

Adaptation and Heterogeneity: how much is too much?

Armin Koch
Medizinische Hochschule
Institut für Biometrie
Carl-Neuberg-Str. 1
D-30625 Hannover

The views expressed in this paper are those of the author and not necessarily those of the BfArM, SAWP, or Hannover Medical School



**Medizinische Hochschule
Hannover**

Introduction –
we had been happy then?

Interim analyses are often an ethical mandate:

- even after sound research in phase II sufficient uncertainty about treatment effects will often remain at the beginning of phase III.

Group sequential designs

- had been developed to avoid inflation of the type-I-error associated with repeated testing of accumulating data,
- allow for a rather flexible number and timing of interim analyses,
- allow to stop the trial early for efficacy or futility.

Why did we need more?

- wish to increase sample size

Introduction – unlimited opportunities:

Landmark paper:

- P. Bauer and K. Köhne: Evaluation of experiments with adaptive interim analyses. *Biometrics* 50:1029-1041, 1994.
- *Idea*: understand study as a “pre-planned meta-analysis” of two sub-studies;
- *Flexibility*: because only P-values from stages are combined in the end;
- *Simple decision rule*: type 1 error is controlled at 2.5% (one-sided) if $P_1 \times P_2 < 0.0038$.

Is there a need to limit flexibility?

Early imagined “Viagra-type” examples:

<i>primary endpoint</i>	<i>Treatment</i>	<i>Control</i>	<i>Risk Diff. 95% CI</i>	<i>P-Value (1-s)</i>
<i>Angina responder (stage 1)</i>	249/631 (39,5%)	228/645 (35,4%)	4,1% (-1,2%; 9,4%)	0,064
<i>Sexual function responder (stage 2)</i>	30/62 (48,4%)	24/69 (34,8%)	13,6% (-3,3%; 30,5%)	0,056

$$P = P_1 \times P_2 = 0,00358$$

Introduction – unlimited opportunities?

From the original paper by Bauer & Köhne:

"Hence even the reduction in the number of components of a multiple endpoint may be a desirable goal for a protocol adaptation..."

"Finally it has to be mentioned that the interim analysis may result in strong grounds for completely redesigning the trial...."

cautious remark:

"If only the sample-size is open to adaptation, the interpretation of the decision rule is simple ...However, if the intersection hypothesis is rejected in the final global analysis this would lead to the acceptance of the respective alternative ... The resulting problem of the interpretation is the price to be paid for changing essential features of the design.

Reflection paper -
discussion about adaptive designs in *confirmatory* trials:

DRAFT

REFLECTION PAPER ON METHODOLOGICAL ISSUES IN CONFIRMATORY
CLINICAL TRIALS WITH FLEXIBLE DESIGN AND ANALYSIS PLAN

- has been out for consultation, received many useful comments, and is now finalized:

Doc. Ref. CHMP/EWP/2459/02

COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
(CHMP)

REFLECTION PAPER ON METHODOLOGICAL ISSUES IN CONFIRMATORY
CLINICAL TRIALS PLANNED WITH AN ADAPTIVE DESIGN

available from:

<http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf>

Adaptation of design specifications: Minimal requirements and general principles

- control of a pre-specified type-I-error,
- availability of corresponding methods to estimate a treatment effect and a confidence interval with correct coverage
- the additional identification problem must be addressed (differences in effects due to chance / communication of interim results / design change?)
- it is not upon regulators to question homogeneity (sponsors must justify combinability)
- the body of evidence that justifies the treatment recommendation must be identifiable (rejection of an intersection null-hypothesis may not be sufficient, no small steps)
- too many design modifications question the confirmatory nature of the trial

Reporting back from the emea/EFPIA meeting on Adaptive Designs:

Many colleagues had complaints, some argued:

Structure of my response:

1. Analogy between meta-analysis and adaptive designs
2. "Natural reasons" for different results being observed in different stages
3. Methodological aspects
4. What is the role of the heterogeneity test?
5. Summary of positions
6. ICH-E9, revisited

Arguments and questions in relation to Meta-Analyses and Adaptive Designs:

Is there analogy between adaptive designs and meta-analysis?

- Yes, in both situations you can end up with findings that are not interpretable (and heterogeneity is the indicator)

With solid procedures in place to restrict information, should we really have the same level of concern and require standards as stringent as in meta-analyses?

No, standards need to be even higher:

- MA is observational research, only.
- Although we compute P-values, there is no type 1 error.
- AD have (at least in principle) the potential to be confirmatory.

Analogy between meta-analysis and adaptive designs?

MA strategy: test for heterogeneity, if $P < 0.1$ (or $P < 0.15$) don't combine studies, is well accepted.

- great, I have promoted this for years, it costs you something, but nothing in life is free.

Will this (e.g. heterogeneity is assumed to be substantial if $P(\text{Het}) < 0.05$) be considered too high a standard (from a regulatory perspective)?

- This will lead to a too late reaction (we all know, that the test has low power)

We do not have conventional standards for acceptable homogeneity in other contexts.

- This is partly true, however, some standards exist to define, how much is too much (see above).

"Natural reasons" for different results observed in different stages

Isn't the main concern information leakage?

- No: the main concern is information leakage *and* study results that can not be interpreted in the context of drug licensing (Viagra revisited)

Gallo & Maurer (BiomJ 2006) : change in treatment effect can be a consequence of:

- *a time trend, a learning curve*
- *change / better selection in / of patient population, exhausted patient pool affecting estimate*
- *change in centre-composition*
- *different batches*

OK, there are many reasons, but does this mean, we can ignore the problem?

Methodological aspects:

Should a signal for heterogeneity not be dependent on observed overall effect strength (i.e. a signal is considered present if between stage effect difference is larger than some fraction of the pooled overall effect)?

- An excellent idea that is worth to be investigated from methodological grounds. How to define thresholds? How does it compare?

Isn't it counterintuitive that a certain degree of heterogeneity can not be compensated by a larger sample size?

- Correctly statistics says no: if treatment effects are grossly different, this should be even more important with larger sample-sizes.
- This is precisely the reason for the "no small steps" minimal requirement.

Methodological aspects:

Change point analysis could suggest change already before interim analysis (T. Friede):

- agreed: a way forward to address the leakage-issue, but doesn't help to understand, why the heterogeneity is there.

The low-power argument:

- somewhat counter-intuitive: shouldn't this help. so that we react only to situations, where treatment effects in stages are grossly different?

Does the direction of a change (first stage effect larger than second stage effect) influence the validity/ interpretation of the results, and if yes, how?

- no: independent from whether the effect is first larger and then smaller, or vice versa, the question is: is this one trial or two?

What is the role of the heterogeneity test?

We need to be very cautious about ascribing any suggested changes to knowledge gleaned from the interim analysis and decision making processes, and potentially invalidating the overall trial results on the basis of such 'bias'.

- of note: information leakage has been chosen in the reflection paper as the most untoward reason for observed differences in treatment effects:
- **trialists** can only provide reassurance that good procedures have been implemented, but you can **never** proof that they have been followed
- if there was information leakage **assessors** don't know, whether we license observed effects or hope / bias.

What is the role of the heterogeneity test?

Willi Maurer: Is this information (description of stages, investigate heterogeneity signal, discuss potential impact of adaptation, discuss and substantiate other potential sources for heterogeneity) sufficient for the regulators?

- yes, it's a signal, it's a signal, it's a signal...don't panic
- discussion is required before we can (hopefully) agree that something is (probably / most likely) a chance finding.

No acceptable strategy:

- regulator: there is a problem in your dataset
- applicant: (probably / definitely) a chance finding

Heterogeneity is nobody's fault, it's just an indicator that there is a riddle in the data that needs to be understood

Summary of positions:

In summary:

- it is difficult to define "negligible heterogeneous" or "sufficiently homogeneous" for stage / trial findings;
- if, however, a classical heterogeneity test, or another similarly (in)competent statistical test indicates discrepancies between stage findings ($P < 0,15$), this can't be overlooked;
- indication for heterogeneity requires thorough discussion;
- and yes, heterogeneity testing is secondary;
- the approach, however, requires thoughtful pre-planning;
- no, this is not double standard, it is just the price to be paid for an interim analysis (with the potential to modify the design),
- and although not primary, heterogeneity can kill a trial:

ICH E9, revisited

With this approach I feel pretty much in line with ICH-E9:

"The statistical model [...] should be described in the protocol. The main treatment effect may be investigated first using a model which allows for centre differences, but does not include a term for treatment by centre interaction. [...] In the presence of true heterogeneity of treatment effects, the interpretation of the main treatment effect is controversial.

If positive treatment effects are found in a trial [...], there should generally be an exploration of the heterogeneity of treatment effects across centers, as this may affect the generalisability of the conclusions.

It is even more important to understand the basis of any heterogeneity characterized by marked qualitative interactions, and failure to find an explanation may necessitate further clinical trials before the treatment effect can be reliably predicted.