EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

<Date of submission>
2017-05-22

Submission of comments on '<document title>' (EMA/…/…)

Guideline on multiplicity issues in clinical trials, Draft, EMA/CHMP/44762/201

Comments from:

Name of organisation or individual
Joint Working Group “Adaptive Designs and Multiple Testing Procedures” of the ROeS and the German Region of the IBS. Contributions by (in alphabetic order): Andreas Faldum, Silke Jörgens, Robert Kwiecien, Rene Schmidt, Susanne Urach

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).

1. General comments

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	The draft guidance on multiplicity issues in clinical trials is an updated version of the EMA Points to Consider on Multiplicity issues in clinical trials (EMA/286914/2012) with an additional section about multiplicity issues in estimation (Section 10). Although it aims at giving some instruction on "how to deal with multiple comparisons and control of type I error in the planning and statistical analysis of clinical trials" (line 55), it mainly deals with the question in which cases one should adjust for multiplicity and not how this should be done. Overall adjustment for multiplicity is recommended whenever there is the opportunity to choose the most favourable result from two or more analyses on which a regulatory claim can be based. Details of all planned analyses and the multiplicity procedure should be laid down in advance in the study protocol. Multiple level alpha tests which split the type I error rate between the null hypotheses are discussed quite generally in lines 173-179. The only explicitly mentioned adaptation method is the Dunnett test on multiple comparisons to a single control (lines 178-179) which in addition makes use of the correlation between the test statistics. Hierarchical testing (lines 225-244, line 319, line 379, lines 402-406) is introduced as both a case where one does not have to adjust for multiplicity and a multiplicity	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	adjustment method. This overemphasis of the hierarchical testing and alpha splitting procedures could be seen as a recommendation to use these methods above all others. The acceptability of multiple testing procedures which take the correlation between endpoints into account would be of interest as this nuisance parameter is usually estimated from the sample and strict type I error rate control can only be achieved asymptotically. Also the regulatory opinion on resampling based methods such as bootstrap or permutation approaches (e.g. the minP test by Westfall and Young) would be worth knowing. Global testing procedures are only shortly acknowledged in line 336 for showing that a drug combination is more effective than either component. Although mainly applied in cases where it is enough to show an effect in at least one of the endpoints, global tests such as O'Brien's GLS and OLS test could lead to more powerful tests of multiple endpoints and together with the closed testing principle allow inferences also for the individual hypotheses.	
	The situation of multiple randomizations per patient should be discussed briefly. Sometimes, multiplicity is ignored, if in the first randomization two treatments are compared (treatment A vs. treatment B), and in the second randomization, in each treatment arm two further treatments are compared (e.g. maintenance treatment A vs. maintenance treatment B). Finally, we	

Stakeholder number *(To be completed by the Agency)*	General comment (if any)	Outcome (if applicable) *(To be completed by the Agency)*
	have then 4 arms. However, sometimes the study designer considers such situation two separate trials, and ignores multiplicity.	
	The situation of drop the loser designs should be discussed briefly. Sometimes multiplicity is ignored, because in the final analysis of a (common) drop the loser design only the remaining “winner”-arm vs. the control arm are compared, i.e. only a single test will be finally performed. However, ignoring multiplicity in this situation is illegal, because the whole set of null hypotheses, that could be potentially rejected in a singly trial / study has to be taken into account. Otherwise, we have again a selection bias of the p-value.	

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)*
87-88 *369-371*		Comment: The term "claim" is defined twice in the document which seems a bit redundant.	
104		Comment: "Examples of both situations will be discussed later." Please refer to a section number here.	
109		Comment: "study-wise type I error in the strong sense". Please add also the widely used synonym "family wise error rate in the strong sense".	
119-122		Comment: It should be clarified that multiplicity issues are not specific to the frequentist approach but are likewise present with alternative approaches in statistical decision theory. If this aspect remains unclarified, the user might be incited to reformulate a frequentist approach e.g. in a Bayesian framework with non-informative prior while pretending that multiplicity issues are "resolved" this way. However, this is incorrect. Proposed change (if any): Add the following sentence in line 122 following "… specified level α.": "However, multiplicity issues are not specific to the frequentist	

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)*
		approach but similarly arise with alternative approaches in statistical decision theory e.g. the Bayesian approach, and have to be addressed adequately (c.f. FDA-Guidance "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials", section 4.9, https://www.fda.gov/MedicalDevices/ucm071072.htm)."	
136-139		Comment: Designs with data-dependent treatment arm selection should be mentioned, since diverse types of multiplicities arise with these design that have to be considered adequately. Proposed change (if any): "For example, there is a rapid advance in methodological richness and complexity regarding interim analyses, with the possibility to stop early either for futility or with a claim for efficacy, or stepwise designed studies, with the possibility for adaptive changes in the trial's next steps including data-dependent treatment arm selection (multi-arm multi-stage designs)."	
135-141		Comment: Multiplicity issues in adaptive designs are only marginally covered in the CHMP Reflection Paper on Methodological issues in Confirmatory Clinical Trials planned with an Adaptive Design. Multiple testing procedures in the context of adaptive designs are only considered for the case of multiple treatment arms.	

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)*
		Proposed change (if any): A paragraph about the testing of multiple endpoints in the context of adaptive designs should be included, especially about the influence of the stopping behaviour on the study-wise type I error rate and therefore on the multiplicity adjustment.	
142-146		Comment: The case of repeated measurements of the primary variable is not considered because it is claimed that usually a summary measure is predefined or the evaluation is reduced to a certain time point. Based on a generalised linear mixed model one would usually perform some global testing procedure like O'Brien's OLS or GLS test to find an overall treatment effect. Proposed change (if any): As global testing procedures are also relevant for composite endpoints and in conjunction with the closed testing principle for multiplicity adjustments, it would be advantageous to include a section about them in the guideline.	
167		Comment: “there are situations where no multiplicity concern arises, for example, having a number of primary hypotheses for a number of primary endpoints that all need to be significant so that the trial is considered successful” This is a conservative approach. In this situation, even relaxed rejection bounds might be possible, if only the rate of the false positive confirmatory conclusion has to be controlled.	

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)
		However, usually it is a hard problem to derive the (asymptotic) joint distribution of multiple test statistics. Hence, the conservative control of the study-wise type I error rate in the strong sense described above is usually the only way.	
173		Comment: "control the overall type I error rate α" The definition of "overall type I error rate" is not clear. There are many concepts of "overall type I error rates". Proposed change (if any): Please replace "overall type I error rate" by "study-wise type I error rate".	
174		Comment: Suggested replacement Proposed change (if any): "type I error rate study-wise" Replace it by "study-wise type I error rate".	
181		Comment: "of the chosen multiplicity procedure should be part of the study protocol and should be written up without room for choice." Please add, that adaptive designs are an exception at this point. The very concept of adaptive designs is to enable flexibility, including some non-data driven decisions during the course of a clinical trial (e.g. modification of multiple testing strategy).	

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)*
193-195		Comment: It is mentioned that sample size planning can become complex due to different alternative hypotheses. Proposed change (if any): It should be mentioned that in case of multiple endpoints there are different power definitions like disjunctive and conjunctive power in addition to the power of the individual hypothesis tests. Which power definition to choose for the sample size planning will depend on the objective of the trial.	
295-298		Comment: The statement that "there is no control of the type I error for a single hypothesis" if the tests for the safety variables are not adjusted for multiplicity is a bit misleading because type I error is maintained for each separate hypothesis test. Proposed change (if any): "there is no control of the study-wise type I error rate for the overall testing procedure"	
335-336		Comment: The phrase “global test strategies” should be explained.	
457-459		Comment: According to the guideline responder analysis may be used if it is difficult to interpret small but statistically significant	

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)*
		improvements in the mean level of the primary endpoints. As dichotomization into responder and non-responder leads to a loss in power and definitions seem kind of arbitrary, it remains questionable if such an analysis should not be discouraged altogether in case there are other primary endpoints available.	
472		Comment: When using a composite endpoint all components should be analysed separately, but it is not stated how this should be done, e.g. is it allowed to state unadjusted confidence intervals for the different components. Especially with time-to-event endpoints there is the problem of competing risks which has to be taken into account.	
478-489		Comment: Only two types of composite endpoints are mentioned, namely rating scales and time to first event, but also binary events can form composite endpoints where either any effect versus none or the number of positive component events are compared. Instead of combining information on a patient level by defining a summary score, multivariate methods such as global testing procedures and GEE approaches can provide distinct advantages for finding an overall treatment effect. In addition non-parametric comparison methods based on multivariate orderings have been proposed.	
520-524		Comment: What are the particular issues of composite endpoints in the	

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)*
		non-inferiority or equivalence setting? The listed problems like a decrease in power due to insensitive components or increased variance also apply for superiority trials.	
558-560		Comment: In case of a successful primary analysis of a composite endpoint, it is often a problem that claims do not properly reflect that a composite was used. An example of what kind of wording should be used would be helpful.	
577		Comment: "lead to an overestimation of the corresponding treatment effect." At this point, it might be crucial to mention, that we especially suffer also a bias in the corresponding p-values via selection. It is important, because in some situations, it was proposed that a bias of the estimate of the corresponding treatment effect can be avoided by publishing all calculated estimates. However, this would not solve the intrinsic problem of the selection bias. Proposed change (if any): "At this point, it might be crucial to mention, that we especially suffer also a bias in the corresponding p-values via selection. Reporting all treatment effects does not solve the problem of multiplicity."	
570-572 *596-599*		Comment: In case confidence regions corresponding to multiplicity	

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)*
		procedures are not available, it is suggested that conservative confidence interval methods such as Bonferroni-corrected intervals are used. At the same time it is said that "Confidence regions, e.g. that are uninformative in the sense that they never exclude the null hypothesis of no treatment effect in order to comply with the multiple testing procedure, would have no relevance in the assessment of the trial results."	
599		Comment: A conclusion section seems to be missing.	

Please add more rows if needed.