29 June 2017

# Submission of comments on 'Guideline on multiplicity issues in clinical trials' (EMA/CHMP/44762/2017)

## Comments from:

| Name of organisation or individual |
| --- |
| International Society for Clinical Biostatistics:<br>Statistics in Regulatory Affairs Subcommittee |

*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).*

7 Westferry Circus ● Canary Wharf ● London E14 4HB ● United Kingdom
**Telephone** +44 (0)20 7418 8400 **Facsimile** +44 (0)20 7418 8416
**E-mail** info@ema.europa.eu **Website** www.ema.europa.eu

# 1. General comments

| Stakeholder number<br>*(To be completed by the Agency)* | General comment (if any) | Outcome (if applicable)<br>*(To be completed by the Agency)* |
|---|---|---|
| | First of all, we thank you for this guideline on multiplicity issues in clinical trials and the opportunity to give our comments. The document is very helpful and addresses the different situations when it is necessary or not to adjust of multiplicity with short summary at the top of each section.<br>However, this guideline is worded in quite general terms. It is only understandable for persons with sufficient knowledge of simultaneous statistical inference. | |
| | There are no formal recommendations on 'how' to adjust. Some ICH and CPMP guidance documents are reported but there are no specific references for each point. For example in section 5.5.3 "There are various methods published in the relevant literature on test procedures with relevance to these studies that can be adapted to the specific aims and that provide the necessary control of the type I error". It is recommended either to explain in more detail the applied technical terms or to give references where the required explanations can be found for each section. | |
| | The issue of interim analyses and early stopping which also causes multiplicity problems is not addressed. It is proposed to add corresponding sections in chapter 5 and | |

| Stakeholder number<br>*(To be completed by the Agency)* | General comment (if any) | Outcome (if applicable)<br>*(To be completed by the Agency)* |
|---|---|---|
| | chapter 10. | |
| | Genomic data in clinical trials are increasingly used. Adjustment in this context of large volume data is not addressed. | |
| | An overview on sample size calculation in multiple testing in different situations (hierarchical procedure, Tukey's range test, Dunnett's test, Sidak correction, Holm-Bonferroni method, closed testing procedure, gatekeeping procedure…) would be a useful complement of the guideline. | |
| | A summary table or decision tree on multiplicity issues would be helpful for the reader. | |
| | A review of available adjustment methods used in their respective context of multiplicity, associated with recommendations for their use (pros and cons for instance) would be very helpful. | |
| | In longitudinal studies with multiple groups (as example 5 follow-up visits and 3 treatments), clinicians are frequently interested by two kinds of hypotheses: 1) test of a group effect for all visits or at specific ones (i.e. between-group comparisons), 2) test of a visit effect for each group (within-group comparisons). Each global test may also involve pairwise comparisons. The guideline does not discuss these type double sources of adjustment (within- and between-group comparisons). | |
| | We do not find discussion on consequences of multiplicity adjustment in the context of independent tests vs. | |

| Stakeholder number<br>*(To be completed by the Agency)* | General comment (if any) | Outcome (if applicable)<br>*(To be completed by the Agency)* |
|---|---|---|
| | correlated tests. Many authors indicate that tests are generally positively correlated which decrease the type I error rate. Could you add some discussions/explanations about this? | |
| | Gatekeeping procedure seems to be very interesting. Any advices on this procedure would be very appreciated. | |
| | Some assessment of multiplicity issues arising in non-inferiority and especially equivalence studies would be much appreciated. Such studies rely heavily on confidence intervals and adjustments might be warranted. For example, approval of a new biosimilar treatment requires a clinical trial with claim of equivalence compared with the originator. Equivalence is claimed if the confidence interval of the treatment difference fully lies within the equivalence margins. Multiplicity issues might arise in such trials, and it is uncertain if the confidence intervals should be adjusted in case of e.g. a sequential testing regime, and how this adjustment should be performed. A simple Bonferroni-correction might be too conservative. | |

## 2. Specific comments on text

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| 64-66 | | Comment:<br>The guideline mentions "alternative statistical methods" but Bayesian methods are not considered in the text. It is a pity that in 2017 such interesting alternative methods to manage the multiplicity issues are not talked about.<br><br>Proposed change (if any):<br>To add at least one paragraph or better a chapter devoted to Bayesian methods and how they solve the multiplicity issues. | |
| 119 | | Comment:<br>Following the preceding comment, this should possibly be modified if Bayesian methods are added to the guideline.<br><br>Proposed change (if any): | |
| 142-146 | | Comment:<br>In case of evaluation of the primary efficacy variable at repeated visits per patient, the guideline indicates that interpretations "usually do not cause multiplicity problems, because in the majority of situations either an appropriate summary measure has been pre-specified or …". We do not understand the use of a summary measure in the majority of situations. When repeated measures are replaced by a single- | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | number summary, there is a loss of information, including dynamic evolution of criteria.<br><br>Proposed change (if any):<br>If possible, the use of a repeated measure model with multiplicity adjustment seems to be more appropriate than a single-number summary. Adjustment in this context would be a useful complement. | |
| 161 | | Comment:<br>The issue of interim analyses is missing.<br><br>Proposed change (if any):<br>Add a section on interim analyses in chapter 5. | |
| 205-207 | | Comment:<br>It is mentioned that other methods are possible but they are neither described, nor even listed<br>In particular, any references regarding other methods of dealing with multiple variables that are more complex and that can be found in the literature would be helpful<br><br>Proposed change (if any):<br>Give at list of these methods or a short bibliography, and at best a dedicated paragraph. | |
| 210-214 | | Comment: | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | We are not aware of good examples in which one would drop a new drug just because one of, say, five endpoints failed to come out significant (without actually coming out unfavourable). A perfectly natural primary outcome that uses a four-out-five criterion of this kind, however, is not covered by the present text, and it would require special statistical guidance to implement without multiplicity problems. | |
| 222-224 | | Comment:<br>Sample size in case of multiple testing is not addressed.<br><br>Proposed change (if any): -<br>Add a discussion or guideline references on how to deal with sample size calculation in case of multiple testing? | |
| 235 | | Comment:<br>Re 'the particular interests of the investigator': The investigator's special interest in an aspect of treatment outcome (e.g., the behaviour of a particular plasma marker) is not our concern. What is meant is (I believe):<br><br>Proposed change:<br>…result from the particular (e.g., geographical or socio-administrative) context of the drug trial. | |
| 256-276 | | Comment:<br>It is correct that a two-step procedure in which at first a | |

| Line number(s) of the relevant text<br><br>*(e.g. Lines 20-23)* | Stakeholder number<br><br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br><br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br><br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | statistical test is used to select the statistical method for the hypothesis test of main interest in the second step does not usually control the type I error and is in general not recommended. However, this situation should be distinguished from the use of statistical methods which are applied to justify the use of a pre-specified statistical method for the hypothesis test of main interest. Examples are given by the investigation of the proportional hazards assumption to justify the use of the Cox model, the investigation of the proportional odds assumption to justify the use of the proportional odds model, the investigation of the linearity assumption to justify the use of a standard logistic regression model with continuous predictors, the investigation of heterogeneity to justify the pooling of study results by means of a meta-analysis or the investigation of heterogeneity to justify a common analysis over certain subgroups. For the investigation of required model assumptions frequently statistical hypothesis tests as well as other statistical approaches, such as graphical tools are available. The result of such investigations of model assumptions can be that the pre-specified method appears to be invalid and should be changed or that no hints are found that the assumptions are strongly violated and the use of the pre-specified method is supported. In the latter case it should not be argued that an adjustment of the p-value for the main hypothesis test is required because in the first step a statistical test has been used (probably in combination with graphical tools) to investigate the required model assumption. | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | Proposed change (if any):<br>Provide a clear distinction between a two-step procedure for model selection (where control of the type I error for the whole procedure is required) and the use of statistical and graphical methods to investigate the model assumptions of a pre-specified method for the hypothesis test of main interest. | |
| 257-258 | | Comment:<br>We do not understand why opposing the Wilcoxon test versus the log-rank test in your examples of different statistical models or statistical techniques "(e.g. parametric vs. non parametric or Wilcoxon versus log-rank test)". Do you mean the generalized Wilcoxon test with different weighs that can deal with censorship?<br><br>Proposed change (if any):<br>Limit the example to "(e.g. parametric vs. non parametric)" or maybe give less confusing example "(e.g. parametric vs. non parametric or Wilcoxon versus Student test)". | |
| 262 | | Comment:<br>Knowledge of assignments is not necessary but it worsens the problem.<br><br>Proposed change:<br>…strategy, not least when based on… | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | | |
| 277-281 | | Comment:<br>It is also correct that "Confirmatory analyses should be fully and precisely pre-defined to exclude the possibility of performing different analyses *post hoc*." However, in line with other guidelines (e.g. the subgroup guideline) a pre-defined analysis (e.g. a common analysis among certain subgroups) should not be performed if the data – against expectation - show clear violations of required model assumptions (e.g. substantial effect modification, which require the estimation of the treatment effect within subgroups).<br><br>Proposed change (if any):<br>It should clearly be stated that precisely pre-defined confirmatory analyses should be changed if the data – against expectation - show clear violations of required model assumptions. | |
| 287-292 | | Comment:<br>The wording gives the peculiar feeling that the limited value of p-values is limited to the adverse effects while the same rhetoric could be applied to the main outcome, in most trials. This is the most surprising if the main outcome is exactly a safety outcome !<br><br>Proposed change (if any):<br>Add a discussion on the logic of the frequentist statistical test | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | and contrast this logic between a main outcome (on efficacy most of the times) and an adverse event. | |
| 293 – 298 | | Comment:<br>This paragraph is a clear illustration of the ambiguousness of the statistical test. Another solution would be to increase the type I error rate and nevertheless perform and adjustment for multiplicity. Another solution would be here to use a Bayesian method, (for instance by computing the probability of a given adverse event) all the more since the line 297 refers to prior knowledge.<br><br>Proposed change (if any): | |
| 313 | | Comment:<br>Re '[reference] product': A user (not producer) viewpoint is required here (and the reference regimen may not even be 'product' at all). This blemish does not occur elsewhere in the text.<br><br>Proposed change:<br>…reference regimen… OR …reference treatment… | |
| 338-343 | | Comment:<br>This part is somewhat fuzzy and unclear.<br><br>Proposed change (if any): | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| 343 | | Proposed change:<br>…is normally not required… | |
| 373 and 378 | | Comment:<br>The sentence states that "statistical tests are of descriptive nature": is it not an oxymoron ? The purpose of a test is to make an inference on the hypothesis tested and not to describe the date, even on secondary outcome, though I glimpse what is meant by descriptive. (lack of power on unspecified assumptions etc).<br><br>Proposed change (if any):<br>Maybe a simple rewording may make things clearer :<br>"CI and statistical tests are, in the specific situation of secondary outcomes, mainly to be understood as exploratory and suggestive of an evidence"<br>Or something like this. | |
| 518 | | Comment:<br>True – unless the beneficial effect can be shown to be practically weight-independent. | |
| 531-532 | | Comment:<br>As appropriate, the guideline indicates that "It is not generally appropriate to handle patients who die before they reach the hospital as censored. It is better practice to study a composite endpoint that includes all important clinical events | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | as components, including death in this example". Another solution could also to use competing risks analysis.<br><br>Proposed change (if any):<br>"It is better practice to study a composite endpoint that includes all important clinical events as components, including death in this example, or to use competing risks analysis." | |
| 561 | | Comment:<br>The issue of interim analyses and early stopping is missing.<br><br>Proposed change (if any):<br>Add a section regarding the bias of effect estimates in trials that stopped early due to very large observed treatment effects. | |
| 562-572<br>and<br>589-599 | | Comment:<br>The statement that informative confidence regions that correspond to multiplicity procedures are frequently not available is given twice (in the beginning of section 10 as well as in section 10.2).<br><br>Proposed change (if any):<br>The paragraph in the beginning of section 10 should be moved to section 10.2. | |
| 579 | | Comment: | |

| Line number(s) of the relevant text<br>*(e.g. Lines 20-23)* | Stakeholder number<br>*(To be completed by the Agency)* | Comment and rationale; proposed changes<br>*(If changes to the wording are suggested, they should be highlighted using 'track changes')* | Outcome<br>*(To be completed by the Agency)* |
|---|---|---|---|
| | | 'may' does not here mean 'is allowed'<br><br>Proposed change (if any):<br>…may still be problematic as it is possibly based… | |
| 579 | | Comment:<br>'correlated' – in this particular dataset (is meant) – rather than in the "true" distribution<br><br>Proposed change (if any):<br>…happens to be correlated with efficacy in the data at hand. | |
| 586 | | Comment:<br>The notion of shrinkage is a typical Bayesian notion. It is an additional argument to talk about Bayesian methods.<br><br>Proposed change (if any):<br>Add at least a small chapter on Bayesian methods and on how they solve the multiplicity issue. | |
| 593-594 | | Comment:<br>The sentence indicates that confidence regions/intervals that correspond to multiplicity procedures are not always available. Could you provide information or references on which methods allow to get the confidence interval and how?<br><br>Proposed change (if any): - | |

Please add more rows if needed.