BfArM
Bundesinstitut für Arzneimittel
und Medizinprodukte

# Heterogeneity:
# how much is too much?

Armin Koch
Bundesinstitut für Arzneimittel
und Medizinprodukte
Kurt-Georg-Kiesinger Allee 3
D-53175 Bonn

The views expressed in this paper are those of the author and not necessarily those of the BfArM

# Overview:

Many questions have been posed by Dres Maurer, Friede, and Anderson.

This is the 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. Example
7. ICH-E9, revisited

# Analogy between meta-analysis and adaptive designs?

***Dr. Maurer:*** *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?

***Dr. Maurer:*** *Adaptive designs which do not make major structural changes have none of these problems (i.e. regions, centers, monitoring standards, change in practice may differ).*

- "The general combination principle allows for early stopping and all sorts of adaptations, including changes between doses […], dropping treatments […], selecting suitable endpoints, and reassessing sample-sizes.

(Brannath, Posch & Bauer, JASA 2002)

So then:

- can we agree that not all changes (that may be possible from statistical grounds) are reasonable in late stage drug development?
- can we agree that there is no room for major structural changes in late stage drug development?

# Analogy between meta-analysis and adaptive designs?

***Dr. Friede:*** *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.

***Dr. Maurer:*** *Will this (e.g. heterogeneity is assumed to be substantial if P(Het) < 0.05) be considered too high a standard (from a regulatory perspective)?*

- No, too low (we all know, that the test has low power)

***Dr. Maurer:*** *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

***Dr. Maurer:*** *The main concern is 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)), Anderson, Friede (here):*** *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:

***Dr. Maurer:*** *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?

***Dr. Friede:*** *Can the power loss 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:

***Dr. Friede:*** *Change point analysis could suggest change already before interim analysis*

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

***Dr. Maurer:*** *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?

***Dr. Maurer:*** *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:
- ***you*** 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 ***we*** don't know, whether we license observed effects or hope / bias.

# What is the role of the heterogeneity test?

***Dr. 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 (Anderson): 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:

# Final example:

Just a comparison of a cream with its vehicle in a multi-center clinical trial (results are change from baseline in a score):

**Table 6:** results for 5 centres each recruiting 10 patients in treatment and control group

| | Treatment | | Placebo | | estim ate | weight | P-value* |
|---|---|---|---|---|---|---|---|
| ID | Mean | Sd | Mean | Sd | | | |
| C1 | 55.4 | 31.39 | 61.0 | 25.39 | -5.6 | 9.21 | 0.6608 |
| C2 | 70.5 | 26.03 | 71.5 | 16.19 | -1.0 | 15.95 | 0.9178 |
| C3 | 48.9 | 23.79 | 37.3 | 31.59 | 11.6 | 9.58 | 0.3536 |
| C4 | 31.3 | 14.56 | 0.2 | 8.19 | 31.1 | 53.70 | 0.0001 |
| C5 | 59.1 | 24.29 | 51.8 | 26.57 | 7.3 | 11.56 | 0.5214 |

* 2-sample t-test

# 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.