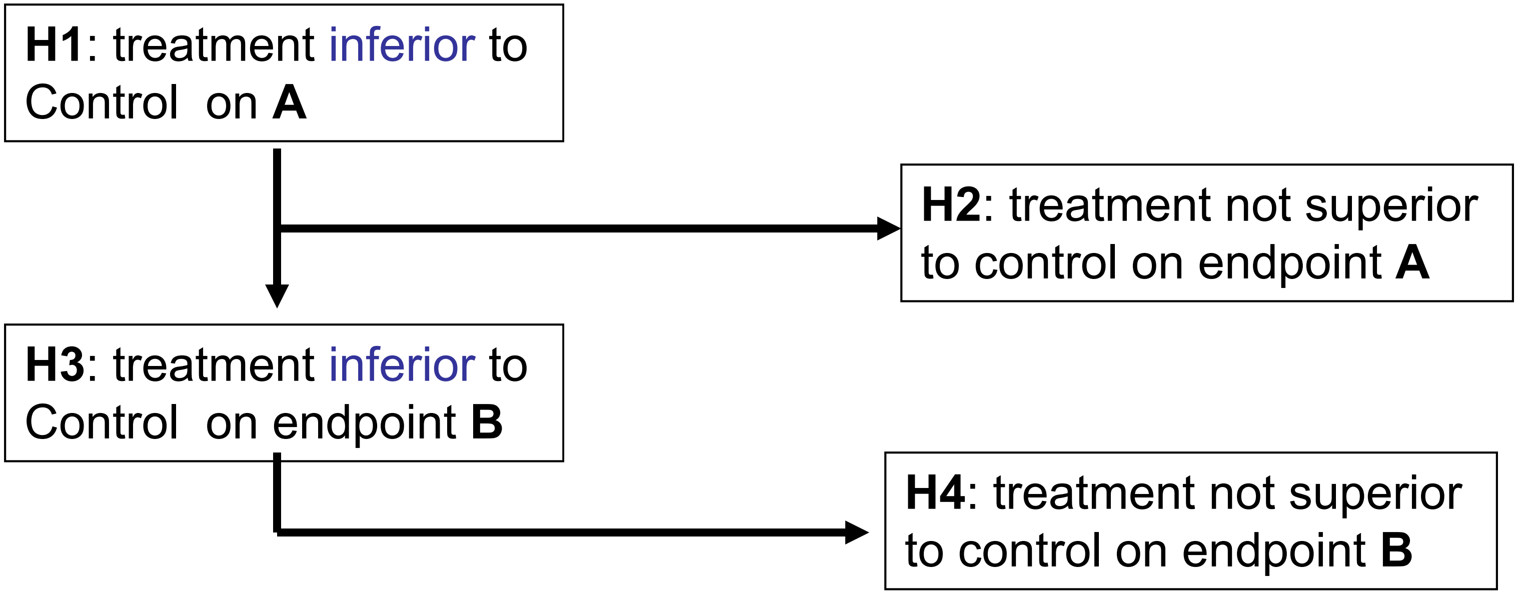
Secondary endpoints

Retesting of PE2 at dose D2 at level α, if PE1 and SE1 at dose D1 are both successful

Then the test of SE1 (dose D2) possible, if retest of PE1 at this dose is successful

Gatekeeping tests w. logical restrictions

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



H1: treatment inferior to Control on **A**

H2: treatment not superior to control on endpoint **A**

H3: treatment inferior to Control on endpoint **B**

H4: treatment not superior to control on endpoint **B**

Test strategy (hierarchical):

If H1 is rejected then test for the family {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 the non-inferiority test by the same 2-sided 95% confidence interval that establishes NI

A simple proof of inflation if each test at 0.025 (1-sided)

- 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 U 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 (E1) = 0.025 + ε > 0.025

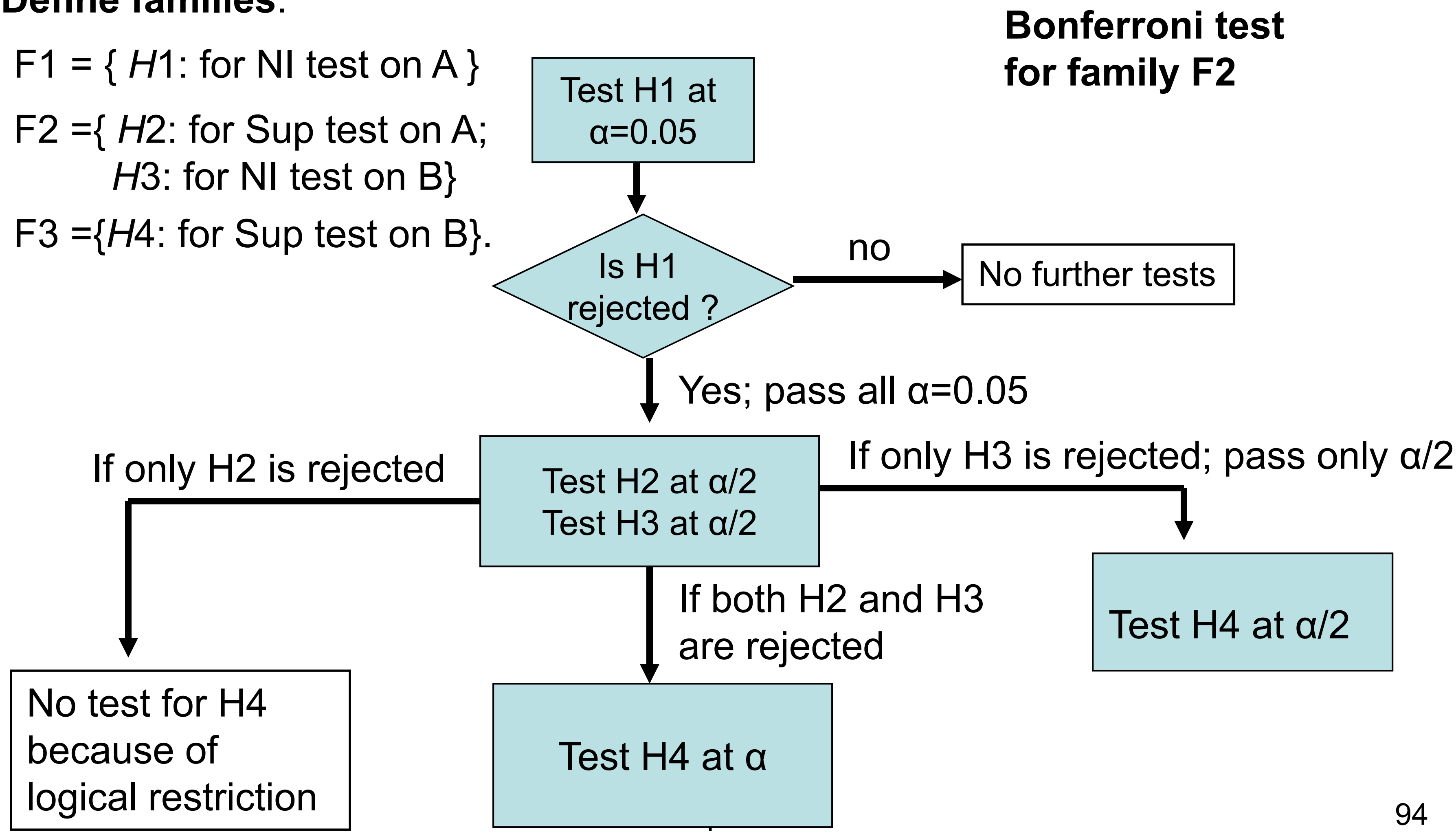
Solution by the Bonferroni gatekeeping method

Define families:

F1 = { *H*1: for NI test on A }

F2 ={ *H*2: for Sup test on A;
*H*3: for NI test on B}

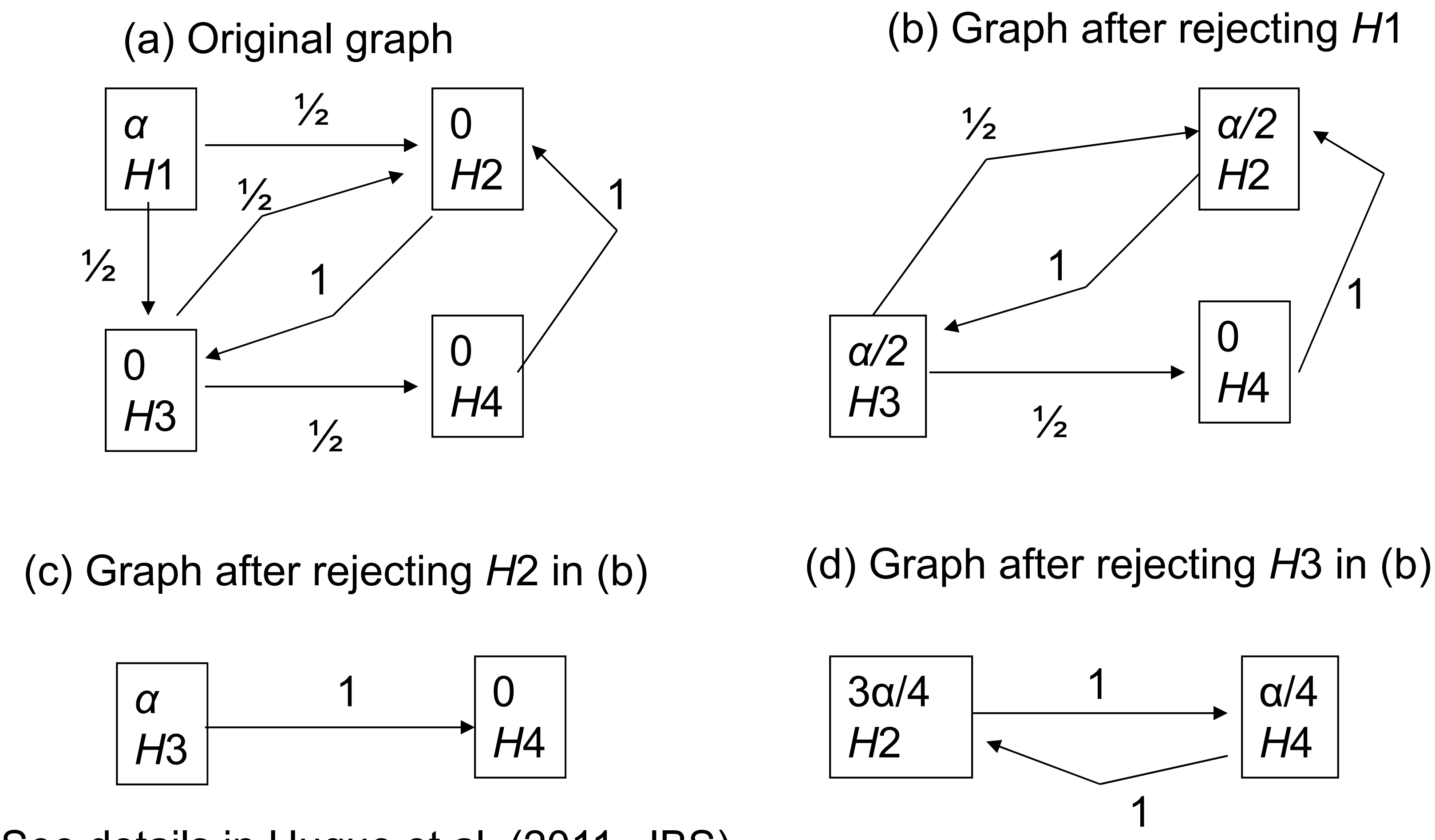
F3 ={*H*4: for Sup test on B}.



Bonferroni test for family F2

Solution by the graphical method

Bretz et al. (2009)


(a) Original graph

(b) Graph after rejecting *H*1

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

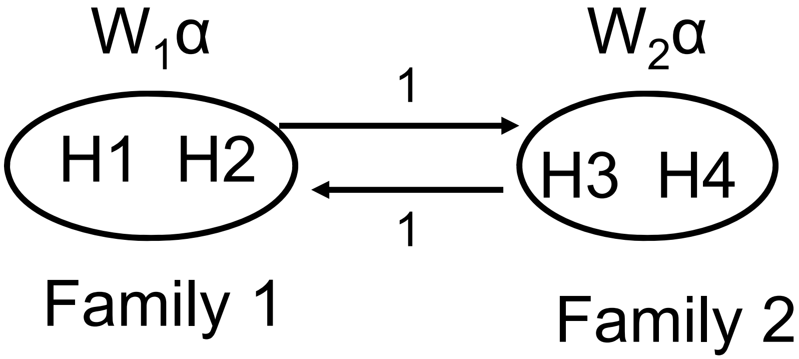
(d) Graph after rejecting *H*3 in (b)

See details in Huque et al. (2011, JBS)

Benefits of the Bonferroni or Bonferroni-based methods

- Simple to explain to non-statisticians
- A finding that survives a Bonferroni adjustment is generally considered a credible trial outcome
- Complex gatekeeping methods simplifies to simple useful shortcut methods.
- Its critical values can combine with the critical values of alpha-exhaustive methods (e.g., Holm’s) leading to (truncated) tests with more power for the primary family
- Confidence intervals computation possible. (Very much needed for benefit-risk assessments)
- Etc.

Gatekeeping approach w. simultaneous testing



$W_1\alpha$

H1 H2

Family 1

1

1

$W_2\alpha$

H3 H4

Family 2

OPP = Overall patient populations

TS = targeted subgroup

$W_1 + W_2 = 1$

H1 for (OPP, High Dose)

H2 for (TS, High Dose)

H3 = (OPP, Lower Dose)

H4 = (TS, Lower Dose)

Kordzakhia and Dmitrienko (2012) provide solution to this problem on using the Truncated Hochberg Procedure under Super Chain Procedures

Use of resampling methods for endpoints with high correlations (e.g. ≥ 0.60)

- A popular 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, …, m: Rejects $H_{(j)}$ associated with $p_{(i)}$ if

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

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

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

  ✓ 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 resampling methods for primary comparisons of a confirmatory trials

- Results approximate, requiring large sample sizes and usually simulations are required to validate the results
- Computation can be difficult (e.g., for time-to-event endpoints)
- Strong control of the overall type I error rate is achieved under the assumption of <u>subset pivotality condition</u> - hard to justify for some cases.
- Permutation based methods also require assumptions.
- Ref:
  - Westfall and Troendle (2008; *multiple testing with minimal assumptions*)
  - Westfall and Young (1993): Resampling based multiple testing
  - Huang et al. (2006; Bioinformatics; *permute or not to permute*)

Additional topics

- Design and analysis issues for trials with event type composite endpoints
- Sample size issues for co-primary endpoint trials
- Subgroup analyses issues for confirmatory trials

Widespread use of composite event endpoints as PEs in clinical trials – some examples

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)

PLATO (NENGL J MED 2009; 361(11): 1045-1057): (death from vascular causes, MI, or stroke)

Setting for an event type composite endpoint as a PE

- Manifestation of the disease is complex and cannot be captured by a single PE
- Multiple clinically relevant PEs (presumably coherent with each other) are needed to capture the disease
- Frequency of individual endpoints in the composite is small, and one expects only a small treatment effect contributed by each component.

Advantages

- Enables better interpretation of results in the presence of complex manifestations of the disease
- Reduces the size of the trial and enables the trial to be done in a reasonable timeframe, when expecting only small treatment effects in each component of the composite
  - If individual components were tested simultaneously (e.g., by the Bonferroni test), when expecting only small treatment effects in each component, then such an approach would not be practical.

Example of sample size reduction

Alpha = 0.025, 1-sided

Endpoints	Control Group Rate	#Treatment Group Rate	Efficacy	Power	Sample size/arm
Mortality	15%	13.2%	1.80%	80%	5865
Hospitalization	19%	16.72%	2.28%	80%	4427
Composite event rates	**34%**	**29.92%**	**4.08%**	**80%**	**2047**

#NOTE: 12% reduction from the control rate

Disadvantages

- Efficacy results of component endpoints may not go in the same direction, causing difficulties in interpretation of the composite endpoint result
- Overall result for the composite can be driven by the less important components which may occur more frequently
  - **when there are null or near null results for important components, or worse, when the results for important components are in the opposite direction.**

Example: Un-interpretable trial results?

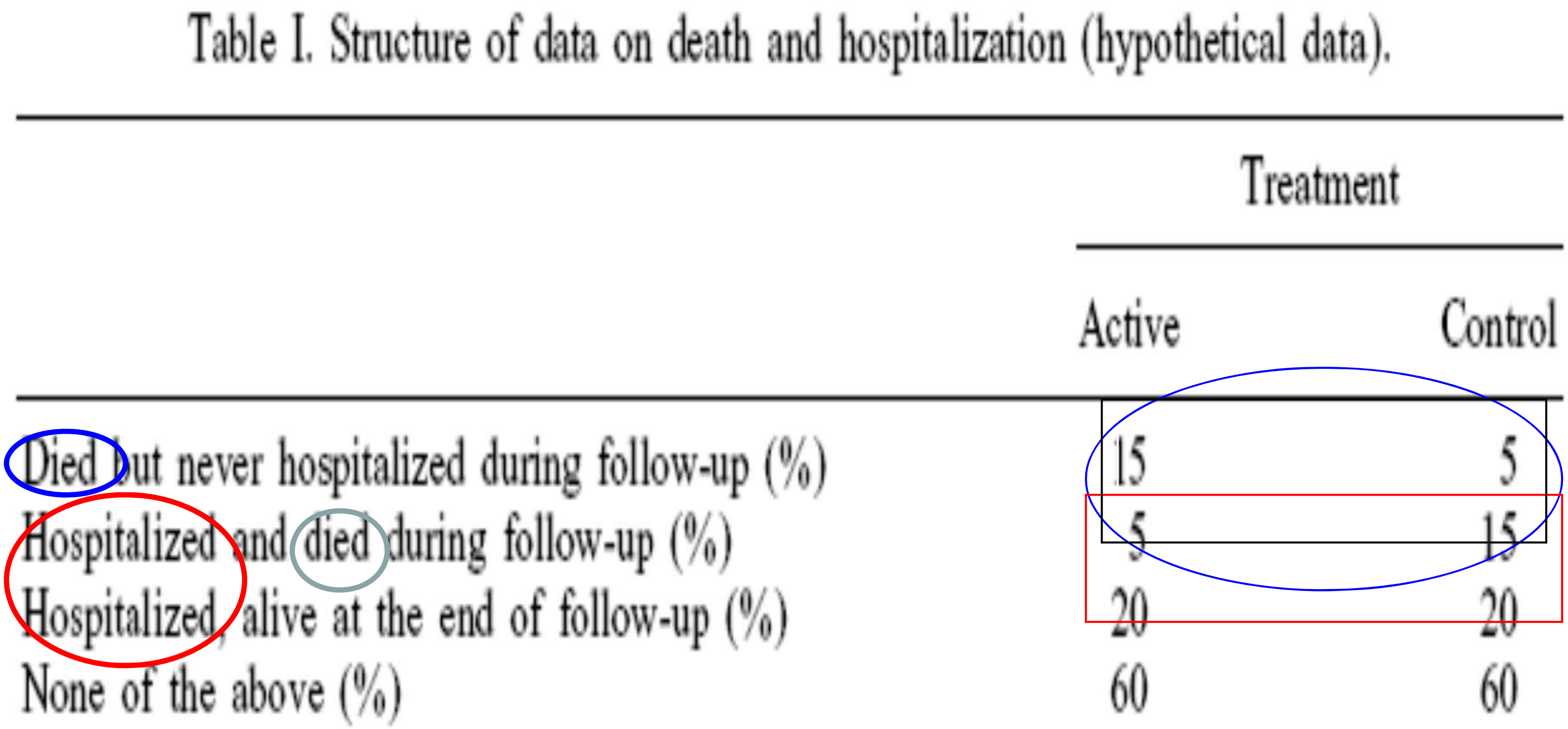


Table I. Structure of data on death and hospitalization (hypothetical data).

	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)

How to address these issues?

- For the purpose of interpreting the composite endpoint result , analysis needs to be done for the composite endpoint
  - With full disclosure of data of individual components, along with relevant analyses of each component, showing appropriate descriptive statistics and graphs
- For important endpoints, such as death, study needs to be designed and statistical testing methodology needs to ensure that there is at least a favorable trend in such endpoints for interpreting the study findings.

Some statistical methods that can address these issues

Methods based on scores

- **A-HeFT trial** (African-American Heart Failure Trial; *J. of Cardiac Failure* 2002)
- Hallstrom et al. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Controlled Clinical Trials* 1992

✓ **Good idea for some situations**

✓ **Scores need to be prospectively defined**

✓ **Difficulties and issues may arise in defining scoring criteria that are clinically meaningful**

Composite Scoring System for A-HeFT

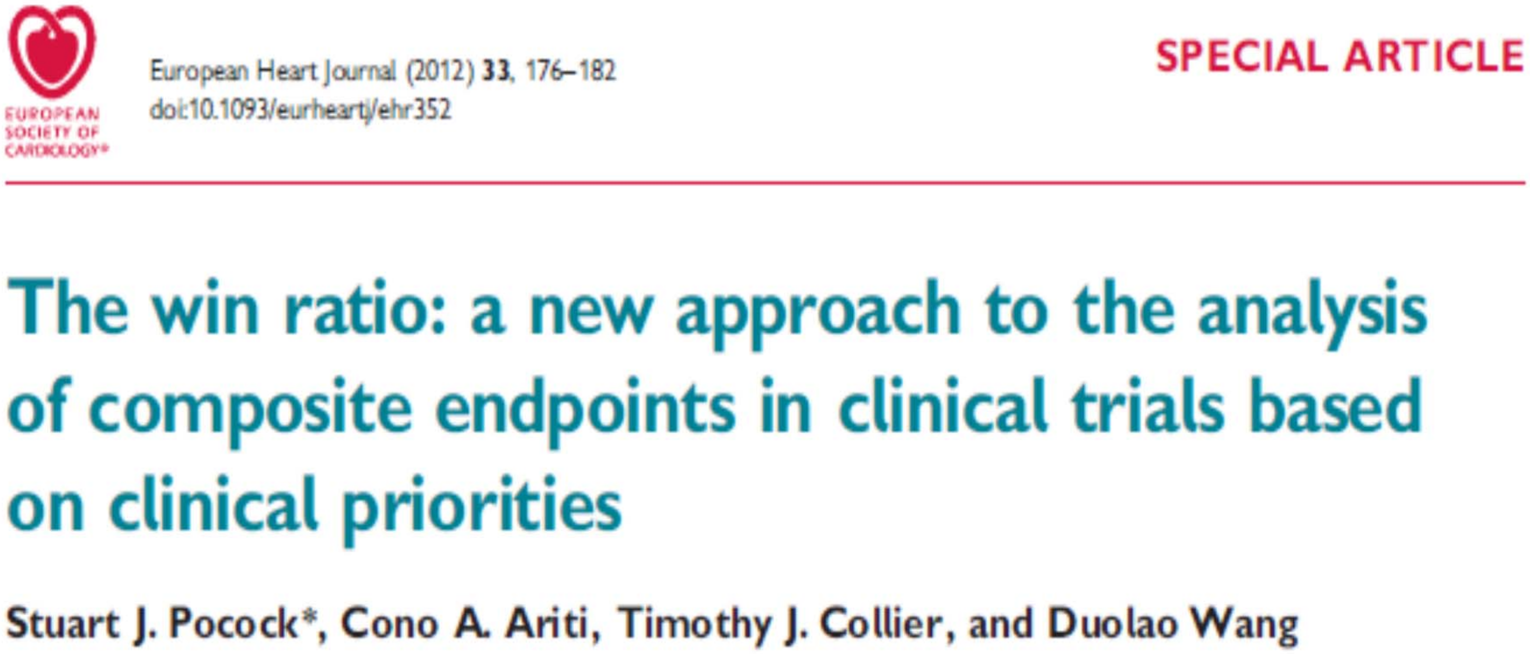
J. of Cardiac Failure 2002; 8(3)

Endpoint scored	Criteria	Score
Death	Death from any cause anytime during trial	-3
	Alive at end of trial	0
Hospitalization	First hospitalization for heart failure	-1
	No hospitalization for heart failure	0
Change in Quality of Life at 6 months*	Reduction by 10 or more units =markedly improved	+2
	Reduction by 5 to 9 units = improved	+1
	Changed by 4 to 4 units = no change	0
	Increased by 5 to 9 units = worsened	-1
	Increased by 10 or more units = markedly worsened	-2

*NOTE: The Minnesota Living with Heart Failure Quality of Life instrument provides a measurement in units. By convention, an increase in units indicates worsening and a reduction, improvement

Methods that consider prioritized clinical outcomes

✓ Pocock et al. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal* 2012



EUROPEAN SOCIETY OF CARDIOLOGY®

European Heart Journal (2012) **33**, 176–182
doi:10.1093/eurheartj/ehr352

SPECIAL ARTICLE

The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities

Stuart J. Pocock*, Cono A. Ariti, Timothy J. Collier, and Duolao Wang

The method accounts for clinical priorities, i.e, CV deaths are considered more important than non-fatal events.

Statistical properties?

The "favor" function

- Consider a CHF trial that compares a treatment (*T*) to control (*C*) using a composite of death and hospitalization

    i =1, …, *n* subjects assigned to treatment *T*

    j =1, …, *m* subjects assigned to *C*

- Consider a pair (*i, j*) where subject *i* belongs to treated group and subject j to the control group then a favor function u_{ij} is defined as in the next slide

The “favor” function (cont’d)

Assign

a)u$_{ij}$=+1: (i) : If *T* favors *C* for death, or (ii) if *T* vs. C comparison NI/NU for death, then *T* favors *C* on hospitalization

b)u$_{ij}$=-1: (i) : If instead *C* favors *T* for death, or (ii) if *C* vs *T* comparison NI/NU for death, then *C* favors *T* on hospitalization

c)u$_{ij}$= 0: (i) : If *C* vs *T* comparisons NI/NU on both death and hospitalization

NI = non-informative, NU = neutral

Win ratio (Pocock et al.; 2012)

Test statistics for treatment difference:

$$\hat{\Delta} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ij}, \quad \Delta = E(\hat{\Delta})$$

Construct *n* matched pairs for the trial:

Win Ratio = $E\left(\sum_{i=1}^{n} 1_{u_{ii}=1}\right) / \left(\sum_{i=1}^{n} 1_{u_{ii}=-1}\right)$

$1_{u_{ii}=*}$ Is an indicator function to show if the comparison for the ith matched pair is positive, negative or neutral

A recent article by Rauch et al. (*Stat in Med* 2014)

- If one assumes for simplicity: 1) components are independent, 2) each time-to- event endpoint follows exponential density, and 3) each patient followed to time f_0. Then

Δ = w_1* (trt. eff. for PE1) + w_2* (trt. eff. for PE2)

where: $w_1 + w_2 \neq 1$ and influenced by the follow-up time f_0

Other methods

- Methods that can rule out treatment harm on a serious component such as death
  - Rauch et al. Consistency-adjusted alpha allocation methods for a time-to-event analysis of composite endpoint. *Computational Statistics and Data Analysis* 2014
  - Röhmel et al. On testing simultaneously non-inferiority in two primary endpoints and superiority in at least one of them. *Biom. J.* 2006
  - Huque and Alosh. A consistency adjusted strategy for accmodating an underpowered endpoint. *J. of Biopharm. Statistics* 2012

Sample size issues for trials with co-primary endpoints

Regulatory definition of co-primary endpoints:

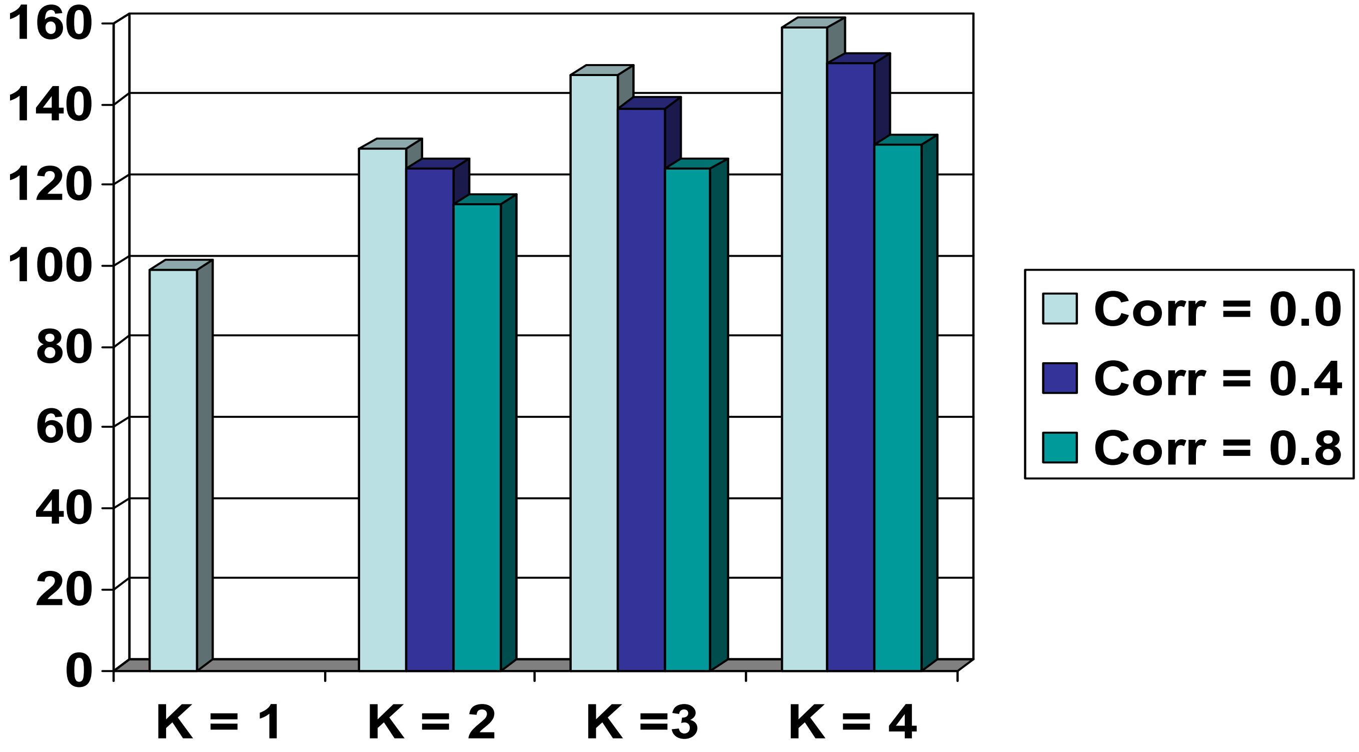
- Two or more specified primary endpoints are said to be co-primary, if each (individually) has to show statistically significant (that is clinically meaningful) treatment benefit at a pre-specified significance level of alpha (e.g., alpha = 0.05)
- How serious is the sample size problem for multiple co-primary endpoint trials?

Sample size issue for co-primary endpoint trials

- No impact on the Type I error rate
  - OK to test each endpoint at the 0.05 alpha level
- But, impact on the Power (Type II error)
  - Consequence: larger sample size for the trial in comparison to a single primary endpoint trial

Sample sizes/ treatment arm
Moderate+ (plus) effect sizes in all K co-primaries

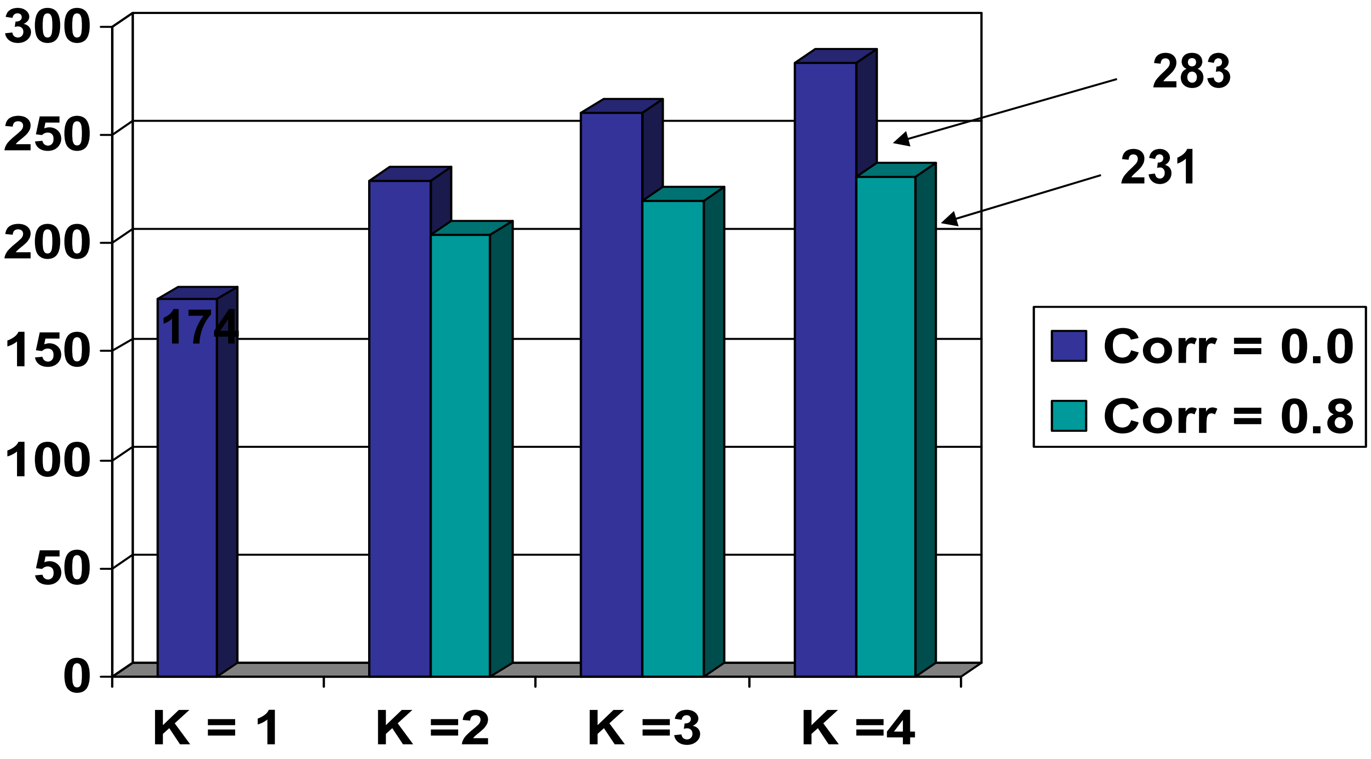
Moderate$^+$ (plus) effect Sizes: (0.40 per unit S.D.)
Power = 80%, α = 0.025 (1-sided)



Calculations using multivariate normal distribution of the test statistics.
Equal pair-wise correlations

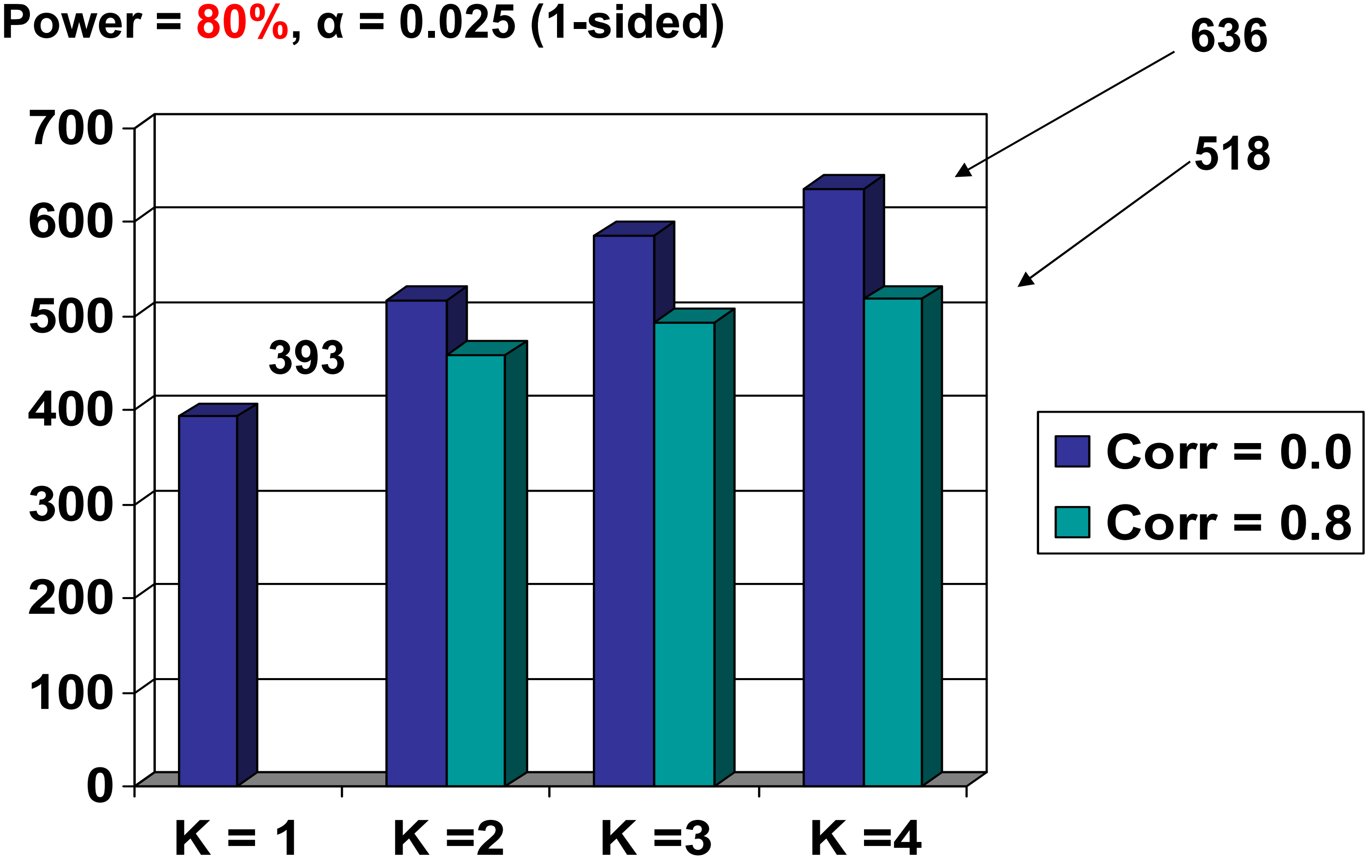
Sample sizes/ treatment arm moderate effect sizes in all K co-primaries

Moderate Effect Sizes: (0.30 per unit S.D.)
Power = 80%, α = 0.025 (1-sided)



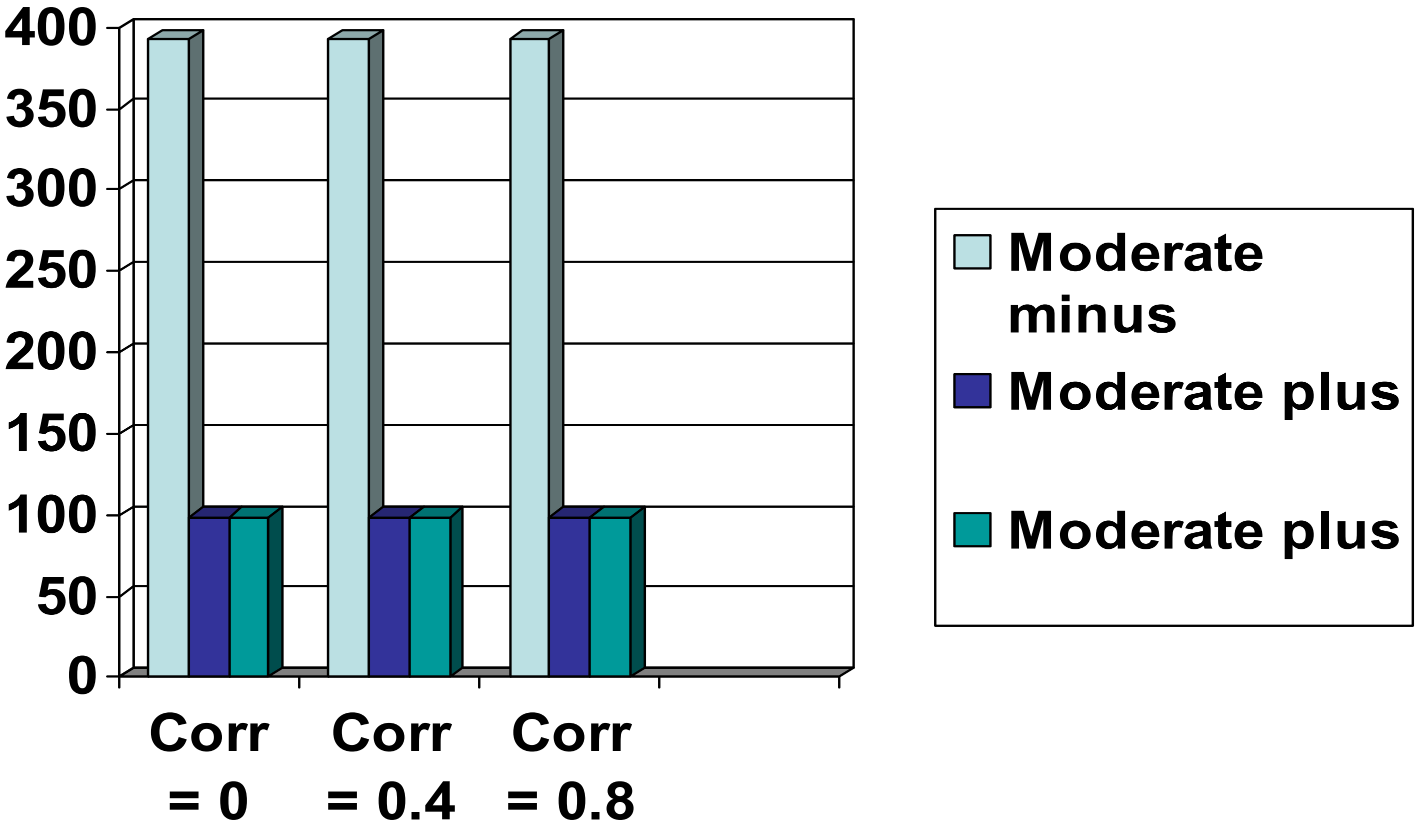
Sample sizes/ treatment arm
Moderate$^-$ (minus) effect sizes in all K co-primaries

Less than Moderate Effect Sizes: (0.2 per unit S.D.)
Power = 80%, α = 0.025 (1-sided)



One of the endpoints has 0.2 per unit S.D. and others have effect sizes of 0.4 per unit S.D.

Trial size = 393/arm, K =3, Power = 80%, α = 0.025 (1-sided)



Huque 2014

Case 1: Migraine trial

Primary efficacy results at 2 hours (ITT)

Endpoint	Treatment N = 364	Placebo N = 360	Treatment Difference	[2]Delta/ S.D.	P-value
Pain Relief	65.1%	28.3%	36.8%	0.66	<0.001
Photophobia free	58.6	36.4	22.2	0.32	<0.001
Phonophobia free	61.3	38.3	23.0	0.33	< 0.001
Nausea free	71.4	64.7	6.7 (0.20[1])	NA	0.007[1]

(1) Analysis adjusted for baseline nausea

(2) Observed delta/S.D.

Huque 2014

Case 2: Alzheimer's trial – mild to moderate disease

Treatment (n = 196), placebo (n = 197)

Primary Endpoints	Treatment Effect Delta/S.D. (observed)	2-sided P-value
CIBIC Endpoint (functions of daily living)	0.32[1] 0.31	0.0013[1] 0.0019
ADAS.Cog	0.33[1] 0.23	0.0012[1] 0.02

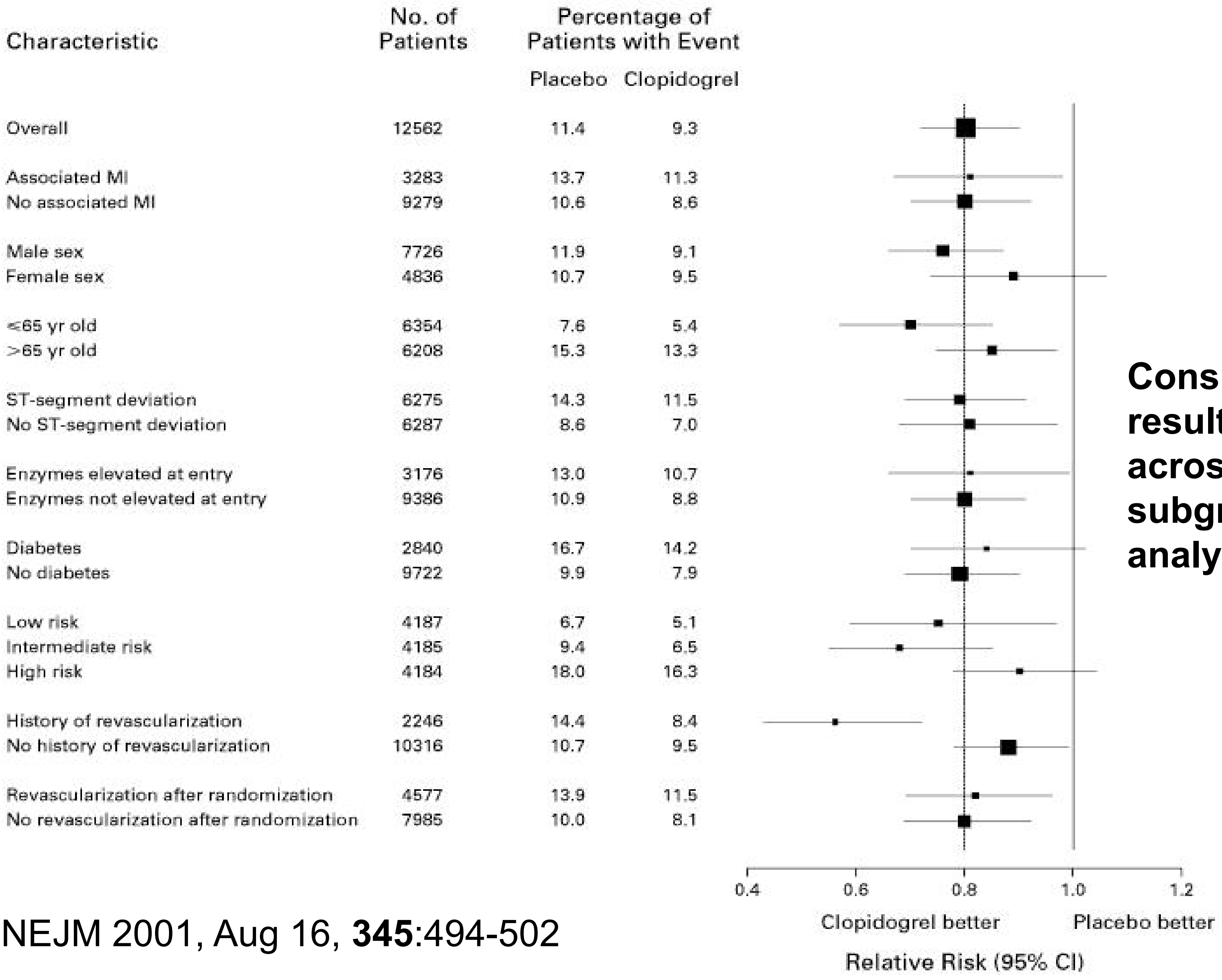
[1] Analysis adjusted for endpoint baseline values

Comments: Co-primary endpoints

- Limit the number of co-primary endpoints to two
  - **if clinically acceptable**
- Use more than two co-primary endpoints:
  - **If clinically necessary to do so** and
  - expected effect sizes are such that trial sample sizes are practical

Roles of subgroup analyses in confirmatory trials

- For claim of efficacy for the overall patient population (OPP) or for targeted subgroups
  - Multiplicity adjustments for Type I error control; sample size considerations for the OPP and for the subgroups
- For adding credibility to the evidence that the treatment is effective for the OPP
  - If the results for the OPP is significant and at the same time results are consistent across various subgroups by baseline and other relevant factors
- For limiting the use of the treatment to a sub-population
  - Examples: Cozaar for stroke; Brilinta for acute coronary syndrome; Valcyte for Cytomegalovirus (CMV) in patients with transplants; Diovan for heartfailure, etc.
- For hypotheses generating



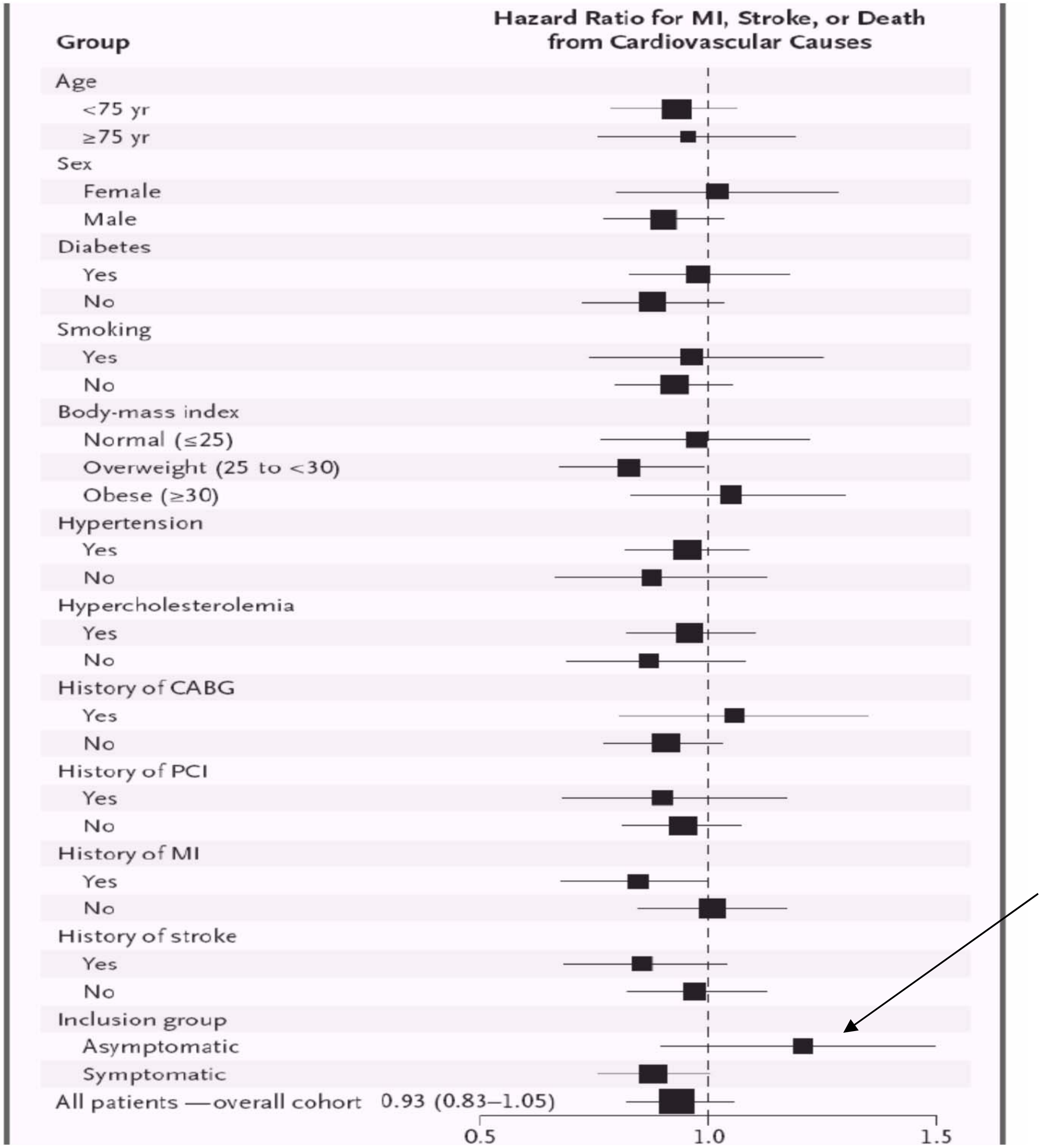
Characteristic	No. of Patients	Percentage of Patients with Event	
		Placebo	Clopidogrel
Overall	12562	11.4	9.3
Associated MI	3283	13.7	11.3
No associated MI	9279	10.6	8.6
Male sex	7726	11.9	9.1
Female sex	4836	10.7	9.5
≤65 yr old	6354	7.6	5.4
>65 yr old	6208	15.3	13.3
ST-segment deviation	6275	14.3	11.5
No ST-segment deviation	6287	8.6	7.0
Enzymes elevated at entry	3176	13.0	10.7
Enzymes not elevated at entry	9386	10.9	8.8
Diabetes	2840	16.7	14.2
No diabetes	9722	9.9	7.9
Low risk	4187	6.7	5.1
Intermediate risk	4185	9.4	6.5
High risk	4184	18.0	16.3
History of revascularization	2246	14.4	8.4
No history of revascularization	10316	10.7	9.5
Revascularization after randomization	4577	13.9	11.5
No revascularization after randomization	7985	10.0	8.1

Consistent results across all subgroups analyzed

NEJM 2001, Aug 16, **345**:494-502

Forest plot of subgroup analyses:

JAMA April 20, 2006
Clopidogrel
+ASA
Vs. ASA



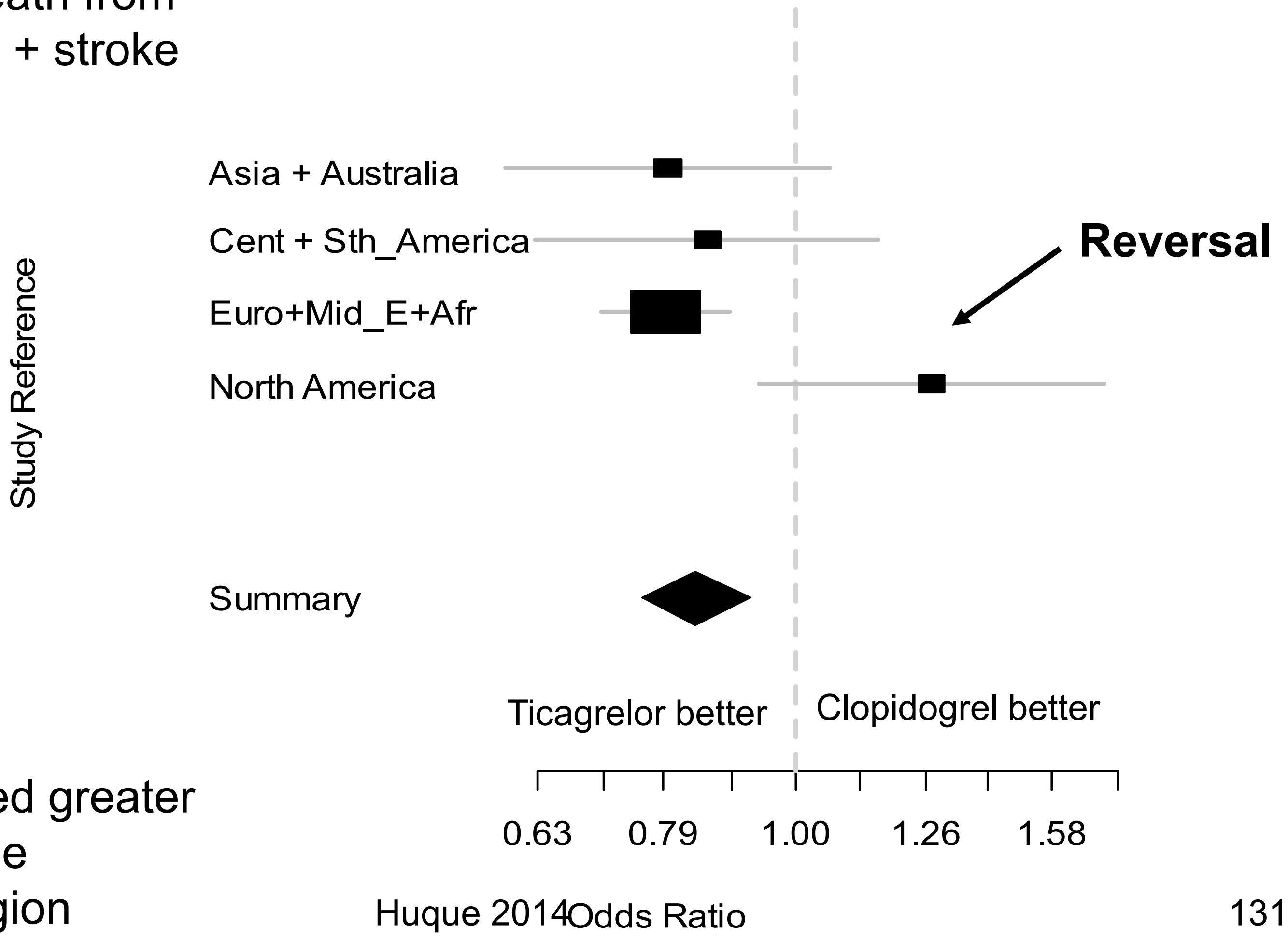
Reversal:

Suggestion of harm in this subgroup

PLATO trial

Ticagrelor vs. Clopidogrel in patients with acute coronary syndrome (NENGL J MED 361(11): 1045-1057, Sep 2009)

PE = composite of death from vascular causes + MI + stroke



Reversal

Note: Data indicated greater use of **aspirin** in the North American region

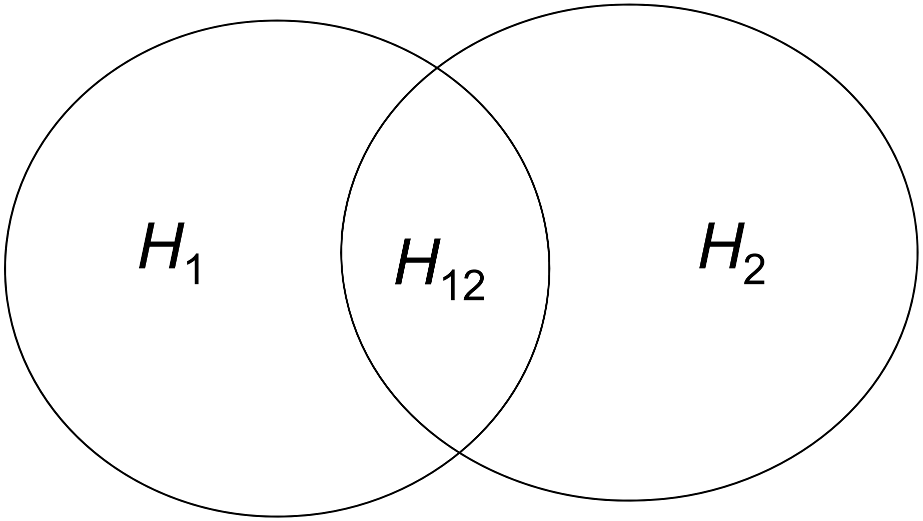
Closed testing and partitioning principles for solving multiplicity problems of clinical trials

Closed testing procedure (CTP)

- Given K elementary hypotheses $H_1, \ldots, H_K$, consider the 2^K -1 intersection hypotheses $H_{\boldsymbol{J}} = \cap_{j \varepsilon \boldsymbol{J}} H_j$ for all non-empty subsets $\boldsymbol{J}$ of $\{1, \ldots, K\}$. Each $H_{\boldsymbol{J}}$ is tested at level α or less.
- An individual null hypothesis H_j is rejected if every intersection hypothesis $H_{\boldsymbol{J}}$ that includes H_j (including H_j itself) is rejected by its local level α test. This controls the FWER in the strong sense at level α (Marcus, Peritz, and Gabriel; 1976).
- A closed testing procedure is α-exhaustive, if the size of each intersection hypothesis test equals α, that is, P(reject $H_{\boldsymbol{J}}$) = α under the null for all subsets $\boldsymbol{J}$ of $\{1, \ldots, K\}$. (Grechanovsky and Hochberg; 1999)

Example: Test of H_1 and H_2 by CTP

Venn diagram for two hypotheses H_1 and H_2 and their intersection H_{12}



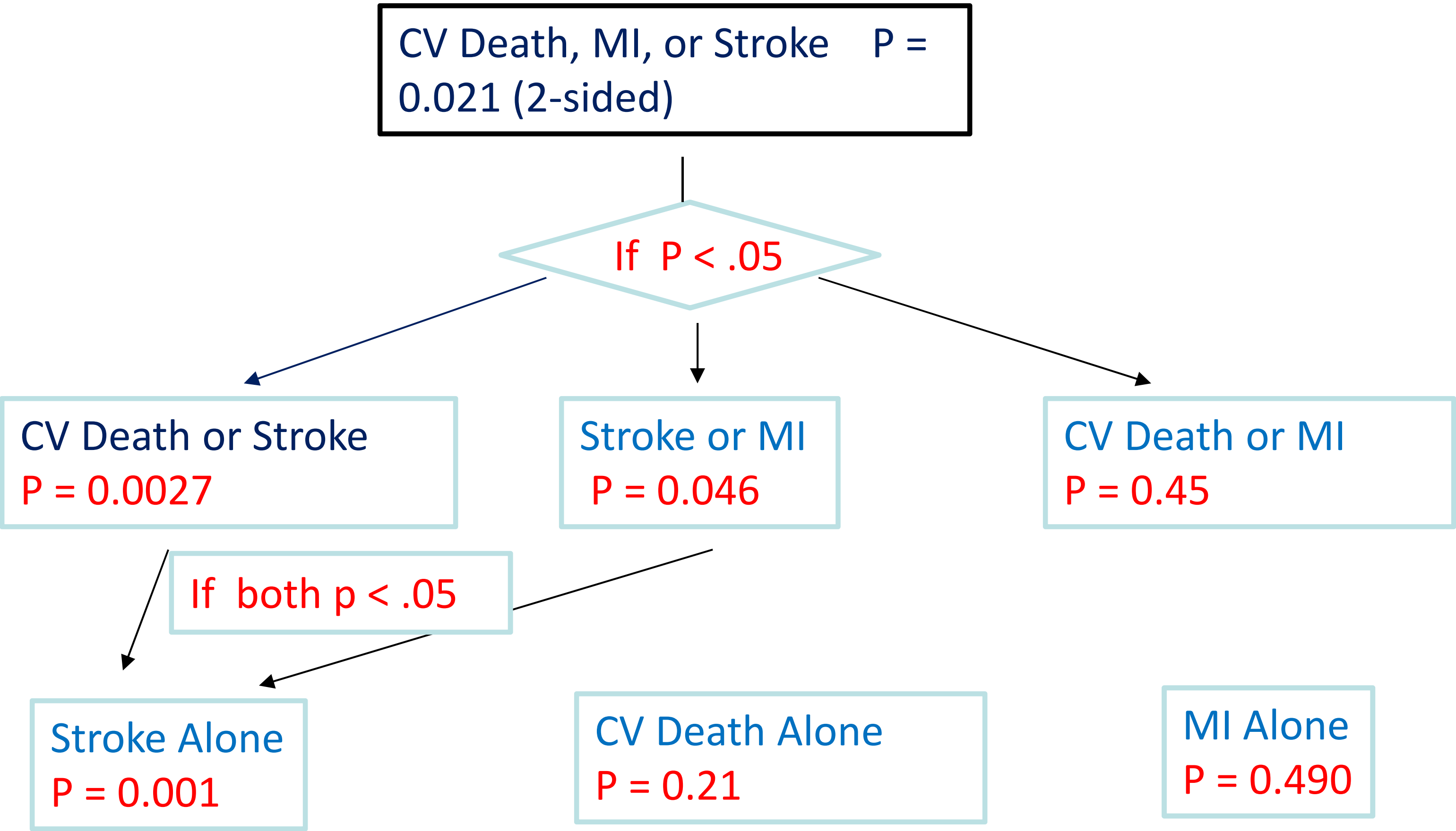
Reject H_1 If:

H_{12} is rejected at level α and also H_1 is rejected at level α

CTPs have good properties

- CTPs are optimal when FWER has to be controlled. (Bauer; SM 1991)
- CTP by construction is coherent (Gabriel; AMS 1969)
  - Because, if H_j is rejected then all intersection hypothesis H_J that includes H_j are rejected as well.
- Coherence avoids interpretation difficulties. E.g., for K = 2, if H_1 is rejected but H_{12} is not, we would have problems in interpreting these results. The CTPs avoids such problems as it implies coherence by construction:
  - It first tests H12, and only if this is rejected, it continues testing the individual null hypotheses.
- Sonnemann and Finner (1988): showed that any non-coherent multiple testing procedure can be replaced by a coherent procedure that never rejects less but may reject more hypotheses. Furthermore, any coherent multiple test controlling the FWER is a closed test.

Closed testing method for explaining the result of acomposite (LIFE trial: Lancet 2002)



CV Death, MI, or Stroke P = 0.021 (2-sided)

If P < .05

CV Death or Stroke
P = 0.0027

Stroke or MI
P = 0.046

CV Death or MI
P = 0.45

If both p < .05

Stroke Alone
P = 0.001

CV Death Alone
P = 0.21

MI Alone
P = 0.490

Partitioning method vs. closed testing

Partitioning method:

- Partition the union of multiple hypotheses into disjoint hypotheses
  - Each disjoint hypothesis can be tested at the same significance level α (e.g., α =0.05)
- A hypothesis H is then rejected if all disjoint hypotheses that overlaps with H are rejected
- Example (next slide)

Example

- Consider partitioning of the union of H_1, H_2 and H_3 into 7 disjoint hypotheses
- Each test T_1 to T_7 is at level α
- H_1 is rejected if A, B, C and D are rejected
- H_2 is rejected if A, B, E and F are rejected
- H_3 rejected if A, C, E, and G are rejected

Disjoint Hypotheses		Test
A:	H_1 H_2 H_3	T_1
B:	H_1 H_2 K_3	T_2
C:	H_1 K_2 H_3	T_3
D:	H_1 K_2 K_3	T_4
E:	K_1 H_2 H_3	T_5
F:	K_1 H_2 K_3	T_6
G:	K_1 K_2 H_3	T_7

K_i is complement of H_i

Partitioning method vs. closed testing

- Standard multiple test procedures derived by the closed testing can also be derived by the partitioning principle
  - Examples: Holm, Dunnett-Tamhane Step-down Procedure, Extended Simes, etc
- This was shown by Pär Karlsson (BASS 2010)

Closing Remarks

- Modern clinical trials often include complex designs for better characterization of risk–benefit profiles of study drugs.
  - In this regard, trials include multiple objectives of different importance seeking answers to a number of specified questions. These answers are generally derived from the results of the planned multiple comparisons of the new treatment to control on multiple primary and secondary endpoints.
  - However, whether the answers to these questions can lead to clinically meaningful benefits of the new treatments is determined by multiple win criteria which introduce multiplicity.

Closing Remarks (cont'd)

- Addressing multiplicity for these trials may require using advanced statistical approaches some of which have appeared only in recent publications.
- Confirmatory trials prospectively plan statistical strategy for addressing multiplicity using methods that are valid and efficient in answering questions of clinical importance.
- In this regard, regulatory agencies generally require an upfront statistical analysis plan (SAP).

Closing Remarks (cont'd)

- There are many considerations for addressing multiplicity in confirmatory trials, but the following three are of paramount importance:
  1. Defining upfront the clinical win criteria of the trial clearly. Statistical results, though significant for certain comparisons, may not have clinical utility, if they do not fit into the clinically specified win criteria of the trial.
  2. Adhering to the principle of prospective planning. The trial may lack validity if the multiplicity problem and their solutions are not worked out in advance.
  3. Protecting the strong control of the FWER in making claims of treatment benefits for the primary as well for the secondary benefits.

Thank You