Multiplicity problems in clinical trials
A regulatory perspective

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

BASS Conference 2010, Hilton Head, SC,
November 8, 2010

Disclaimer

- This presentation expresses **personal views of the presenter** and does not necessarily present the regulatory views or policy of the FDA
- Notes:
  - A regulatory FDA guidance is under preparation on the topic of multiple endpoints in clinical trials
  - Details of this guidance will be discussed when released for public comments
  - Some multiplicity information: visit http://multxpert.com/
  - Book: Multiplicity Testing Problems In Pharmaceutical Statistics, Edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz, CRC press. (Chapter 1, by Huque and Röhmel)

Regulatory standard for effectiveness of a new drug - ("substantial evidence")

- U.S. Food Drug and Cosmetic Act

  "evidence consisting of adequate and well-controlled investigations, including clinical investigations, … to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports …"

- FDA's interpretation of the statute

  At least two "adequate and well-controlled" trials, each convincing on its own, are required to establish effectiveness

Why at least 2 adequate and well controlled studies?

- Minimizes the possibility of regulatory decisions based on bias driven results
- Reduces the false positive error rate to the level that regulatory decisions are meaningful.
- Ensures consistency of study findings through replication of study findings
- Gives confidence in generalizing results to a larger population

The Food and Drug Amendments Act of 1997 (FDAMA)

- FDAMA amended the Act that FDA may consider
  - "data from one adequate and well-controlled clinical investigation and confirmatory evidence" to constitute substantial evidence
  - if FDA determines that such data and evidence are sufficient to establish effectiveness

FDA guidance following FDAMA

- FDA Guidance for Industry(1998): “Providing Clinical Evidence of Effectiveness for Human Drugs and Biologic Products”
  - Describes circumstances in which FDA may rely on a single trial to demonstrate effectiveness for human drugs and biologic products

Characteristics of a single adequate well-controlled trial to support an effectiveness claim

1. A large multicenter trial in which no single site provided an unusually large fraction of the patients and no single investigator or site was disproportionally responsible for the favorable effect seen
2. Consistency of study findings across key patient subsets (e.g., disease stage, age, gender, race)
3. Presence of multiple studies within a single study, such as occurs in a factorial design, which show consistent findings
4. Persuasive evidence on multiple endpoints
5. A statistically very persuasive finding (2-study worth of evidence)

Commentaries on 1 vs. 2 study issues: One by Professor Gary Koch and the other by Mohammad Huque (SIM 2005; 24: 1639 – 1651)

Clinical trials often face problems of bias and inflation of false positive error rate

- Bias
  - Estimation of treatment effects can appear to be better than actually they are as a result of:

    factors or processes that tend to deviate the results of a trial systematically away from the truth

    E.g., excessive missing data in a trial*
- Bias can be addressed by better study design and conduct of adequate well-controlled trials
- Inflation of false positive error rate can be controlled by addressing "multiplicity"

[*NAS Report on "the prevention and treatment of missing data in clinical trials", 2010];
[Sackett DL: Bias in analytic research, J. Chronic Disease 1979]

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 causes the 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.

Multiplicity in a trial can occur in many situations with differing complexity

- Comparing treatments for more than one endpoint and at different time points
- Comparing several doses of a drug to a control
- Comparing a treatment to control for non-inferiority and superiority on each of several endpoints and doses
- Comparing treatments on multiple primary and secondary endpoints
- Analyzing components of a composite primary endpoint for claiming treatment benefits for one or more of its components (e.g., for the mortality component)
- Performing subgroup analysis for efficacy for the total population and for special subgroups of interest
- Conducting Interim analysis
- Making design modifications
- etc.

Good News

- We have many statistical approaches for addressing different aspects of multiplicity
- There has been remarkable innovations in statistical methodology in dealing with all sorts of multiplicity problems of clinical trials
  - Surprisingly, most of these approaches and methods are fairly recent

Last few years – new useful statistical methods on

- Recycling of alpha from one family to the next (on using Bonferroni and truncated Holm's methods)
- 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., gatekeeping testing schemes
- Lower limit for 1-sided confidence intervals for step-up and step-down procedures
- The fallback and adaptive alpha allocation approaches (the 4A)
- "Partitioning principle" based testing strategies
- Methods for planned subgroup analysis
- Consistency ensured (adaptive) methods
- Others (e.g., related to interim analyses and adaptive designs)

Some key statistical principles/ concepts underlying new methods

- Union-Intersection (UI) and Intersection-Union (IU) testing principles
- Closed testing principle
- Partitioning principle
- Gatekeeping principles
- Graphical concept of transporting alpha from one hypothesis to others (has led to improvements in the fallback methods)
- Concepts and methods for recycling of “unused” alpha from one family to the next
- Adaptive alpha allocation concept

Rest of the presentation on:

- Distinction between primary and secondary endpoints
- Concept of clinical win for efficacy
- When is it necessary to adjust for multiplicity and when is it not?
- Types of FWER control for treatment efficacy claims
- Problems and methods for primary endpoints
- Co-primary (and composite[#]) endpoint issues
- The issue of Type I error rate adjustments for secondary endpoints
- Subgroup analysis
- Other issues (3-arm trials, drug combination trials, and safety data analysis)
- Concluding remarks

[#]Second presentation

Distinction between primary and secondary endpoints

- **Primary endpoints:**
  - 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.
  - These endpoints can either form a single family or multiple hierarchical families depending for example on their relative importance and power considerations, and the win criteria
  - 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

- 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

Efficacy win criteria

- Simply triaging endpoints to primary and secondary is not sufficient.
- The trial should specify a 'win scenario' for the set of primary endpoints that determines whether or not the trial has met its efficacy objectives.
- Examples of efficacy win criteria:

1) All specified primary endpoints needs to show clinically meaningful and statistically significant treatment efficacy
2) At least one of the specified primary endpoints need to show clinically meaningful and statistically significant treatment efficacy
3) A pre-specified subset of primary endpoints need to show clinically meaningful and statistically significant treatment efficacy.

(More examples in Chapter 1 of the book: Multiple testing problems in pharmaceutical statistics; Edts., Dmitrienko, Tamhane and Bretz, 2010, CRC Press),

Can a primary endpoint be called a " secondary" or "key secondary" because of power considerations?

- One or more primary endpoints characterize clinically meaningful benefits of the treatment
- Secondary endpoints by definition do not have this ability in the absence of demonstration of clinically meaningful benefits on one or more primary endpoints
- Calling such a primary endpoint a "secondary" or "key secondary" does not seem appropriate
  - It can still be called a primary endpoint and can take a lower position in the hierarchy of primary endpoints in the gatekeeping framework of primary endpoint families

When multiplicity adjustments are not necessary?

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)

When is it necessary to adjust for multiplicity?

- When the type I error rate inflates as a result of multiple ways to achieve a successful outcome
- Example:
  - CHF trial with 2 PEs (death, hospitalizations)

    Success criterion: superiority of the treatment to control for at least one of the two endpoints;

    Each endpoint tested at the 0.05 level
  - FWER can be as high as 0.0975, an unacceptable trial alpha level for making regulatory decisions.

FWER: Type I error rate concept when testing a family of *m* hypotheses

- Probability of (win by chance: reject at least one true null hypothesis out of the *m* given hypotheses)
  - Which hypotheses are true and which are false are unknown
    - Calculate under the assumption that all hypotheses are true (often known as global or complete null hypothesis)
    - Calculate under all possible null hypotheses configurations and take the maximum
- Familywise error rate (FWER) = max Pr (wining by chance) under all possible null hypotheses configurations (all possible scenarios of true states)
  - Example: (δ_1, δ_2, δ_3) are unknown treatment effect values for the three endpoints
  - Null hypotheses configurations are: (0, 0, 0), (δ_1, 0, 0), (0, δ_2, 0) (0, 0, δ_3), (δ_1, 0, δ_3), (0, δ_2, δ_3), (δ_1, δ_2, 0) for all possible values of deltas and correlations

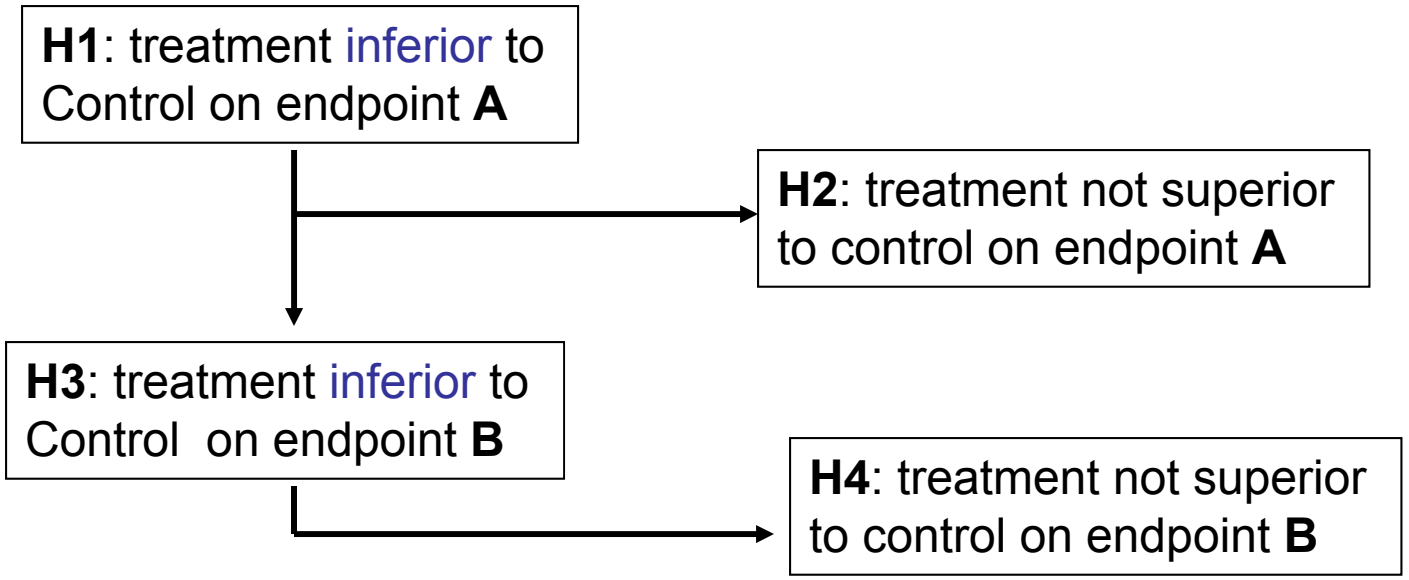
Exercises on FWER calculations

- Exercise 1: A trial compares a treatment to control for efficacy on two primary endpoints A and B.
  - Win criterion: show statistically significant result on at least one of the two endpoints (at-least-one win criterion)
  - Test strategy: test each endpoint at level 0.05;
  - FWER = $1- (0.95)^2 = 0.0975$

  Exercise 2: Same set-up as Exercise 1
  - Win criterion: show statistically significant result at level 0.05 on both endpoints (all-or-none win criterion)
  - FWER = 0.05 x 0.05 (when both null hypotheses true, and tests independent)
  - FWER = 0.05 x power (when one hypothesis is true and the other is false, and tests independent)
  - Maximum (FWER) = 0.05

Exercise 3: FWER = ?

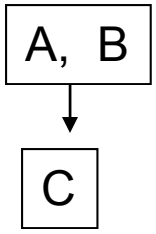
1. **Endpoint A**: Compare treatment **T** to **Control** for non-inferior at level 0.05. If **T** is non-inferior to **Control** then compare **T** to **Control** for superiority at the same 0.05 level
2. **Endpoint B**: Once **T** is at least non-inferior to **Control** on A, proceed to test for **B** in the same manner with each test at the same 0.05 level



Note: this exercise will be addressed in the second presentation

Exercise 4: FWER inflation

- Consider a trial that compares a new treatment to control on endpoints A, B and C
- 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 type 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)

Case to use the truncated Holm's based gatekeeping test strategy (Dmitrienko et al , Biometrical Jr. 2008)

FWER control terms†: "weak" and "strong"?

- Terms confusing even for statisticians without training in multiplicity
- Confusing for clinicians

† Terms defined in the book by Hochberg and Tamhane (1987): Multiple Comparison Procedures, Wily, New York

Weak and strong FWER control approaches differ in critical respects

- Weak FWER control:
  - Control of alpha at level 0.05 for the conclusion that some endpoints (or dose levels), either individually or collectively, have treatment effects. Null hypothesis is global: no effect in any endpoint (or dose level).
  - No intention to identify or to claim which endpoints (or dose levels) have treatment effects (or which win scenario makes it).
- Strong FWER control:
  - Control of erroneously finding a significant result for an endpoint (or dose group) regardless of the size of the treatment effects in other endpoints (or dose groups).
  - Intention: to claim about specific outcomes (e.g., which specific endpoints or which dose groups have treatment effects)

Regulatory applications

- Generally, require strong FWER control for the primary as well as secondary families
  - Except perhaps in rare situations for serious diseases, or for special situations, when weak FWER control may be OK
  - E.g., “treatment of stroke” trials; Tilley et al., 1996)
  - E.g., test for positive slope (of dose response) in a multi-dose trial without placebo

Which analysis methods 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

Graphical approach (Bretz et al., 2009)

- A useful tool
- Makes the test strategy transparent
- Easy to communicate to clinicians as to how alpha adjustments are taking place in a test strategy
- Potentially useful at the planning stage
  - Easy to create different versions of a graph for creating different test strategies and then selecting the one that is most tailored to the objectives of the trial

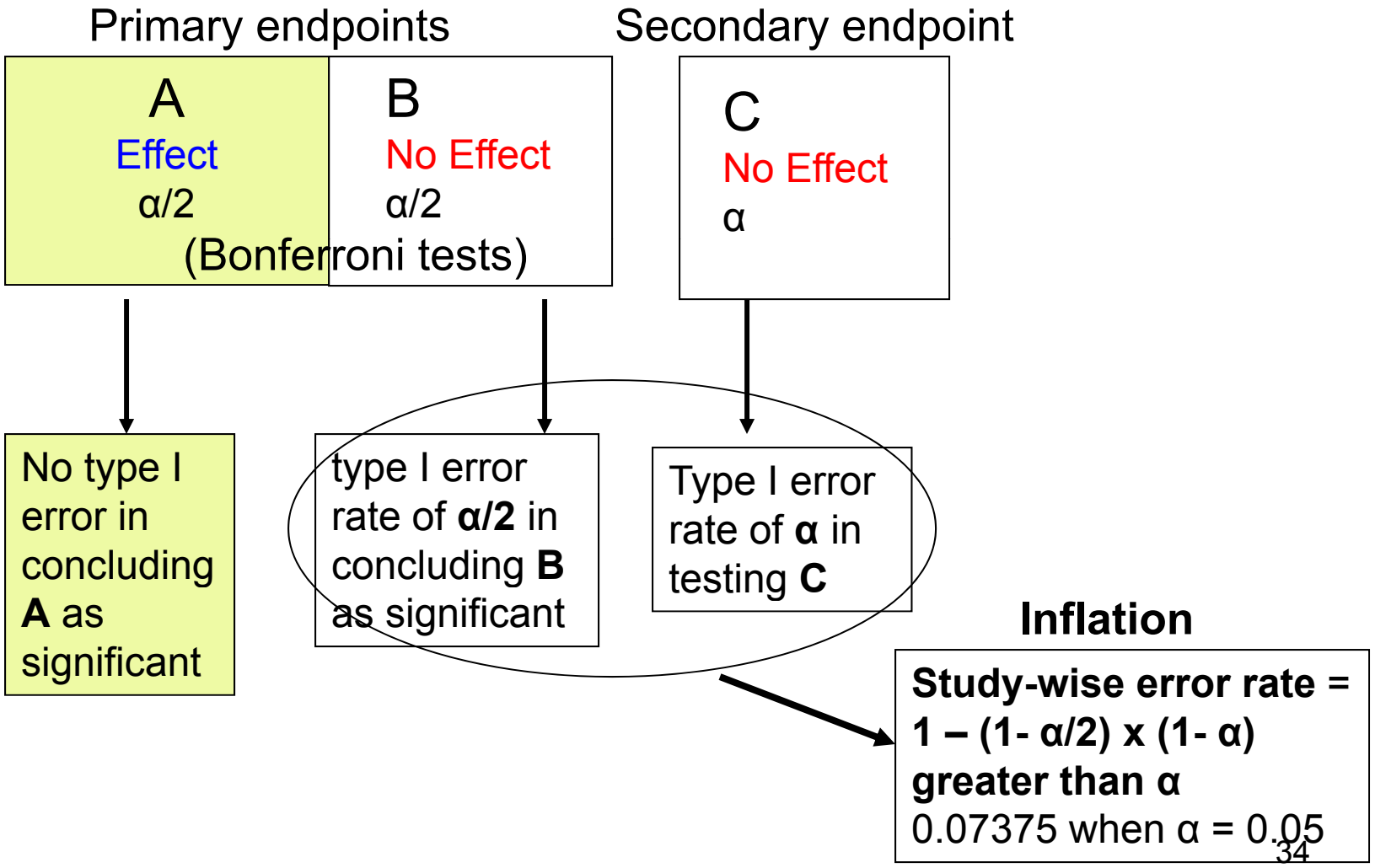
The issue of FWER control for the primary and secondary endpoint families

- Should there be a separate FWER control for the primary endpoint family and separate FWER control for the secondary endpoint family with the condition that

  the secondary endpoint family is to be tested only when one or more primary endpoints shows statistically significant and clinically meaningful results in a manner that the treatment can be indicated for use in the patient population studied?

- For example:
  - Allocate alpha = 0.05 for the primary endpoint family
  - Allocate separate alpha = 0.05 for the secondary endpoint family
  - Test secondary endpoints only after statistically significant and clinically meaningful evidence of treatment benefit on one or more primary endpoints

Issue: Inflation of the study-wise error rate



Family of primary endpoints

Family of secondary endpoints

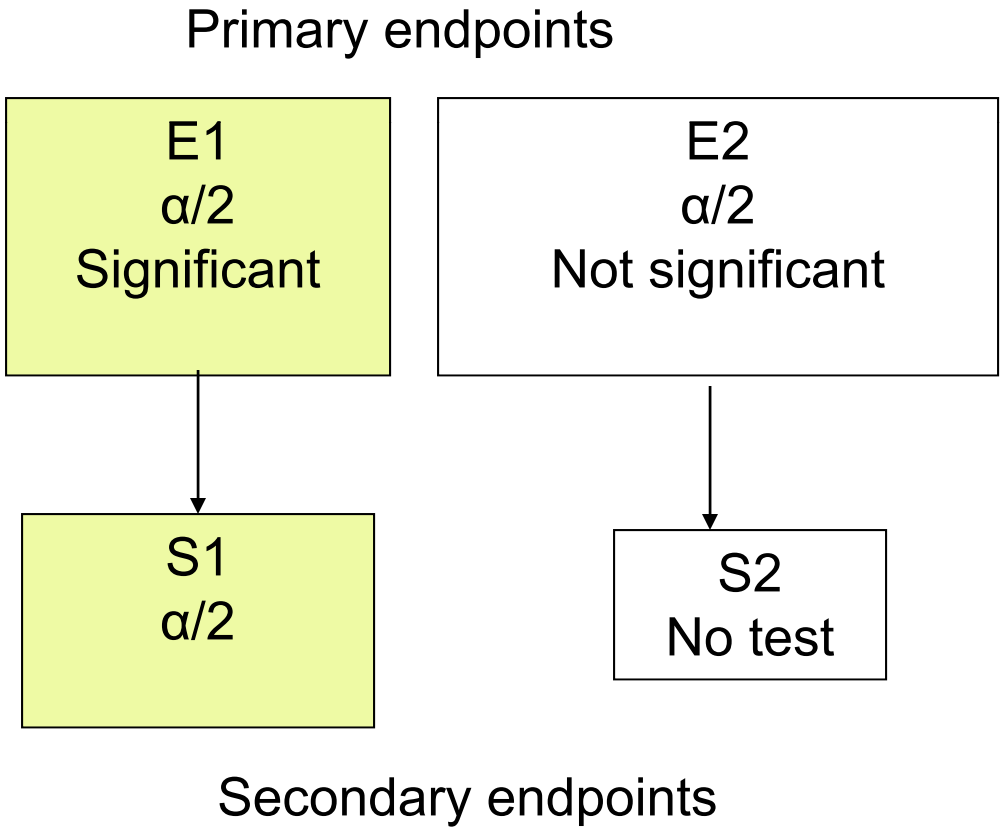
FWER if **locally** controlled at level α by the Bonferroni method

FWER **locally** controlled at level α, e.g., by the Bonferroni or the Holm's method

- **Study-wise error rate is not controlled at level α**, unless use Bonferroni gatekeeping principle[#] (or other gatekeeping principle, e.g., based on truncated Holm's method) is followed, i.e.,
  - [#]Transfer to the secondary endpoint family: α – (sum of alphas of those primary endpoints which are not significant)

Note: Method of transferring alpha are different for other methods

Issue: A secondary endpoint can not be tested if its logically related PE is not significant



Primary family tests by the Bonferroni method

Bonferroni or the Bonferroni base methods

- Not all that conservative
  - When the number of endpoints *m* = 2 to 5, and correlation = 0.3 or less
- Type II error for some situations can be small:
  - if success criterion is to win in at least one of the *m* endpoints.
  - Example (2-arm trial, 2 endpoints):
    - Consider a single endpoint trial: alpha = 0.025, test = 1-sided Z-test, **power = 90%,** and delta (per unit s.d.) = 0.5, then *n* = 84 per treatment arm.
    - Consider a 2-endpoint trial, each endpoint test at level alpha = 0.025/2 = 0.0125, delta1 = delta 2 =0.5, *r* = 0.6, assume *n* = 84 per treatment arm, then

    Power (win in at least one of the two endpoints) = **92.7%**

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 easy. (Very much needed for benefit-risk assessments)
- Etc.

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

- A popular a resampling based step-down procedure:

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

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

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

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

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

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

  ✓ 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 subset pivotality condition - hard to justify for some cases
- Ref:
  - Westfall and Troendle (2008; *multiple testing with minimal assumptions*);
  - Huang et al. (2006; Bioinformatics; *permute or not to permute*)

Co-primary endpoints

- Regulatory requirement:
  - Test each endpoint at 0.05 level to control FWER at the 0.05 level
- Inflation of the type II error recognized.
  - Limit the number of co-primary endpoints to 2 for the claimed indication **(if clinically acceptable)**

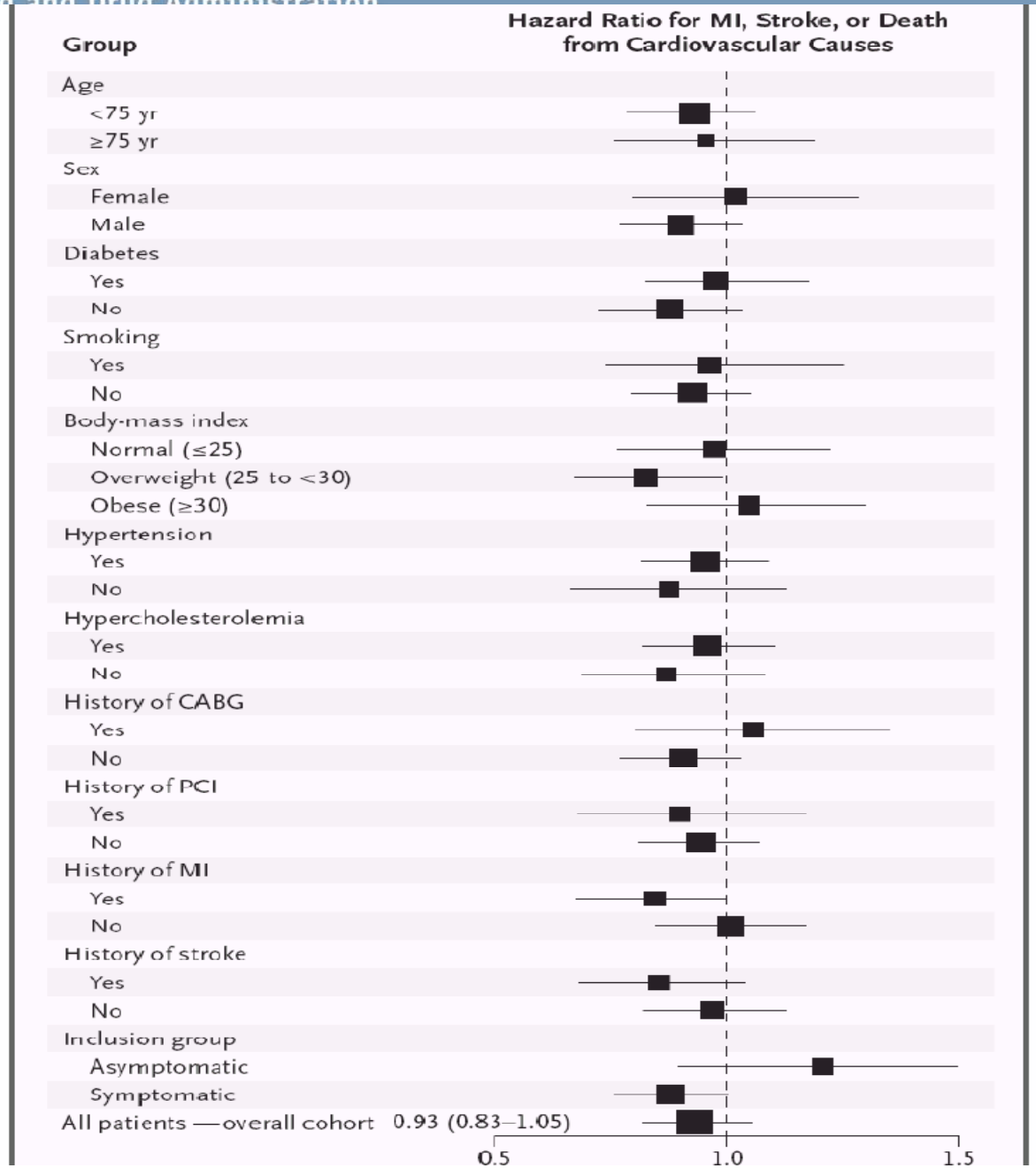
Co-primary endpoints (cont'd)

- More than two co-primary endpoints:
  - When clinically necessary to do so
  - Expected effect sizes are such that trial sample sizes are practical.
  - Cases of strong treatment effects in some (e.g., p-value < 0.01), but weak in some (i.e., p-values slightly > 0.05):
    - OK on case-by-case basis if replicated evidence or presence of other clinical evidence.

Subgroup analyses quite prevalent and considered necessary for clinical trials

- Regulatory guidance (FDA & European) expect some analyses by gender, race and age for Phase III trials
- Forest plots:
  - Common for visual display for showing consistency of results across various subgroups and baseline factors
  - These plots have statistical issues:
    - Treatment comparison result within a subgroup needs to be adjusted for imbalances in other variables, similar to those done in epidemiologic studies

Forest plot of subgroup analyses: Very common in clinical trials



Unplanned subgroup analysis – (Potential Limitations)

- P-value difficult to interpret
  - confounding, bias and multiplicity issues
- Type I error rate un-tractable; usually very high
- Steve Lagakos (*N Eng J Med* 2006, April 20):

  "Subgroup analyses ….. are commonly over interpreted and can lead to further research that is misguided or, worse, to suboptimal patient care."
- Peter Sleight (Current Control Trial Cardiovasc. Med. 2000 1(1): 25-27)

  "Subgroup analyses in clinical trials: fun to look at- but don't believe them!"

Planned subgroup analysis:
(Trial enrichment/targeted subgroup)

- Novel concepts and trial designs by Richard Simon et al. for oncology trials with a biomarker positive targeted subgroup; visit the website: brb.nci.nih.gov
  - Build a "predictive model" for identifying potential responders to the study treatment.
  - Validate the model (for adequate sensitivity and specificity)
  - Use this model to enrich the trial by "potential responders" forming a targeted subgroup in the trial

Drugs combination trials

- Usually a factorial trial with treatments A+B, A, B and placebo; A and B are approved components
- Important comparisons: A+B vs A, A+B vs B, and A+B vs. placebo
  - No multiplicity adjustment. Each test is performed at the 0.025 level by 1-sided test
- Other comparisons: A vs. placebo, and B vs. placebo
- Statistical test strategy much more involved:
  - when multiple drug combination (e.g., drugs A, B, C combined), multiple doses and multiple endpoints

3-arm trials with a new treatment, an active control, and placebo

- Comparisons of interest:
  - Is new treatment superior to placebo?
  - Is the active control superior to placebo?
  - Is the new treatment at least non-inferior to active control with a specified margin of non-inferiority?
- Approaches:
  - Koch and Rohmel (2004); Hauschke and Pigeot (2005); Rohmel and Pigeot (2009)

Three-tier approach for safety data analysis

- **Tier 1**: analyses of AEs associated with specific hypotheses formally tested in a clinical study on addressing both type I and type II error rates
- **Tier 2:** analyses of common AEs
- **Tier 3:** analyses of rare serious AEs (occurring in the range of 1/100 to 1/1000) may require large data base and evaluation by specialty area experts.
- (Mahrotra & Hayes, 2004): proposed the use of false discovery and double false discovery rate methods for Tier 2 analysis.

A suggestion for addressing Type II error rate concerns in analyzing safety events

1. Specify an error rate for failing to identify at least k (e.g., k =1 or 2) unwanted events out of the total K such events that will be analyzed.
2. For a specified alpha workout the size of the database needed for satisfying this error rate
3. If the size is limited then adjust the alpha above accordingly.

Concluding Remarks

1. 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.
2. 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
3. Clinical trials also pose multiple testing situations when multiplicity adjustment in not necessary

Concluding Remarks

4. 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

4. Multiplicity adjustment approaches (cont'd)
   - ✓ Graphical methods useful as explained
   - ✓ Truncated Holm's method – for more power for the 1st primary family
5. Co-primary endpoints issues:
   - ✓ Control of alpha necessary at 0.05 level. Some flexibility on the case-by-case basis when number of co-primary endpoints greater than 2 with some additional sources of evidence
6. Subgroup analysis
   - ✓ Unplanned subgroup analysis has serious limitations
   - ✓ Planned subgroup analysis: Novel approaches for enrichment of a trial by targeted subgroup, and analysis of the subgroup and the overall population
7. Analysis approaches for efficacy and safety are different

Ref: Useful references on multiplicity