Recent Development in Addressing Multiplicity Issues in Clinical Trials

Mohamed Alosh, Ph.D.
Division of Biometrics III,
OB/OTS/CDER FDA

Disclaimer: Views expressed in this presentation do not represent those of the FDA

Outline

I. Background: multiple testing in clinical trials
II. Some commonly used multiple testing procedures
III. Recent advances in multiple testing strategies:
  a) Fallback and extensions
  b) Adaptive Alpha Allocation Approach (4A)
  c) α-spending function
  d) Consistency Adjusted Alpha Adaptive Strategy (CAAAS)
  e) Gatekeeping multiple testing strategies
  f) Separable MTP
  g) Graphical approaches and recycling α
IV. Applications
V. Some concluding remarks

I. Background: multiple testing in clinical trials

- Most clinical trials collect efficacy and safety data on multiple endpoints, or time points, treatment groups, subgroups or a combination of these "endpoints".

- The objective of assessing the multiple endpoints might be to support an efficacy claim for an investigational therapy along with supportive evidence for labeling.

- With the current trend of accelerating drug development it is likely that:
  a) this evaluation of various endpoints will continue in the future, and
  b) less information will be available before the initiation of Phase 3 trials.

- For each endpoint tested there is a chance of making a false conclusion (claiming a benefit with the new treatment when it is not real) and such chance increases with the number of comparisons made.

I. Background: multiple testing in clinical trials

- Many Multiple Testing Procedures (MTPs) have been proposed in the literature to control the false positive rate (FPR) and a comprehensive review will be difficult (see e.g. Dmitrienko et al., 2009 or Hochberg and Tamhane,1987, for a general discussion).

- In parallel with accelerated drug development, new MTPs are introduced with flexibility, aimed to overcome the disadvantages of some commonly used methods.

- Several of these new methods have not been tested yet in applications. However, in general, there is no single method which uniformly out performs other methods, and differences between different MTPs can be subtle.

- There is a need for careful consideration to understand the best method for a given situation, if any, and whether its advantages are worthwhile for a given setting.

II. Some commonly used multiple testing procedures

Notation:

Let : $E_1, E_2, \ldots, E_k$ *denotes the k "endpoints", and*

$H_1, H_2, \ldots, H_k$ be the corresponding null hypotheses,

$H^{(1)}, H^{(2)}, \ldots, H^{(k)}$ *ordered hypotheses, set a priori, based on perceived "natural" ordering*

$\alpha_1, \alpha_2, \ldots, \alpha_k$ *corresponding allocated Type I error rates, with* $\alpha_i = w_i \alpha$ *, such that* $w_i \geq 0$ *and* $\sum w_i \leq 1$,

$T_1, T_2, \ldots, T_k$ *corresponding* test statistics,

$p_1, p_2, \ldots, p_k$ *corresponding p-values, and*

$p_{(1)}, p_{(2)}, \ldots p_{(k)}$ *ordered p-values, with* $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$

$H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ hypotheses corresponding to the ordered *p*-values.

I. Background: multiple testing in clinical trials:

Some general definitions:

Strong control of the familywise error rate (FWER): is the probability of rejecting at least one true hypothesis among several tested hypotheses, disregarding how many are true (Hochberg and Tamhane,1987).

Global hypothesis is the intersection of all given hypotheses H_i ($\cap H_i$). The global hypothesis will be rejected if anyone, or more, of the individual hypotheses are rejected.

Closed testing principle: one in which the k individual hypotheses are grouped together in 2^K-1 non-empty subsets of intersection hypotheses. Every intersection hypothesis is considered with a global test at the level α. An individual hypothesis is rejected if every intersection hypothesis containing that hypothesis is rejected. As long as the global test for every intersection hypothesis is carried out at the level α, the FWER for the family of individual hypotheses is controlled at the same level α. (Marcus et al, 1976)

II. Some commonly used multiple testing procedures

The Bonferroni Test:

- ✓ Test H_i at the level α/k, for i=1,2,….k.
- ✓ More generally, test H_i at $\alpha_i = w_i \alpha$, *with* $\sum w_i = 1$ *(weighted Bonferroni).*

Advantages:

- *Simple to apply and r*obust (always controls the type I error rate strongly).
- Allows for testing each endpoint.

Disadvantages:

- Conservative with large number of endpoints and/or correlated endpoints (loss in power).

➢ Useful when insufficient information about magnitude of treatment effect of different endpoints and results of each endpoint can stand on its own.

II. Some commonly used multiple testing procedures

Data Driven Procedure:

The hypotheses are tested in an order determined by the magnitude of their *p*-values, that is: $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$.

Step-up (step-down) procedures test the hypotheses associated with the largest (smallest) *p*-value first; “up” (“down”) refers to increasing (decreasing) test statistics which correspond to decreasing (increasing) *p*-values.

II. Some commonly used multiple testing procedures

Holm test (Holm,1987):

- ✓ A commonly used step-down approach which tests $H_{(i)}$ by comparing $p_{(i)}$ *with* $\alpha/(k-i+1)$ *as long as previous hypotheses are rejected.*
- ✓ *Holm also proposed a weighted test.*

Advantages:

- Simple to apply and controls the Type I error rate strongly.
- An extension of the Bonferroni test, as it 'updates' the significance level after rejecting the previous hypothesis taking into account the remaining hypotheses to be tested → Improving the power of the Bonferroni.

Disadvantages:

- Still conservative

- ✓ Can be useful in the absence of good information about magnitude of treatment effect of various endpoints

II. Some commonly used multiple testing procedures

Fixed sequence approach:

✓ Hypotheses are tested in an order set *a priori*, $H^{(1)}$, $H^{(2)}$,…., $H^{(k)}$, *where* $H^{(i)}$ *is tested at the full α as long as all previous hypotheses are rejected.*

Advantages:

- Simple and controls the Type I error rate strongly.
- Optimal when early tests in the sequence have largest treatment effect and performs poorly when early hypotheses are true (Westfall and Krishen, 2001).

Disadvantages:

- Once a hypothesis is not rejected no further testing is permitted.

➢ An approach to alleviate this problem is the "Fallback" method.

III. Recent advances in multiple testing strategies:

a) The Fallback method (Wiens, 2003 and Wiens and Dmitrienko, 2005).

- ✓ α is partitioned among the endpoints (with $\alpha_i = w_i\,\alpha$, allocated for testing H_i (i=1,2,…k) such that $\sum w_i = 1$; (as in the Bonferroni);
- ✓ The hypotheses are ordered prospectively with testing carried out sequentially (like the fixed sequence).
- ✓ $H^{(i)}$ *is tested at the level* α^i,

  with: $\alpha^i = \alpha_i$ if $H^{(i-1)}$ *is not rejected*

  $= \alpha_i + \alpha_{i-1}$ if $H^{(i-1)}$ is rejected.

Advantages

- α^i "adapts" to previous findings (recycles "unused" α) leading to higher power than the Bonferroni.
- It allows testing for all hypotheses, unlike the fixed sequence, which is a special case of the fallback by setting: $\alpha_i = 0$ for i >1.

➢ Useful when there is insufficient information prior to the conduct of Phase 3 trials.

III. Recent advances in multiple testing strategies

- Additional properties for the fallback test are gained by casting it as a closed testing procedure (Wiens and Demitriko, 2007), which include:
- Control of the Type I error rate for more than 2 hypotheses
- With identical weights used for the fallback and weighted Holms the two procedures are quite similar
- *The test is not alpha exhaustive and can be made so to gain additional power, but this may change some of its desirable properties.*

An alpha exhaustive test is a closed test in which each intersection hypothesis is tested at the full alpha level. If an intersection hypothesis is tested at level less than α, the test is not alpha exhaustive and can be made uniformly more powerful by testing that intersection at the full α level. (Grechanovsky and Hochberg, 1999).

III. Recent advances in multiple testing strategies

- ✓ An extension of the fallback method, often called in the literature as "parametric fallback" method, to account for correlation which leads to some power improvement is given in Huque and Alosh (2008).

- ✓ A more general approach for adapting the significance level for testing the second based on the findings of first hypothesis is the Adaptive Alpha Allocation Approach (4A) procedure (Li and Mehrotra, 2008).

III. Recent advances in multiple testing strategies

b) Adaptive Alpha Allocation Approach (4A): (*Li and Mehrotra, 2008*)

Introduced to handle potentially underpowered endpoints.

- ✓ Like the fallback, *allocate* $\alpha_i = w_i\alpha$ *for testing* $H^{(i)}$, $i=1,2$ *and* $w_1 + w_2 = 1$
- ✓ *Test* $H^{(1)}$ *at* α_1*; if rejected test* $H^{(2)}$ *at the level* α *(i.e., accumulate "unspent" alpha) otherwise test* $H^{(2)}$ *at the level min* $(\alpha_t / p_1^2, w_1\alpha$, *where:*

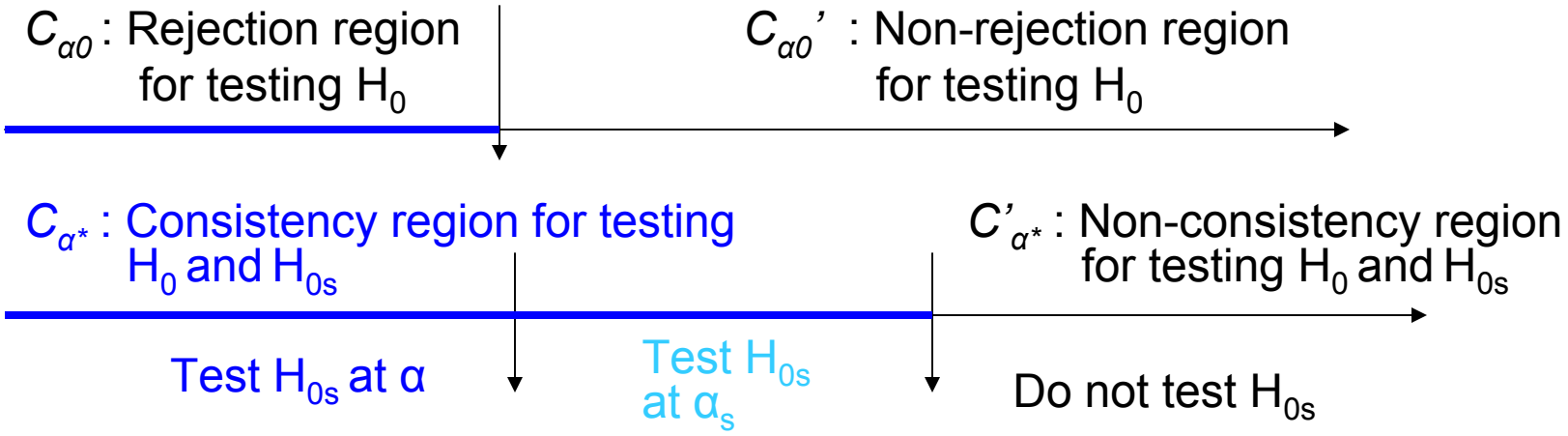
$$\alpha_t = w_1\alpha\,(\alpha - w_1\alpha)/(1 - w_1\alpha) \quad \text{if } w_1\alpha + (w_1\alpha)^2 - (w_1\alpha)^3 > \alpha$$
$$= w_1\alpha\,\{1 - [\,2\,(w_1\alpha - \alpha - (w_1\alpha)^2) / w_1\alpha\,]^{1/2}\,\} \quad \text{otherwise}$$

- Li and Mehrotra proved that the 4A controls the FWER for two independent endpoints strongly. They also adapted the approach to account for correlation.
- For more than two endpoints that are not independent they recommended further research.

III. Recent advances in multiple testing strategies

c) Consistency concept:

- For two endpoints to be used interchangeably for an efficacy claim a certain degree of consistency, at a minimum, in their finding is expected.
- This concept came originally when testing for total population (H_0) and subgroup (H_{0s}) [Song and Chi, 2007 and Alosh and Huque, 2007& 2009] where testing H_{0s} is recommended only when H_0 meets a pre-specified 'consistency' criterion, (say, α^*) such that $0< \alpha - \alpha_0 < \alpha_0 < \alpha^* << 1$.



III. Recent advances in multiple testing strategies

c) α-spending function (Zhao et al. 2010):

Introduced in the context of subgroup analysis to determine the α–level for the targeted subgroup (α_s) using an α-spending function $f(x)$, with:

(i) $0 \le f(x) \le 1$ if $\alpha_0 \le x \le 1$; and

(ii) $P_r (p_0 > \alpha_0 ,$ and $p_s \le \alpha f(p_0)) = \alpha - \alpha_0$ under the global hypothesis.

The decision rules for testing H_0 *(total population) and* H_{0s} (subgroup):

Reject H_0 if : (1) $p_0 \le \alpha_0 ,$ or (2) $\alpha_0 < p_0 \le \alpha$ *and* $p_s \le \alpha f(p_0)$

Reject H_{0s} if: $p_0 \le \alpha_0$, and $p_s \le \alpha$ or (2) $p_0 > \alpha_0$ and $p_s \le \alpha f(p_0)$.

Zhao et al. also considered restricting $f(x)$ to establish consistency requirements between the subgroup and the overall study findings.

Can be viewed as a generalization of the 4A and the flexible approach in Alosh and Huque (2009).

III. Recent advances in multiple testing strategies

d) Consistency Adjusted Strategy (CAS): (Huque and Alosh, 2010)

- ✓ Closely related to the "α-spending function" of Zhao et al. approach by:
- Considering the consistency requirement, and
- Allowing any non-increasing function for adaptation of the significance level for testing $H^{(2)}$

- ✓ However, unlike α-spending function, CAS allows re-testing $H^{(1)}$ if it is originally missed slightly and the result for $H^{(2)}$ turns out to be relatively strong.

III. Recent advances in multiple testing strategies

d) Consistency Adjusted Alpha Adaptive Strategy (CAAAS), Alosh and Huque, 2010; is a generalization of CAS to allow for:

- Two levels of consistency criteria (α^*, α^{**}) with $0 < \alpha 1 \leq \alpha^* \leq \alpha \leq \alpha^{**} \leq 1$;
- Different functions for adaptation of α_2 *as a function of* p_1 in each interval $(0, \alpha 1]$, $(\alpha 1, \alpha^*]$ and $(\alpha^*, \alpha^{**}]$;
- The criteria for the adaptation functions ($f_i(.)$ for i =1,2,3) is that the weaker the efficacy of E_1 the stronger the evidence required on E_2 for convincing evidence (e.g. linear, exponential, etc..).

✓ It unifies most previous approaches as by appropriate selection of the consistency criteria (α^*, α^{**}) and the adaptation functions ($f_i(.)$ for i =1,2,3) these approaches arise as special cases of CAAAS.

III. Recent advances in multiple testing strategies

An outline for CAAAS: Allocate α_1 (with $\alpha_1 < \alpha$) for testing $H^{(1)}$: and initially allocate α_{21}, α_{22} and α_{23} for testing $H^{(2)}$ ($\alpha_{23} \le \alpha_{22} \le \alpha_{12}$) *and for* consistency criteria $\alpha^* \le \alpha \le \alpha^{**} \le 0.5$ for 1-sided test; then consider:

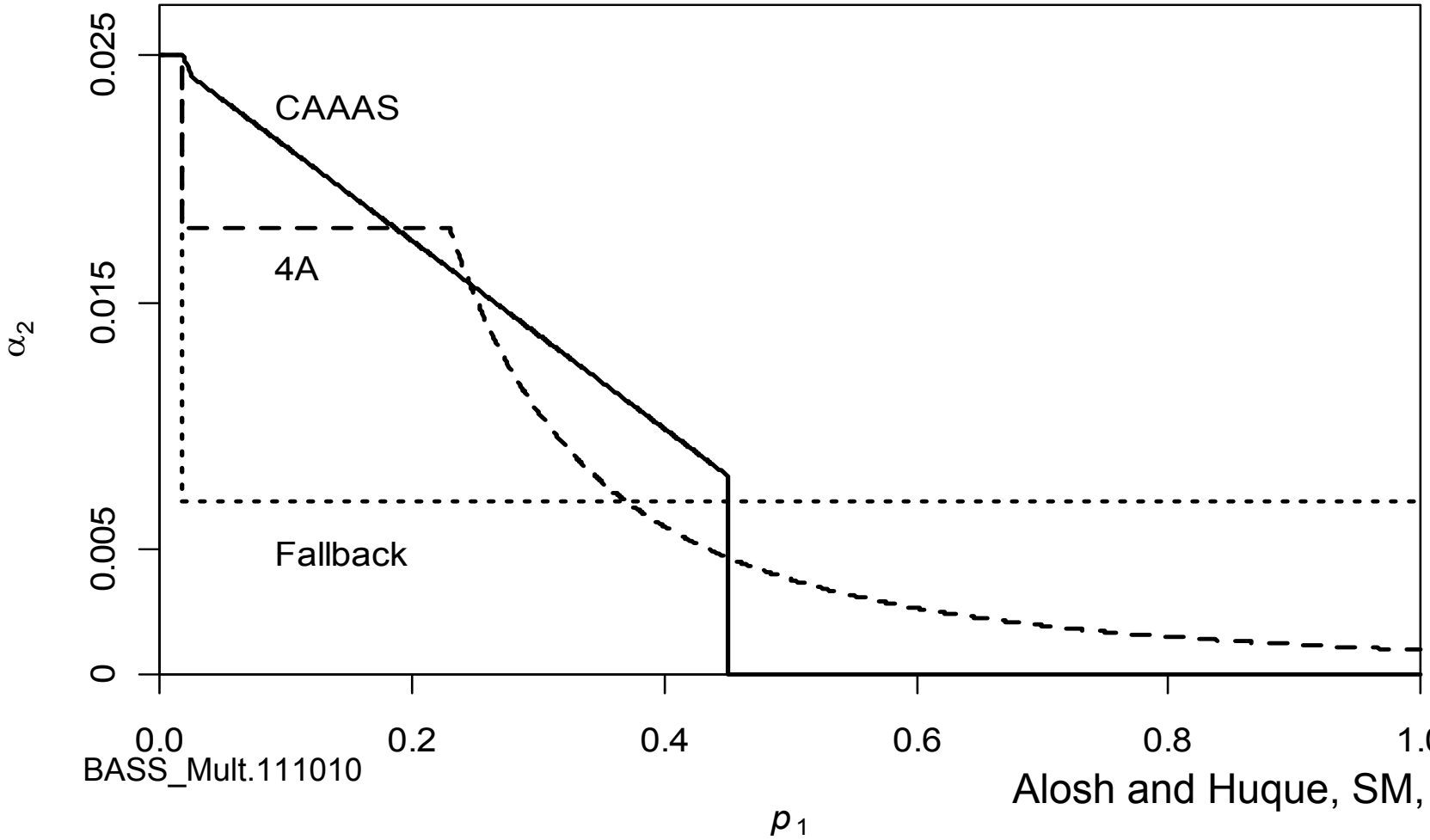
Step 1: Test H_{01} at the significance level α_1
If H_{01} is rejected, then test H_{02} at the significance level $\alpha_{21} = f_1(p_1) = \alpha$;
If H_{01} is not rejected at the level α_1, then go to Step 2.

Step 2: If p_1 satisfies the condition: $\alpha_1 \le p_1 < \alpha^*$, then:
Test H_{02} at the significance level $\alpha_{22} = f_2(p_1)$;
If H_{02} is rejected in Step 2 reject H_{01} also.
If H_{02} is not rejected go to Step 3.

Step 3: If p_1 satisfies the condition: $\alpha^* \le p_1 < \alpha^{**}$, then:
Test H_{02} at $\alpha_{23} = f_3(p_1)$ for rejecting **only** H_{02}.
Otherwise, no further testing is permitted if $p_1 \ge \alpha^{**}$.

III. Recent advances in multiple testing strategies

Figure 1: Comparison of Significance Level for H_{02} (α_2) for Different Methods
($\alpha=\alpha^*=0.025$, $\alpha_1=0.018$, and $\alpha^{**} = 0.45$, all 1-sided)

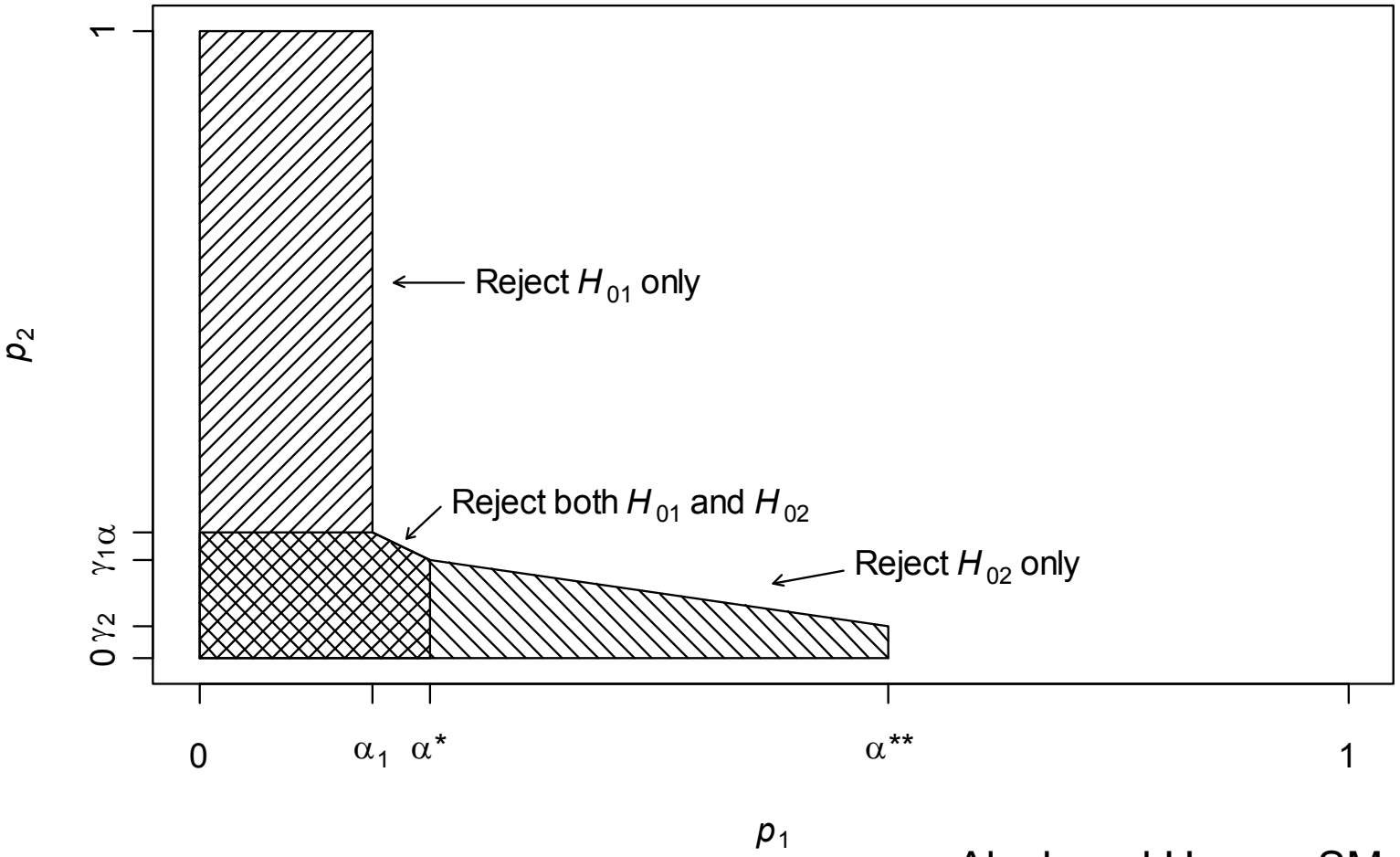


Alosh and Huque, SM, 2010

III. Recent advances in multiple testing strategies

Figure 4: CAAAS Piecewise Linear Rejection Region

($\gamma_1 \leq \alpha$, $\alpha^*=\alpha$)



Alosh and Huque, SM, 2010

III. Recent advances in multiple testing strategies

Comments about α*:

- The idea is that for E_1 and E_2 to be used interchangeably for establishing an efficacy claim when a certain degree of efficacy is expected, otherwise it would be difficult to interpret study findings.
- A study which establishes efficacy under a strict consistency criterion provides more persuasive evidence than one with a relaxed criterion.
- How to select α*: Closeness of the two endpoints (measuring the same symptoms !), correlation and clinical determination.

III. Recent advances in multiple testing strategies

e) Gatekeeping strategies:

A flexible approach for grouping multiple endpoints in blocks depending on their "relative importance".

- ✓ Hypotheses with similar objectives are grouped in families, and
- ✓ Families are tested sequentially, where an early family in the sequence serves as a gatekeeper for testing the next family (a generalization of the fixed-sequence approach).

Example: H_1, H_2 ,….., H_k can be for primary, secondary and tertiary endpoints.

F_1 = { *primary objectives);*

F_2 = *{secondary objectives},*

……..

F_m= *{other endpoints .....* }

III. Recent advances in multiple testing strategies

e) Gatekeeping strategies:

✓ When testing hypotheses within a family the criterion for passing the gate can be:

- Each hypothesis in the family is to be rejected (serial gatekeeing), Bauer et al, 1998 and Westfall and Krishen 2001). Note in this case each hypothesis would be tested at the full α (e.g. case of co-primary endpoints).

- At least one hypothesis in the family is rejected (parallel gatekeeing), Dmitrienko et al, 2003 based on *the closure principle; Guilbaud,(2007) showed that a stepwise parallel gate keeping procedure can be directly constructed without appeal to the closure principle*.

- *Tree-structured gatekeeping unifies and generalizes the* serial and parallel g*atekeeping procedures. (Dmitrienko et al, 2007) to address multidimensional hypotheses testing problem.*

III. Recent advances in multiple testing strategies

e) Gatekeeping strategies:

Comments:

- ✓ Selection of the type gatekeeing (serial, parallel or tree structured) would be driven by hierarchy of the endpoints and the success criteria for passing the gate.

  In oncology trials:

  F_1:(overall survival) → F_2 (progression free survival)
  → F_3 (time to treatment failure)…

  In psoriasis trials, rejection of at least one hypothesis in *F_1 = (success on the investigator global evaluation or PASI 75)* might be sufficient *to win → F_2 = (secondary objectives: puritis, erythemia, scaling).*

III. Recent advances in multiple testing strategies

e) Gatekeeping strategies:

Comments (cont):

- ✓ *Generally, FWER is controlled at the level α for $F = U F_i$ of all hypotheses and different MTPs can be used for different families as long as the FWER is controlled.*

- ✓ Might assign weights when testing within a family and across families . Within family the relative weights might reflect the relative importance of the individual hypotheses. Similarly weights across families might assign greater importance to a certain family and consequently increase the likelihood of rejecting the hypothesis in the family.

III. Recent advances in multiple testing strategies

f) Separable MTP: (Dmitrieko et al 2008).

- ✓ Introduced in the context of developing a general multistage procedure for arbitrary gatekeeping problem, by carrying forward the Type I error rate for any rejected hypotheses to test hypotheses in the next ordered family.

- ✓ For any subset of hypotheses $I \sqsubseteq K \{1,2,\ldots k\}$, define an error rate function *e(I)* to be the maximum probability of making at least one Type I error in the subfamily $\{H_i, i \in I)$ of an MTP, and let *e*(I)* be the upper bound for *e(I).*

  For example for the Bonferroni, which rejects H_i if $p_i < \alpha/k$, the upper bound is $e^*(I) = \alpha |I| /k$; where $|I|$ is the cardinality of set I.

- ✓ Then for a MTP at the level α and for an index set (A) of its accepted hypotheses; the part of α that is unused and can be carried over to test the hypotheses in the next family is $\alpha - e^*(A)$.

III. Recent advances in multiple testing strategies

f) Separable MTP: (cont.).

- ✓ The Bonferroni MTP is separable, but others such as Holm, Hochberg, and the Fallback are not.

- ✓ Dmitrieko et al. proposed a modification of these tests by taking a convex combination of their critical constants with the Bonferroni critical constant, so that the truncated versions of these tests are separable and more powerful than the Bonferroni.

- ✓ In particular for the modified (or truncated) Holm test the critical constant for comparing $p_{(i)}$ is:

  $$w_i \alpha = [\gamma / (k-i+1) + (1-\gamma)/k]$$

  for a specified γ $(0 \leq \gamma < 1)$ called the truncation fraction.

- ✓ The power of truncated Holm is increasing γ. For γ =0 and γ=1 the modified Holm reduces to the Bonferroni and Holm, respectively.

III. Recent advances in multiple testing strategies

g) Graphical approaches and recycling α: (Bretz, et al. 2009 and Burman et al 2009).

Main idea:

- ✓ An iterative approach to construct and perform Bonferroni-based MTPs, with α being split between the different null hypotheses with α_i allocated for H_i (i=1,2,…k).
- ✓ Whenever a null hypothesis is rejected its nominal level alpha (α_i) may be recycled to the testing of other hypotheses.
- ✓ The recycling MTPs are closed testing procedures, thus it controls the type I error rate strongly.
- ✓ The class of such MTPs includes, for example, serial and parallel gatekeeping, fallback and Holm procedures.
- ✓ Graphical displays for such MTPs make it easier to communicate an underlying MTP which relies on closed test principle, and may facilitate the tailoring of MTPs for different purposes.

III. Recent advances in multiple testing strategies

g) Graphical approaches and recycling α:

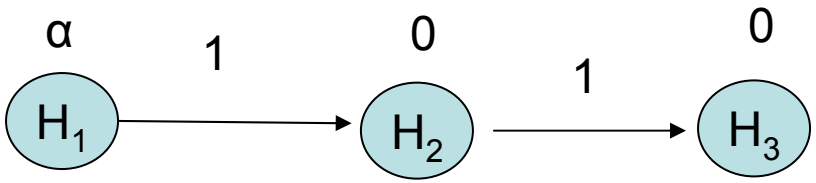
Some details:

- ✓ The MTPs is represented by directed, weighted graphics, where nodes corresponds to the elementary hypotheses
- ✓ The elementary hypotheses are represented by a set of vertices with associated weights representing local significance level.
- ✓ The weights associated with a directed edge between two vertices indicates the fraction of the (local) significance level at the initial vertex (tail) that is added to the significance level at the terminal vertex (head) if the hypothesis at the tail is rejected.
- ✓ An algorithm to generate such graphs while sequentially testing the individual hypotheses.

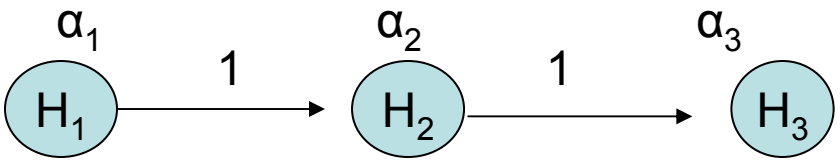
III. Recent advances in multiple testing strategies

Simple examples of graphical approaches for MTPs

Fixed sequence:

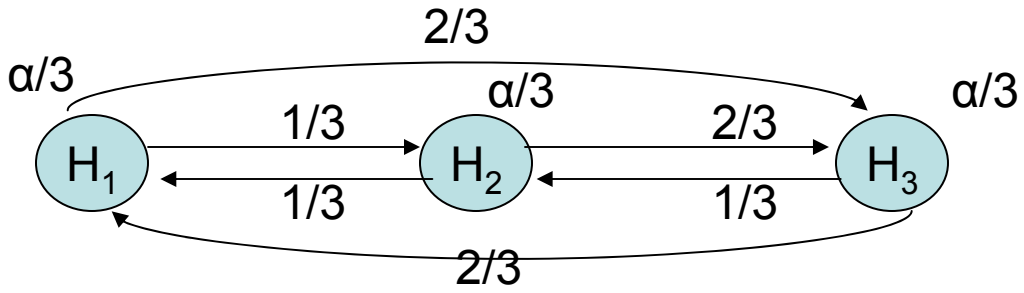


Fallback:

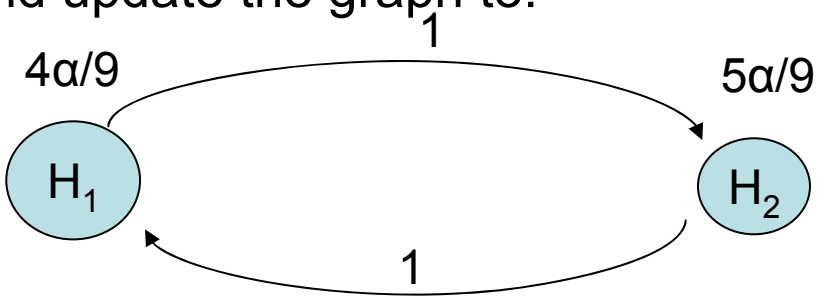


III. Recent advances in multiple testing strategies

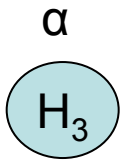
Graphical approach for sequentially rejective Bonferroni and recycling α:



Assume $\boldsymbol{p}$ = (0.02, 0.01, 0.06) and for α=0.05, then in the first step H_2 will be rejected and update the graph to:



In the second step H_2 is rejected and update the graphic to:



As p_3= 0.06 > 0.05, stop

IV. Application: Retrospective analyses on two clinical trials

Setting: Use 1-sided test with $\alpha = 0.025$, $\alpha_1 = 0.02$ and $\alpha^* = 0.10$ and $\rho = 0.50$. $\rightarrow \gamma_2 = 0.0161$ (table from computation).

IV.1 The ATLAS trial:

- The "Assessment of Treatment with Lisinopril And Survival" (ATLAS) trial (Packer, et. al.1999) comparing efficacy and safety of low and high doses of ACE inhibitor Lisinopril in patients with congestive heart failure.

- The primary endpoint: all-cause mortality; and
  'principal' secondary endpoint: all–cause hospitalization.

- Results: $p_1 = 0.064$ and $p_2 = 0.001$; (1-sided).
- Conclusion: Trial is not positive on E_1 ; but since $p_1 = 0.064 < \boldsymbol{\alpha}^*$
  $\rightarrow$ test E_2 at $\gamma_2 = 0.0161$; which led to a positive trial by CAS.
  Also the same conclusion with the 4A method, weighted Bonferroni and the Fallback lead to the same conclusion.

IV. Application to a published clinical trial data:

The PROactive Trial

- Study designed to assess the efficacy of pioglitazone in reducing cardiovascular events and mortality in type 2 diabetes patients.
- The primary endpoint: time from randomization to first occurrence of: all-cause mortality, non fatal MI, stroke, acute coronary syndrome (ACS), major leg amputation (above the ankle), leg revascularisation, cardiac intervention.
- The main secondary endpoint: time from randomization to first occurrence of: all-cause mortality, non fatal MI (excluding silent MI), stroke.

- Results: p_1= 0.0475 and p_2 = 0.0135.

- Conclusion: Trial is not positive on E_1 ; but since p_1 = 0.0475 < $\boldsymbol{\alpha}$ *
  $\rightarrow$ test E_2 at γ_2 =0.0161; which led to a positive trial by CAS. Also the same conclusion with the 4A method.
  However, both the weighted Bonferroni and the Fallback lead to a different conclusion.

V. Some concluding remarks:

- ✓ Several MTPs have been proposed recently in the literature to accompany the accelerated drug development.
- ✓ Flexibility of the proposed approaches allows:

(i) Recycling the significant level of a rejected hypothesis to other hypotheses*; leading a* higher significance levels for testing the later hypotheses in the sequence. This process can be re-iterated for Bonferroni-based MTPs until no further rejection of hypotheses can be made.

(ii) Adapting the significance level for testing a hypothesis in a sequence to the findings *of testing of the preceding hypothesis in the sequence* even if the later was not significance but still meet a pre-specified consistency requirements.

(iii) Modification of commonly used MTPs by truncating them to make them separable while maintaining their power advantage over the Bonferroni MTP.

V. Some concluding remarks:

- ✓ In general, no single method is expected to uniformly out perfom other methods and the differences between different methods can be subtle.
- ✓ Need better understanding of the properties of each method when applied to a particular situation.
- ✓ For several of these methods, selection of the weights, truncation factors, splitting α_i's and selection of consistency criteria, should be considered whenever applicable. This might be driven by the objective of the MTPs, whether it is increasing the power of the trial (conditional vs marginal power), increasing the chance of a positive trial vs increasing the number of hypothesis rejected, etc…

Some references

Alosh, M. and Huque, M.F. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 2009, **28**: 3-23.

Alosh, M. and Huque, M.F. A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine* 2010, **29**:1559-71.

Bretz, F, Maurer, W, Brannah, W and Bosch, M (2009). A graphical approach to sequentially rejective multiple test procedures Statistics in Medicine586-604.

Burman,CF, Sonesson C, Guilbaud, O (2009). A recycling framework for the construction of Bonferroni-based multiple tests. Statistics in Medicine, 28(5) 739-61

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

Dmitrienko, A. Bretz, F.Wesfall, P.H. Troendle,J. Wiens, B, Tamhane,AC,.Hsu,J.C (2009) Multiple Testing methodology. *Multiple testing problems in pharmaceutical Statistics*. Dmitrienko, Tamhane, A.C, Bretz, F (editors). Chapman and Hall/CRC Press. New.

Dmitrienko, A. Tamhane, AC Wiens, B .(2008) General multistage gatekeeping procedures. Biometrical Journal, 50 667-677.

Dormandy, J. et al. on behalf of the PROactive investigators, Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive study (PROspective pioglitazone clinical trials in macrovascular events): a randomized controlled trial. *Lancet* 2005, **366**: 1279-89.

Guilbaud, O. (2007) Bonferroni Parallel Gatekeeping-Transparent Generalization , Adjusted P-values and Short direct proofs. Biometrical Journal, 49, 917-927.

Grechanovsky, E. and Hochberg, Y. (1999). Clsoed procedures are better and often admit a shortcut. *J. Statistical Planning and Inference; 76*: 79-91.

References (cont)

Hohberg, Y. Tamhane, AC (1987) Multiple comparison procedures. Wiley, New York

Holm, S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 1979; 6, 65-70

Huque, M.F. And Alosh, M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *J. Statistical Planning and Inference* 2008*;* 138: 321-35.

Huque, M.F. and Alosh, M. A consistency-adjusted strategy for accommodating an underpowered primary endpoint, 2010, *J. Biopharmaceutical Statistics* (to appear)

Li J, Mehrotra D. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine* 2008; **27**:5377--5391

Marcus, R. Peritz, E and Gabriel, KR (1976) On closed testing procedures with special reference to ordered analysis of variance, Biometrika 63, 655-660.

Song, Y. and Chi, G. Y. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 2007; **26**: 3535-49.

Westfall,PH, Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. J. Statistical Planning and Inference, **99**, 25-40

Wiens, B. L. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003;2: 211-15.

Wiens, B. L. and Dmitrienko, A. The fallback procedure for evaluating a single family of hypotheses. *J. Biopharmaceutical Statistics* 2005; 15: 929-42.

Wilcox, R., Bousser, M-G, Betteridge, D.J, Guntram, S., Pirags, V., Kupfer, S., and Dromandy, J. Effects of Pioglitazone in patients with or without previous stroke. *Stroke* 2007, **38**: 865-73.

Zhao, YD, Dmitrienko, A and Tamura, R. Design and analysis considerations in clinical trials with a sensitive subpopulation, Statistics in Biopharmaceutical Research, 2010, **2**, 72- 83.