Considerations on guideline requirements in relation to homogeneity

Willi Maurer, Paul Gallo

Novartis Pharma AG, Novartis Pharmaceuticals

EMEA/EFPIA WORKSHOP ON ADAPTIVE DESIGNS IN CONFIRMATORY CLINICAL TRIALS

Dec 14th 2007

NOVARTIS

Agenda

- Reflection paper on adaptive designs: Justification of combination
- Meta analyses and heterogeneity within clinical trials
- Formal statistical investigation of homogeneity
  - Possibilities, issues and consequences
- What to do? Plan and investigation
- Conclusions

CHMP reflection paper; the concern

- "Studies with interim analyses where there are marked differences in estimated treatment effects between different study parts or stages will be difficult to interpret.
  - It may be unclear whether the estimated treatment effects differ just by chance, as a consequence of the intentional or unintentional communication of interim results, or for other reasons."
- "This problem can be even greater if the study design has been changed as a result of an interim analysis."

CHMP reflection paper; justification of combination

- "From a regulatory point of view, whenever trials are planned to incorporate design modifications based on the results of an interim analysis, the applicant must pre-plan methods to ensure that results from different stages of the trial can be justifiably combined."
  - "….studies with adaptive designs need at least the same careful investigation of heterogeneity and justification to combine the results of different stages as is usually required for the combination of individual trials in a meta-analysis."
- "Depending on the nature of the design modification, the simple rejection of a global null hypothesis across all stages of the trial may not be sufficient to establish a convincing treatment effect."

Comparison to meta-analysis standards

- Studies included in a meta-analysis may differ in important ways:
  - regions, centers, investigators following different protocols and procedures, monitoring standards, calendar times, information from other trials out in the open, no common randomization, therapeutic drift (change in medical practice/standard of care), etc.
- Adaptive trials which do not make major structural changes have ***none*** of these problems
  - In systematic reviews the meta-analysis is post hoc, in AD it is pre-planned and same data collection methods are used
  - Changes in confirmatory adaptive designs (e.g. sample size , dropping an arm) do not alter the primary hypotheses
  - The main concern is information leakage
- 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?

Implementation issues

- We have not been in the habit of routinely looking for this type of ‘drift’ within trials.
- We probably can expect to find signals frequently.
  - Subsets of trial data tend, of course, to be highly variable. This arises in other contexts:
    - Li, Chuang-Stein, Hoseyni (DIJ 2007) on subgroup effects
    - Senn (1997), on qualitative interactions in multicenter trials
  - Both quantified that outcomes in opposite directions can occur with surprising frequency *even when there is true homogeneity*.
  - Aren’t some changes natural? (drift in patient population, “learning curve”, etc.)
- **We do not have conventional standards for acceptable homogeneity in other contexts.**

Real-life example

- Long-term (non-adaptive) trial, monitored by a DMC, no information released.

"Stage"	# events		% benefit
	Control	Experimental	
1	146	150	-2.7
2	574	473	17.6
3	81	100	-23.5
Total	801	723	9.7

- Interpretation?

Formal statistical investigation of homogeneity

- Are there formal statistical approaches that could be applied ?
  - Pre-specify a plan for dividing the data into stages according to when adaptations took place
  - Apply some pre-agreed thresholds for lack of homogeneity across stages that would suggest bias (or at least 'non-combinability' of results).
- High level of control for both potential types of errors (re. heterogeneity signal) is not possible
  - Use conventional statistical significance level for proof of efficacy (e.g. 5%) for a "test" of interaction?
    - Will this be considered too high a standard (from a regulatory perspective)?

Formal statistical investigation of homogeneity (2)

- Use a „meta-analysis“ level of significance for heterogeneity ?
  - If a heterogeneity signal subsequently leads to an invalidation of an otherwise ‘significant’ treatment effect it has a serious impact on power
    - If heterogeneity is declared e. g. at a significance level of 15%, the power to confirm an overall effect would be reduced by 15% of the original power even if there is in fact no heterogeneity!
  - Such an automatic procedure would prevent one from conducting an adaptive design!
  - It would be particularly disturbing if the observed effect sizes in both stages are strong
    - in this case there is no good reason to doubt that there is a clear overall effect
- 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

Formal “test” and causality

- What is the role of such a „test“?
  - A final primary requirement automatically leading to an invalidation of an otherwise significant finding ?
  - A secondary, more exploratory requirement that could, after further investigations of the type, magnitude, and possible causes, lead to a potential change in estimation or interpretation of effects?
- 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’.
  - There are other possibilities besides a causal relation to the adaptation
    - General time shift in influencing factors (learning, centers, population..)
    - Chance...

Direction of a heterogeneity signal

- Assume that a heterogeneity signal in the treatment effect is present
  - If the main concern is bias caused by the design adaptation (analysis details „leaked“ to trial participants, or information deduced from changes that cannot be concealed) then
    - smaller efficacy effects after adaptation than before, do not suggest an ‚intentional‘ bias or a bias favorable to the sponsor
    - if the effects after adaptation are better, one might argue that this might be similar in a future phase III trial where much more information about the results of preceding trials would be available?
  - Does the direction of a change influence the validity/ interpretation of the results, and if yes, how?

Additional practical implementation issues

- Can the “cutpoints” be unambiguously defined?
  - e.g., data used for the IA, time of the DMC meeting, announcement of decision, implementation of adaptation, etc. – these might be very different
- *Patients* vs. *outcomes*: is it clear how we should handle patients enrolled prior to the IA, but whose outcome is assessed after the adaptation?
- What factors should be adjusted for in this investigation?
  - Should we search for explanatory covariates? If we find them: does this resolve the concern? or have we just characterized the problem?
- Time-to-event outcomes: this issue will be highly confounded by any non-constancy of hazard ratios.

So what do we do ?

- ***Not*** ignore the issue.
  - This is an interaction question, and as with certain other types of interactions, should be considered to be an important secondary concern.
- Effective processes to restrict access to the interim data are critical
  - In the presence of suggestive heterogeneity, concerns about leakage and bias should be far less if it can be documented how the information was controlled.

So what do we do upfront ?

- A Proposal:
  - Implement and document all possible measures to protect confidentiality
  - * Consider whether the nature of the adaptation inherently conveys too much information, and how this might compromise trial results
    - * Discuss *mechanisms* by which we expect that 'bias' might be introduced,
  - * Describe staged analysis plan to investigate possible heterogeneity and consequences
  - * Discuss and define choices for analysis details
    - * e.g., definition of analysis cutpoints, homogeneity analysis method and model (incl. choice of covariates), handling of patients on trial across multiple stages, etc.
- Is this* feasible? Can we move to a more standard approach?

What could be done at final analysis ?

- Conduct primary analysis according to plan, e.g.
  - Present descriptive statistics by stage
  - Investigate whether a heterogeneity signal is suggested in the primary efficacy parameters, and if so
    - Investigate if a change occurred in one of the areas that was considered as potentially being affected by the adaptation (e.g. a modified risk population)
    - Explore if these changes could be due to a 'natural' time effect e.g. via a changing covariate independent of adaptation
  - Discuss relevance on overall claim of an effect and best approach for estimating the effects (pre-planned adaptive methodology, inclusion of covariates for estimates, concentration on subgroup results...)
- Is this information sufficient for the regulators?

Conclusions

- Investigation of heterogeneity is an important secondary concern
  - Is an additional „incentive“ to implementing and documenting processes to maintain trial integrity
- Planning and conducting such a homogeneity analysis is not straightforward
  - There are fundamental differences between AD and meta-analyses /systematic reviews
- We must find a reasonable balance between
  - identifying ‚real‘ heterogeneity and its consequences on the interpretation of results and
  - unjustified invalidation of a pivotal trial showing a true treatment effect.
- Further discussion, experience and a consensus on how to do that is necessary

Backup and draft slides

CHMP reflection paper; justification of combination

- "From a regulatory point of view, whenever trials are planned to incorporate design modifications based on the results of an interim analysis, the applicant must pre-plan methods to ensure that results from different stages of the trial can be justifiably combined.
  - The justification should include consideration of the impact of the modification on patient population, schedule of assessments, patient management and other features of the trial, which, if affected, would complicate the interpretation of the trial.
  - In this respect, studies with adaptive designs need at least the same careful investigation of heterogeneity and justification to combine the results of different stages as is usually required for the combination of individual trials in a meta-analysis."
- "Depending on the nature of the design modification, the simple rejection of a global null hypothesis across all stages of the trial may not be sufficient to establish a convincing treatment effect."
- "Addressing such issues at the planning stage of the trial is essential to avoid post-hoc discussions about whether observed data may indicate that combining results from different stages of the trial is justified or not.
  - Hence, trials should not be planned to make many changes along a series of small steps with limited numbers of patients at each step. Insufficient information will be available from such trials to justify the consistency of the treatment effect after a design modification."

Measures and types of heterogeneity

- To shape ideas, assume:
  - 2-stage design, roughly equal amount of information (e.g. sample size) per stage (S_1 and S_2)
  - Primary effect size is measured by standardized statistics Z_1 and Z_2 in comparison to a control
  - Pooled standardized effect size $\quad Eff = (Z_1 + Z_2)/\sqrt{2}$
  - Measure of heterogeneity of effect size; e.g. $\quad Int = (|Z_2 - Z_1|)/\sqrt{2}$
- Questions:
  - Sensible threshold for „Int“ to signal heterogeneity ?
  - Is $Z_2 > Z_1$ more relevant than $Z_1 > Z_2$?
  - If heterogeneity is flagged, what type is it: „qualitative“ or „quantitative“, continuous time effect or a ‚jump‘ at adaptation time?

Consequences of a heterogeneity signal

- Reduction of power for an efficacy claim:
  - Assume a „formal“ heterogeneity signal (e.g. Int > $z_{0.85}$) would turn an otherwise „significant“ effect into „non-conclusive“
  - Reduced power for efficacy =
    (Power of efficacy)*(Probability of no heterogeneity)
    since „Eff“ and „Int“ are independent (**irrespective of true overall effect**)
  - If e.g. (‚true‘) power is 95% and the type I error rate for heterogeneity signal is 15%, then reduced power is 0.95*0.85 = 80.8%

Initial „inhomogeneity signal“: dependent on effect size?

- The invalidation of a positive efficacy signal by a false inhomogeneity signal
  - is particularly disturbing if the observed effect sizes in both stages are strong
    - in this case there is no good reason to doubt that there is a clear overall effect, even if there is in fact some possibility of operational bias
- Rejection of homogeneity should be, in some way, dependent on overall effect
  - A possibility: Signal for inhomogeneity if Int > c*Eff (e.g. with c = 0.7)
    - In particular for stronger true treatment effects, false inhomogeneity signals are more rare and the power loss on efficacy would become more acceptable
      - e.g. for c = 0.7, initial power = 95%, the reduced power would be ~92.5%