EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

25 July 2014

Submission of comments on '<Guideline on the investigation of subgroups in confirmatory clinical trials>' (EMA/539146/2013)

Comments from:

Name of organisation or individual
Stephen Senn, Franz König, Ralf-Dieter Hilgers and Geert Molenberghs on behalf of the IDEAL team (IDEAL is a research project funded by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552, http://www.ideal.rwth-aachen.de/) (Contact: Stephen Senn stephen.senn@crp-sante.lu)

Please note that these comments and the identity of the sender will be published unless a specific justified objection is received.

When completed, this form should be sent to the European Medicines Agency electronically, in Word format (not PDF).

General Remark (not only related to this guideline)

Also according to the EMA webpage the overview of comments received during the consultation period [*] should be accessible via the webpage. In light of the EMA initiatives on transparency, we are wondering why for the most recent guidelines, concept paper or reflection papers on statistical topics (adaptive designs, multiplicity, missing values, adjustment for baseline covariates, extrapolation, …) the comments sent to EMA and also the outcome (last column in the comments template, to be added by EMA) have not been published yet. We would ask EMA to publish the comments and outcomes (if and how the comments have been addressed in the final documents) for transparency reasons also for statistical guidelines.

[*]
http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000043.jsp&mid=WC0b01ac05800240cb

1. General comments

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	Although this draft guideline[1] makes many interesting points there are a number that are controversial and even some that do not appear to be well thought-out. This is no doubt a reflection of the fact that this difficult field is one for which methodologists are still struggling to agree rational recommendations. Of course this is an unsatisfactory situation but it raises the question as to whether a guideline of this sort is not premature. Although one may feel that it is better to have some guidance than none, the danger is that ill thought-out recommendations become frozen as standard practice. Our general feeling is that this guideline needs to go back for major revision and that it may need face-to-face meetings between regulators, sponsor and other interested parties to come to a sensible conclusion.	
	The guideline as a whole risks allowing 'the best to become the enemy of the good' (to use a phrase employed by Klim McPherson in another context[2]). If a treatment has convincingly shown that it is superior to control *on average* then although it is true that this does not prove that it is superior to control for every subgroup, it is irrational to prefer to continue to use the control for that reason since there are no grounds at all for believing that the control is better for every subgroup. Therefore, it is a mistake to take the point of view that it is necessary for the sponsor to demonstrate efficacy in key subgroups. The guideline needs to recognise that some values needs to be placed on the inferentially conservative position that in the absence of	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	convincing evidence to the contrary it is rational to act as if a treatment that is better on average is the right one to choose. This does not preclude investigating subgroups that respond differently but it does suggest that one should be cautious and it also suggests that this task of personalising treatment may be rather different from the general one of deciding whether a license should be granted or not.	

2. Specific comments on text

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
Lines 4-5		Comment: The title of the draft reflection paper seems misleading in some respects. The draft title of “guideline on the investigation of subgroups in confirmatory clinical trials” suggests that confirmatory conclusions would be the main (or at least an important) part of this paper. Disappointingly the draft guidance document does not tackle this internationally controversially discussed topic, but refers to the more general reflection paper on multiplicity only Proposed change (if any): The guidance should either be expanded by section on confirmatory subgroup testing or the title should be changed in such a way that it clearly indicates that the draft reflection paper focuses on exploratory subgroup analyses only (“Guideline on the exploratory investigation of subgroups in confirmatory clinical trials”).	
Lines		Furthermore, the guidance document should include categories on how the different levels of such exploratory subgroup analyses should be labeled in the study protocol, SAP and final report. Otherwise, everyone will generate and use different terms clouding the different levels of evidence which can be derived by the different approaches of a-priori specification or	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		post-hoc analysis. Proposed change (if any): Define categories such as: • A-priori specified confirmatory subgroup analysis: the factor and anaylsis was pre-specified for the confirmatory primary analysis in the study protocol. This subgroup analysis is incorporated in the control of the studywise type I error rate. • Fully pre-specified exploratory subgroup analysis (the factor and analysis was specified as exploratory analysis in the study protocol. Though the subgroup analysis was not incorporated in the control of the studywise type I error control, this specification would allow assessing the risk of false positive conclusion in the family of hypotheses given by all fully pre-specified exploratory subgroup analyses). • Pre-specified subgroups: means that the subgroups were specified in advance, but not the details how the subgroups will be analyzed statistically. • Fully (post-hoc) exploratory subgroup analysis: Neither the subgroup nor the analysis strategy have been	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		stated in the study protocol.	
Line 225-233 & elsewhere		Comment: Although the guideline recognises on occasion (see for example, the first paragraph of section 4.2) the crucial distinction between heterogeneity in the target population and heterogeneity in the clinical trial, it does not really come up with practical guidance to the sponsor. The danger is that 'out of sight is out of mind' a homogenous but unrepresentative trial may offer little in the way of reassurance that there will not be heterogeneity of response in the target population. Proposed change (if any): The language used should be more careful in this respect. Advice on designing trials should be given. The EMA may have to agree to give specific advice on inclusion criteria when trial are designed if exploration of heterogeneity is to be made workable.	
Lines 58-59 & elsewhere		Comment: The guideline fails to distinguish adequately between covariates (and subgroup classifications) that are *prognostic* on the one hand and *predictive* on the other (to use the distinction that is made in the biomarker literature). The latter are sometimes referred to in the epidemiological literature as *effect modifiers* and the former as *confounders*. In	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		the statistical literature one might distinguish between main effects of covariates and covariate-by treatment interaction. For example P2 lines 58-59 mentions heterogeneity and covariate adjusted analysis, as if the latter addressed the former whereas, in fact, analysis of covariance is most commonly used to improve precision by taking account of the main effect of covariates. Another example is section 5.3 where it is mentioned that centre was recommended as a stratification factor in ICHE9[3] because strongly prognostic, argues in favour of replacing this approach with the use of countries but then talks about consistency, which is essentially an aspect of stratum- by-treatment interaction rather than prognostic difference between strata. Proposed change (if any): A clear distinction is made between main effect differences of covariates & sub groups and interactions of these with the effect of treatment.	
Lines 400-402		Comment: The report recommends, lines 400-402 'that a sufficient number of patients are recruited to the subgroup to ensure an estimate of effect that can be made with reasonable precision' as if this was a practically realisable objective. Frequently it is not, for the simple reason that in many disease indications	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		adequate precision is only achievable for the trial as a whole. If the relevant subgroup is not common, a strategy of oversampling from the subgroup, with separate randomisation lists is needed. This will typically mean that recruitment will have stopped from other groups while continuing in the rare subgroup thus producing cohorts of patients who are not coeval and adding to problems of interpretation quite apart from delaying the trial. In our opinion this is a 'throw-away' recommendation that has not been thought out. It should be backed up by some actual calculations that show a) what is considered adequate precision b) how small the subgroup can be for this still to be a reasonable goal and c) what the likely impact on time to complete and overall cost of such a recommendation might be. Proposed change (if any): This requirement is unrealistic and impractical and should be dropped	
		Comment: There is no mention of transformations in the document, yet appropriate transformations can often reduce heterogeneity. Skilful exploitation of this fact can mean that heterogeneity on the clinically relevant scale can be powerfully addressed using an additive scale. Thus, (for example) analysis on the log-odds scale might reveal little difference between subgroups but if they have different levels of	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		background risk would translate into relevant differences on the risk difference scale[4, 5]. Proposed change (if any): This possibility should be explicitly recognised in the guideline.	
Lines 320-321		Shrinkage methods can be useful but are difficult to implement unless a hierarchical model with many observations at the higher level of hierarchy can be used (for example when there are many subgroups that are deemed exchangeable such as might be the case for centres in a multi-centre trial). Proposed change (if any): The distinction between cases where a shrinkage factor can be reasonably estimated from the data-set and those where it cannot should be made. The analogy to random effect meta-analysis could be made.	
Lines 497-570		Comment: The discussion of consistency is inadequate. Although it points to the dangers of multiplicity it fails to realise the seriousness of the problem or, for example, that if the number of subgroups is reasonably large effect reversal are almost inevitable. For example, with 6 subgroups a type I error rate and an overall power of 80% the probability of at least one subgroup showing having a sign reversal (treatment	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		estimate worse than control) even though there is no heterogeneity is at least one half[6]. Proposed change (if any): Use either a formal test of interaction or a Galbraith plot [7, 8] to judge heterogeneity.	
Lines 531-532		Comment: Forest plots are a bad way to investigate heterogeneity Proposed change (if any): Galbraith plots[7, 8] should be mentioned as a superior alternative	
Lines 548-551		Here and elsewhere *ad hoc* solutions are provided with little or nothing in the way of logical justification and no attempt to investigate the consequences. Calculations (available on request) suggest that the difference between the effect in a sub-group and the effect in all other patients must be nearly 8 times the standard error of the overall treatment effect to trigger a warning. Proposed change (if any): Despite what is stated elsewhere in section 6.1 (for example lines 506-507), formal tests of lack of heterogeneity (backed	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		up by suitable graphical methods such as the Galbraith plot) should remain the main way of judging heterogeneity. The guideline is right to state that absence of evidence is not evidence of absence. However, that is beside the point. The question is whether when a treatment has been proved effective on average it is appropriate to deny patients its use simply because one cannot prove that all of them will benefit. Thus it may be logical to act as if there is no problem even if the most one can do is fail to prove there is a problem.	
Lines 81-82, 351-358, & elsewhere		It is a surprise that in a methodological guideline so much emphasis is put on in principle not always empirically provable issues (for example “biological rational”, “external evidence”, …). It seems that here regulators introduces a Bayesian way of thinking to incorporate such expert opinion in the evaluation of data of confirmatory clinical trials, though it has not been explicitly formalized. Proposed change (if any): In a methodological guideline it should be explained how this different sources of information should be combined in a formalized, reproducible way.	
Lines 380-384 and 433-435		The CHMP PtC on multiplicity issues in clinical trials is much clearer for which (limited) number of important factors exploratory subgroup analyses should be provided. The draft guideline lacks on giving concrete examples of which subgroups would be of key interest for exploration of the treatment effect. But one would expect such a concrete	

Line number(s) of the relevant text *(e.g. Lines 20-23)*	Stakeholder number *(To be completed by the Agency)*	Comment and rationale; proposed changes *(If changes to the wording are suggested, they should be highlighted using 'track changes')*	Outcome *(To be completed by the Agency)*
		guidance in a guideline on this topic. If the guideline remains unclear in this respect, a reference that this should be discussed with assessor at the planning phase could be critically questioned for economic reasons[9] Proposed change (if any): At least for important indications concrete examples of important subgroups should be given.	
Lines 645		The way how the three scenarios are presented may suggest that all three scenarios are of equal importance, which they are not. Scenario 2 and 3 should remain the exception, which can only be decided on a case by case basis. Neither should an assessor restrict the indication too easily [10], nor should the door be opened to rescue a failed trial by post-hoc subgroup analysis. However, the text in lines 712 to 717 indicates that regulators are open for this. If the scenarios 2-3 are really current practice in regulatory decision making, it would be important to cite real case examples and whether they turned as good or bad decisions later on or not. Proposed change (if any): Shorten text for Scenario 2. Delete lines 712 to 717. Include real case examples of scenarios 2 and 3 illustrating the way of decision making when approving drugs in Europe.	

References

1. European Medicines Agency, *Guideline on the investigation of subgroups in confirmatory trials*. 2014, European Medicines Agency: London. p. 20.
2. McPherson, K., *The Cochrane Lecture. The best and the enemy of the good: randomised controlled trials, uncertainty, and assessing the role of patient choice in medical decision making.* J Epidemiol Community Health, 1994. **48**(1): p. 6-15.
3. Lewis, J.A., *Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline.* Statistics in Medicine, 1999. **18**(15): p. 1903-42.
4. Glasziou, P.P. and L.M. Irwig, *An evidence based approach to individualising treatment.* British Medical Journal, 1995. **311**(7016): p. 1356-9.
5. Senn, S.J., *Added Values: Controversies concerning randomization and additivity in clinical trials.* Statistics in Medicine, 2004. **23**(24): p. 3729-3753.
6. Senn, S.J., *Individual Therapy: New Dawn or False Dawn.* Drug Information Journal, 2001. **35**(4): p. 1479-1494.
7. Galbraith, R.F., *A Note on Graphical Presentation of Estimated Odds Ratios from Several Clinical-Trials.* Statistics in Medicine, 1988. **7**(8): p. 889-894.
8. Galbraith, R.F. and J.I. Galbraith, *On the graphical presentation of a collection of means.* Journal of the Royal Statistical Society Series a-Statistics in Society, 1996. **159**: p. 611-613.
9. Wise, J., *European Medicines Agency is attacked over proposal to allow technology assessment bodies to sell advice to drug industry.* BMJ, 2014. **349**: p. g4674.
10. Eichler, H.-G., et al., *The risks of risk aversion in drug regulation.* Nature Reviews Drug Discovery, 2013.