Decomposition of a Composite Endpoint for Assessing Treatment Benefits for its Components

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

BASS Conference 2010, Hilton Head, SC,
November 10, 2010

Use of composite endpoint as a primary endpoints in clinical trials - widespread

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

Composite (primary) endpoint

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

Composite endpoint topics[#] wide; and trials with CE often face difficult issues

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

[#](topics listed by Joachim Röhmel, 2004)

This presentation

- Background on composite endpoints
- Why assess treatment effects on the components of the composite?
- Issue of heterogeneity
- Issues when mortality is a component
- Use of a composite endpoint for an enrichment design
- Statistical methods and approaches
  - multi-branched test strategies for the composite endpoint and its components
  - graphical method of Bretz et al. (2009)
  - adaptive methods
  - consistency ensured methods
- Concluding Remarks

Three types of composite endpoints: (Chi, 2005)

1. An index or a responder type endpoint that is a weighted or an un-weighted combination of multiple item scores, counts or other endpoints.
   - E.g., HAM-D total score for depression trials and the ACR20 (or ACR70) endpoint for the rheumatoid arthritis trials
2. Compound failure rate endpoint that captures clinically relevant treatment failures of different types into single failure rate endpoint.
   - E.g., an organ transplant trial, in treating a patient during the first six-months after transplantation, may call the treatment for that patient as a failure if there is a biopsy-proven acute rejection or graft loss or death.

Three types of composite endpoints (Chi, 2005)

3. The third type, very common for cardiovascular trials, combines several binary (or time-to-event) type endpoints into a single composite endpoint usually counting in a patient the component event that occurs first.
   - E.g., a CHF (congestive heart failure) trial may define a composite endpoint of death and hospitalizations as first occurrence of any of these two events, i.e., if a patient was hospitalized and then died after a few days, then this composite endpoint would count the event as hospitalization. However, a separate analysis of components would capture both events.

Several motivations (Moye, 2003)

- Reduces multiple endpoints to a single PE
- Can reduce the size of the trial if
  - Components increase the number of events in non-overlapping manner (i.e., an event is not a direct consequence of the other)
  - Some homogeneity of treatment effects across components, or components jointly enhance the overall treatment effect
- Can addresses a broader aspects of a multifaceted disease
- Can change the focus of the trial
  - To discovering clinically meaningful small treatment effects that collectively demonstrate a statistically significant and clinically meaningful benefit of the treatment

An example of sample size reduction
(consider a 2-arm CHF trial)

- Trial size with the CHF mortality endpoint:
  - Assume risk reduction of 12% in the 18 month incidence rate of CHF mortality (treatment vs. to control)
  - Assume the 18 month CHF mortality rate is 18% for the control group, and the power is 90%, 2-sided test, type I error of 0.05.
  - Trial sample size = 12,653 patients
- Trial size with the composite endpoint of CHF related death and CHF related hospitalizations
  - Assume that control group event rate increases to 36% (because of additional hospitalization events)
  - The trial size reduces 5,032 patients for detecting the same low level of efficacy of 12% risk reduction.

#Moyé (2003): similar example in his book, *Multiple Analyses in Clinical Trials* (p. 224)

Some key considerations

- Clinical relevance of the composite endpoint
  - there should be a prior empirical evidence of this given the type of disease and drug studied.
- Prospectively defined
  - the endpoint itself and all its components along with their ascertainment methods
- Proper choice of components
  - input from disease area experts, special studies and evaluation of historical data pertinent to the type of drug and disease under study.
  - decisions about the number and type of components and their relative importance usually have implications in the interpretation of the outcomes of the composite endpoint.

Some key considerations (cont'd)

- Quality and sensitivity of a composite endpoint depends on the quality and sensitivity of the component endpoints
- A trial may not use a single composite endpoint as a primary endpoint
  - when large variations are expected to exist among the components with respect to importance to patients, frequency of outcomes, and the size of the effects.
- A trial with a combined primary outcome of two (or more) separate unrelated sets of components (clinically and physiologically unrelated), although sufficiently powered, may fail to show a joint treatment effect
  - when one set is sensitive to the treatment effect and the other set is not and can move in the wrong direction (NEJM 2010; 362:1959-1969).
- In such cases, however, breaking the single composite to suitable sub-composites that amends these anomalies at the sub-composite level can be helpful.

Principle of full disclosure

- Outcomes of the component endpoints must be fully displayed along with the composite endpoint outcomes for allowing a meaningful interpretation of the results of the composite and its components
- These displays can be done in multiple ways for proper understating of patterns of outcomes and how they are distributed in the treated and the control groups.

Displays of the outcomes of the components

- A simple first step is to display the first occurrences of the composite by components. This shows how the component events make up the first composite events.
- This type of component-wise displays is not appropriate for evaluating comparative efficacy at the component level, as it may distort the result of a serious component if the probability of a serious component depends on the probability of a less serious component.
- It is necessary to display the component-wise data, or conduct the analysis thereof, on capturing all events of a component including those that are not first events.
- It is also useful to display the comparative event data of a trial in a manner that also shows occurrences of all component events on a patient and the order in which those events occurred.

Adequate follow-up of patients

- An improper practice:
  - Some investigators may remove patients from a trial on observing a less serious component on them and not allowing them to stay in the trial long enough for observing other more serious components on them, even if they were rescued with other medications.
  - This may cause bias in the results of component-wise analysis, because, after the occurrence of a less serious event on a patient, the probability of occurrence of a serious event on that patient may me high during an adequate follow-up period.

Why assess treatment effects on the components of the composite?

- Statistically significant treatment effect benefit on a composite endpoint does not necessarily mean benefits on all its components.
  - Need to know which components are impacted and drive the result
  - LIFE trial: composite = CV deaths + MI + stroke; Results: Composite, $p < 0.021$ (2-sided); CV death alone, $p = 0.206$; MI alone, $p = 0.491$; stoke alone, $p = 0.001$.

  (*The Lancet*, Volume 359, Issue 9311, Pages 995-1003)
- Statistically significant treatment effect benefit on a composite endpoint may not have clinical utility
  - If a serous component such a mortality suggests harm to patients by the treatment.

Claims of treatment benefits on components

- Claims of treatment benefits on one or more key components (e.g., for mortality) in the total population or in a target subgroup of patients

  (a) When the composite endpoint is impacted significantly in favor of the treatment

  (b) When the result for the composite endpoint is not statistically significant, but trends in the right direction

  (c) When the composite endpoint is not significant but the mortality endpoint shows a robust statistically significant and clinically meaningful result

When do multiplicity issues arise in composite endpoint trials?

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

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

Issue of heterogeneity – example

Is there a treatment benefit for the "hospitalization" endpoint?

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)

Issue of heterogeneity across components:

(mortality trending in the wrong direction)

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

The usual questions then are:

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

  (Following are some ideas – next 3 slides)

1. Assign clinical utility weights

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

Comments

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

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

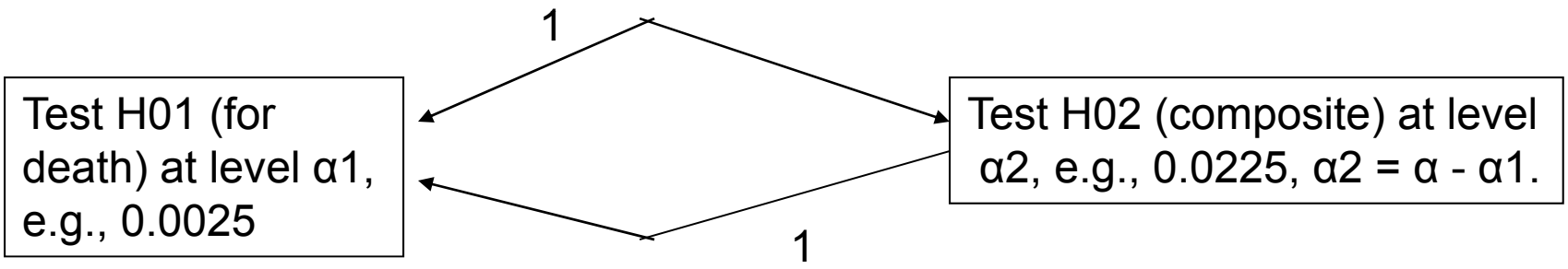
- Set a margin for acceptable inferiority for critical components, e.g., the upper CI for the mortality odds-ratio not to exceed 1.2

  Comments

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

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

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



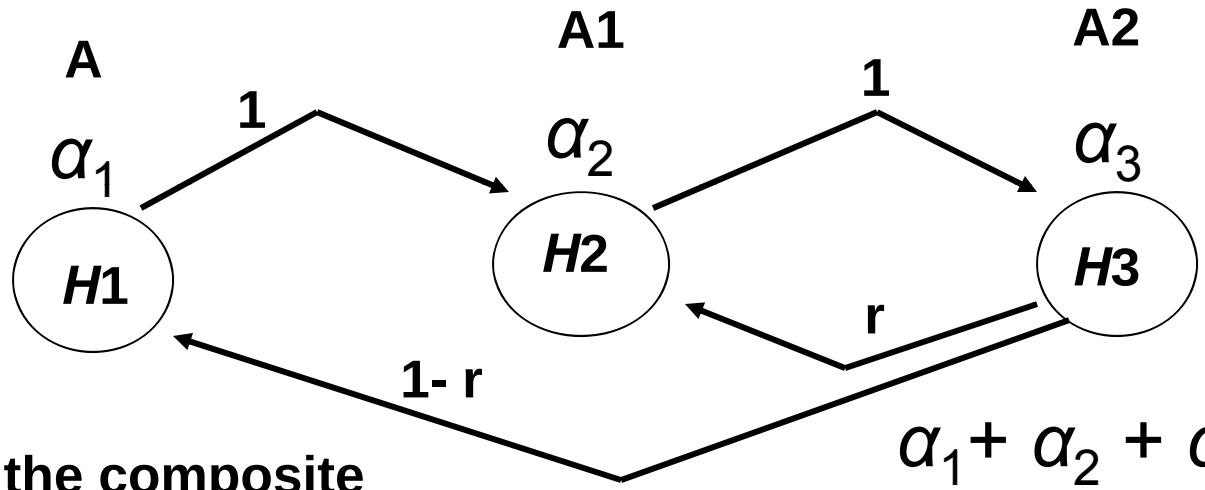
1. If H02 is rejected, then test H01 at the full significance level of 0.025 and accept the result for H02 if the 1-sided p for death < α* (e.g., α* = 0.50 or 0.55 to satisfy consistency of direction of effect)
2. If H02 is not rejected then test H01 at level α1
3. One can also start the test on the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

Statistical methods

(for testing a composite endpoint and its components)

- Example 1:
  - Statistical tests of a composite and its components by the improved fallback method
  - 2-arm trial designed to compare a treatment to control for the composite endpoint A and for two of its components A1 and A2
- Example 2:
  - Statistical tests of a composite and for two of its components for non-inferiority and superiority by a (multi-branched) tree-structured gatekeeping method
  - 2-arm CHF trial with a composite of mortality and hospitalization endpoints as components

Example 1
(test by the improved fallback method)



$\alpha_1 + \alpha_2 + \alpha_3 = \alpha$

***H*1 is for the composite**
***H*2 and *H*3 for its components**

If *H*1 is rejected, its *α*1 is passed to *H*2, then test *H*2 at *α*1 + *α*2
If *H*2 is rejected, then test *H*3 at *α*
If *H*1 is not rejected, then test *H*2 at *α*2, and so on
If *H*3 is rejected, then its alpha is distributed to *H*1 and *H*2 for retest

(See, Bretz et al., 2009)

Example 2: A CHF trial

- **Composite endpoint *c*** = death (*d*) + hospitalizations (*h*);

  T**est strategy:**

  1. **Test for *c***: Compare treatment **T** to **Control** for finding that **T** is at least non-inferior. If **T** is non-inferior to **Control** then find if **T** is also superior to **Control.**
  2. **Test for *d***: Once **T** is at least non-inferior to **Control** on ***c***, proceed to test for the most relevant component ***d*** (death endpoint) in the same manner.
  3. **Test for H**: if **T** is at least non-inferior to **Control** on endpoint ***d***, then do the same for the hospitalization component ***h***.

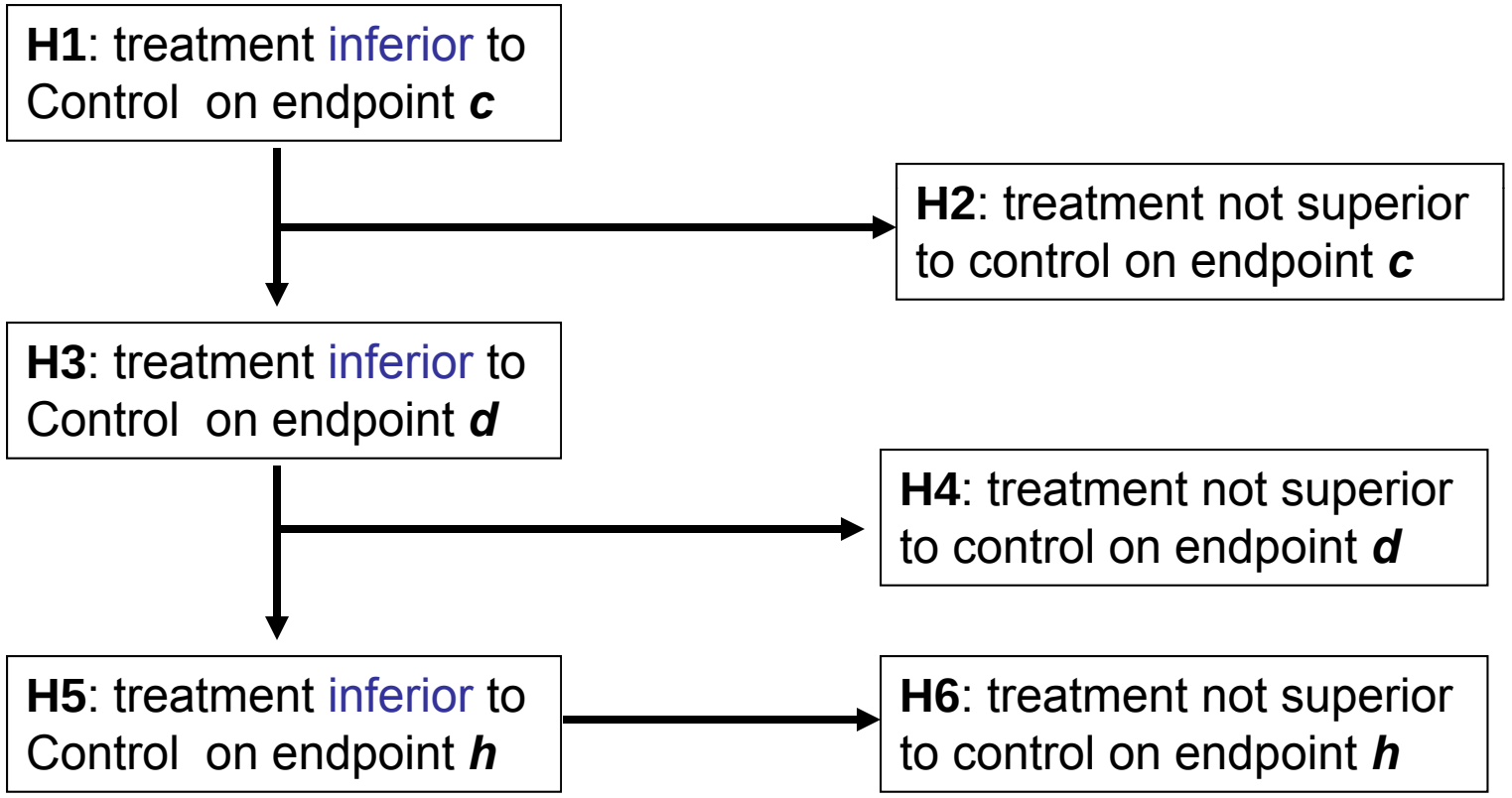
- There are also logical restrictions:
  - Test endpoint **d** after non-inferiority for endpoint ***c*** is first established; similarly, test endpoint ***h*** only after non-inferiority for endpoint ***d*** is first established
  - Test for superiority on an endpoint only after non-inferiority for that endpoint is first concluded..

(Multi-branched) tree-structured gatekeeping test strategy[#]?

- Multiplicity problems involving a composite endpoint and its components can be of multi-dimensional hierarchical structure.
  - One dimension may represent the composite endpoint and its components, another dimension to multiple doses
  - Another dimension to multiple analysis objectives, such as non-inferiority and superiority tests for each endpoint at each dose.
- The total number of hypotheses to be tested can becomes large with even just 2-3 options for each dimension.
  - The hierarchal nature of the problem comes from considerations that the composite endpoint and its most relevant components should be evaluated first,
  - A superiority test should follow a non-inferiority test, and higher doses for multi-dose trials should be considered before lower doses.

[#]Dmitrienko, et al., 2007

Flow diagram of the test strategy (previous slide)



Method of tree-structured gatekeeping

- **Define families of hypotheses:**

  F1 = { $H1$ }, F2 ={ $H2$, $H3$ }, F3 ={ $H4$, $H5$ }, and F4 = { $H6$ }.

- **The test strategy:**

1. Test first $H1$ in F1 at the level α (e.g., $\alpha = 0.05$).
2. Test for $H2$ and $H3$ in F2 and passing of alpha to F3
   - Once the result for $H1$ is significant at level α, testing proceeds to the hypotheses $H2$ and $H3$ in F2 with the alpha that was not lost within the F1 family, which in this case is α
   - Test of $H2$ and $H3$ in F2 can be by the Bonferroni test. That is, one would test $H2$ and $H3$, each at level $\alpha/2$.
   - If both $H2$ and $H3$ are rejected then a total of $\alpha/2 + \alpha/2 = \alpha$ transfers to F3, and if only $H3$ is rejected then only alpha of $\alpha/2$ transfers to F3.
   - If $H3$ in F2 is not rejected (even if $H2$ is rejected), then there is no passing of alpha from F2 to F3. Consequently, there are no further tests because of the logical restriction.

Method of tree-structured gatekeeping

- **The test strategy** (cont'd):

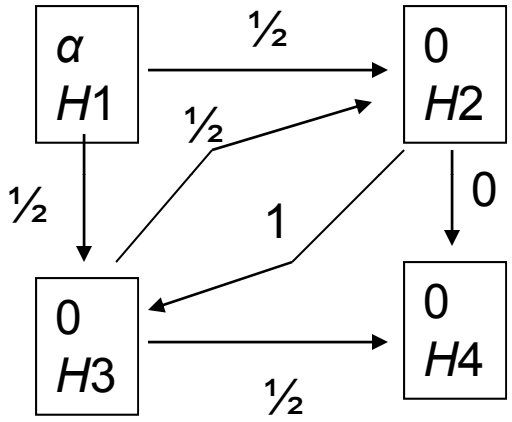
  3. Test for *H*4 and *H*5 in F3
     - Suppose that a non zero $\alpha 1$ ($\alpha 1 = \alpha$, or $\alpha 1 = \alpha/2$) passes from F2 to F3, then one would similarly test *H*4 and *H*5 in F3, each at level $\alpha 1/2$.
     - If both hypotheses in F3 are rejected then the test for *H*6 is at level $\alpha 1$. However, if only *H*5 is rejected, then this test is at level $\alpha 1/2$
     - If *H*5 is not rejected then there is no test for *H*6 because of the logical restriction.

  Note:

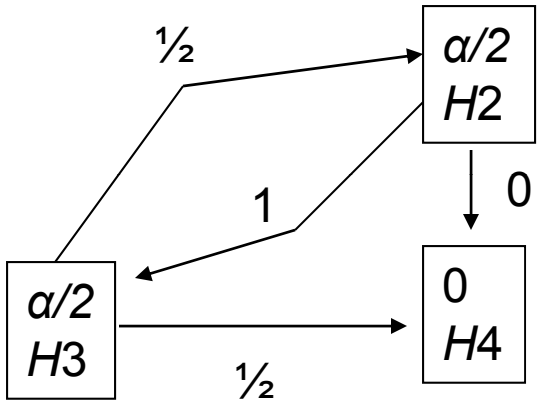
  One can also use the truncated Holm's tests in place of Bonferonni tests. The regular Holm's test does not apply because it is α-exhaustive (Dmitrienko et al., 2008).

Graphical method: NI/SU tests for the composite (d +h) and the death component (d)

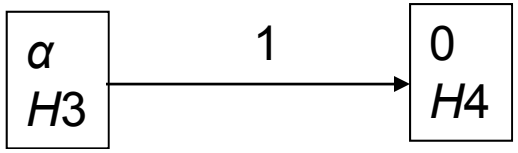
(a) Original graph



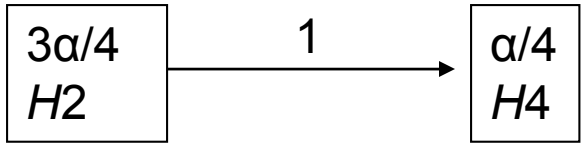
(b) Graph after rejecting *H*1



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



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



(Ref: Bretz et al., 2009)

Calculation of arrow weight *w*34 in Graph (c) when H2 is rejected in Graph (b)

- Algorithm for the new weight *w*34 for the arrow going from *H*3 to *H*4 in Graph (c)

  = (old *w*34 + A)/(1- B)

  - A = additional weight for *H*3 to *H*4 going through the rejected hypothesis *H*2 = *w*32 x *w*24 = (1/2) x (0) = 0 (calculated as proportion of a proportion)
  - B = adjustment for the arrow going from *H*3 to *H*2 and returning back to *H*3 = *w*32 x *w*23 = (1/2) x (1) = 1/2.
  - Therefore, updated *w*34 in Graph (c) = (1/2 + 0) / (1 – 1/2) = 1.

(See, Bretz et al., 2009)

Calculation of arrow weight *w*24 in Graph (d) when *H*3 is rejected in Graph (b)

- Algorithm for the new weight *w*24 for the arrow going from *H*2 to *H*4 in Graph (d)

  = (old *w*24 + A)/(1- B)

  - A = additional weight for *H*2 to *H*4 going through the rejected hypothesis *H*3 = *w*23 x *w*34 = (1) x (1/2) = 1/2 (calculated as proportion of a proportion)
  - B = adjustment for the arrow going from *H*2 to *H*3 and returning back to *H*2 = *w*23 x *w*32 = (1) x (1/2) = 1/2.
  - Therefore, updated *w*24 in Graph (d) = (0 + 1/2) / (1 – 1/2) = 1.

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

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

Results at the completion of the trial

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

The #4A method for such a trial

(adaptive alpha allocation approach)

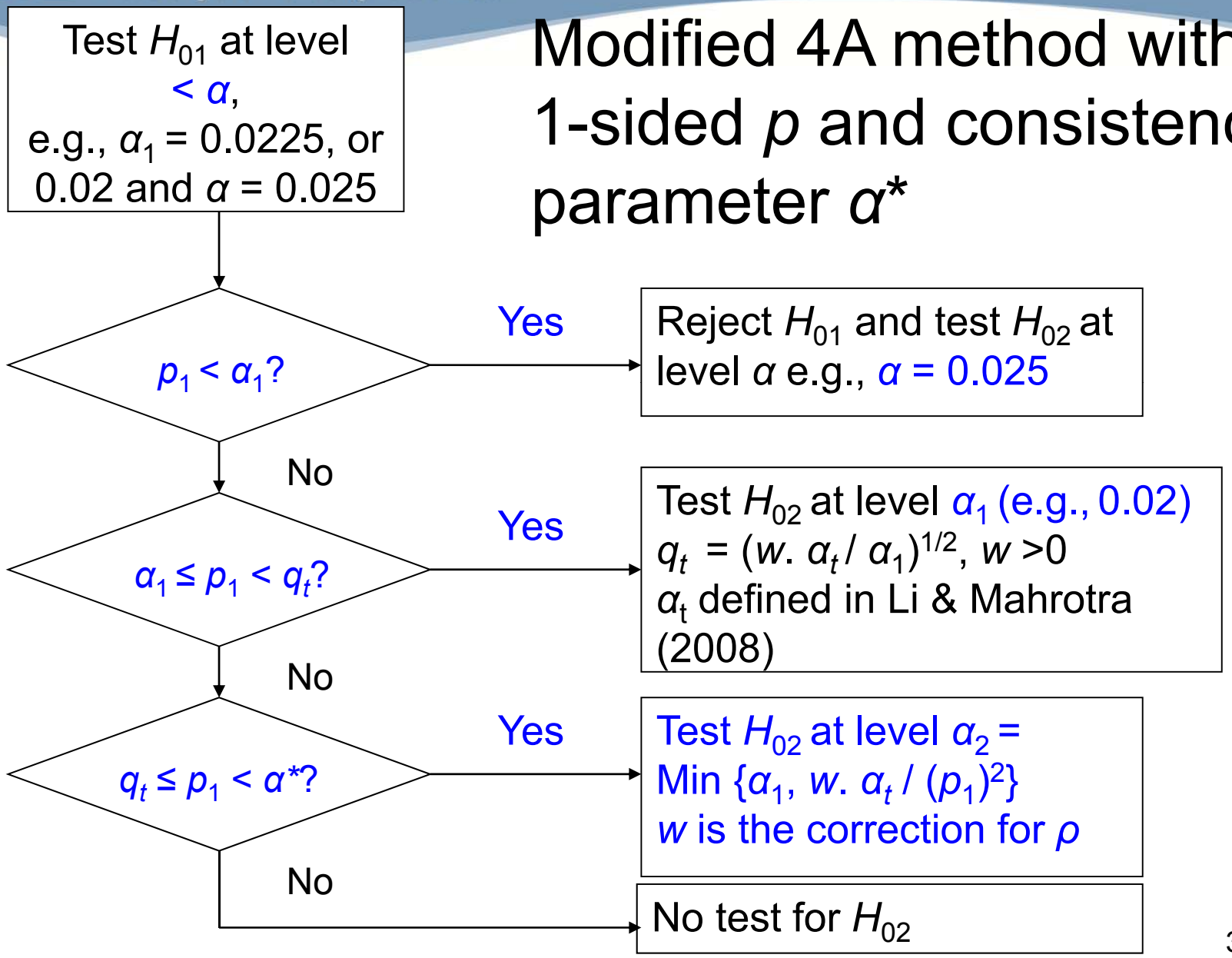
p1 = p-value for endpoint (c); p2 = p-value for endpoint (s)

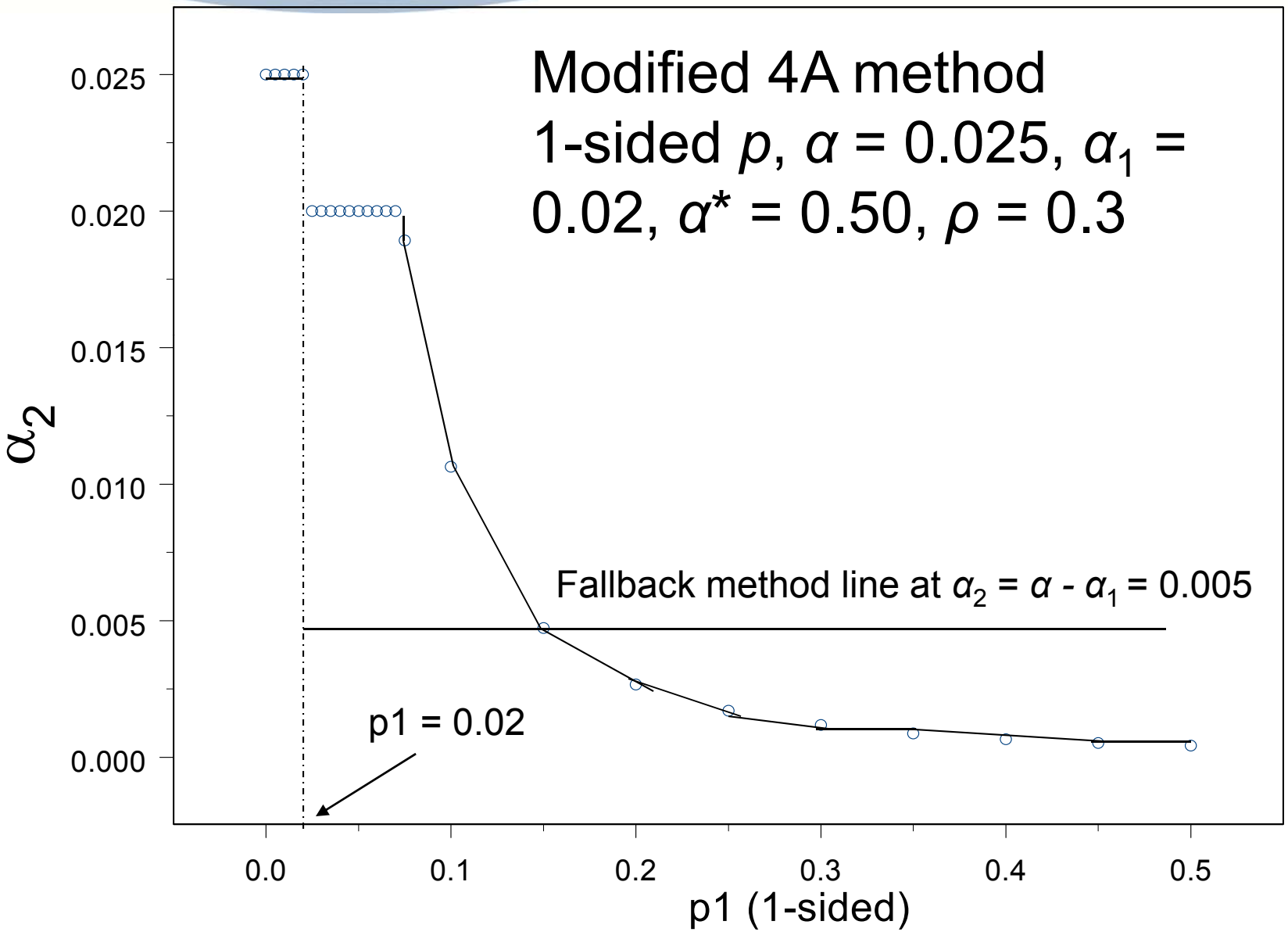
- **The 4A method**
  - Split alpha ($\alpha 1$, $\alpha - \alpha 1$)
  - E.g., (0.04, 0.01)
- If p1 < 0.04, then test p2 at level 0.05
- If p1 ≥ 0.04, then test p2 at level $\alpha 2$ (adaptive):
  - (a) $\alpha 2$ in the interval [0.04, 0.01) for small values of p1 but ≥ 0.04
  - (b) $\alpha 2 \leq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

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

#Reference: Li & Mehrotra, SIM 2008

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




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

The method of #CAS (consistency assured strategy)

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

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

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

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

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

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

Conclusions for the trial results

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

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

Testing of a composite primary endpoint for an enrichment design

- Treatment effects on a composite or on its components can be much larger for example in a biomarker positive subgroup than in the general patient population.
- Trials during randomization can be enriched to include a subgroup of patients (sensitive subpopulation) who are likely to respond better to the treatment than the rest of the patients of the trial.
- Such a trial of enrichment design can increase the success of the trial and can make the test more powerful at least for the enriched subgroup depending on the extent of enrichment.

Statistical tests for the total populations and a targeted subgroup

- Methods that ignoring correlation between the test statistics of total population and the sub-group
  - Bonferroni
  - Bonferroni Holm's
  - Fallback, fallback with a loop-back strategy
- Methods that account for correlation between the test statistics of total population and the sub-group; assumes bivariate normal probability model
  - Method similar to CAS: Alosh and Huque (2008)
  - CAAAS: Alosh and Huque (2010)
  - Zhao, Dmitrienko and Tamura (2010)
  - Others (e.g., parametric fallback)

Concluding Remarks

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