30 June 2017

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

Comments from:

| Name of organisation or individual |
| --- |
| EFPIA – Sandra Rodrigues (sandra.rodrigues@efpia.eu) |

*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 | Outcome (if applicable)  *(To be completed by the Agency)* |
| --- | --- | --- |
|  | EFPIA welcomes the publication of the EMA Draft “Guideline on multiplicity issues in clinical trials” and is pleased to offer the following comments and recommendations. |  |
|  | **Comment:**  There seems to be no formal definition of the terms “primary”, “key secondary”, “secondary” and “exploratory/supportive” endpoints. It is described that secondary endpoints require multiplicity control if they aim to provide confirmatory conclusions, but not if they only provide supportive evidence. There seem to be different viewpoints on the use of these terminologies. For example, the FDA Draft Guidance for Industry on “Multiple Endpoints in Clinical Trials” seems to suggest that all secondary endpoints require multiplicity adjustments. Clarification and international harmonization would be helpful for consistent applications in practice.  See also the specific comments on Lines 383-389 and 390-394, among others. |  |
|  | **Comment:**  The document covers the key principles of multiplicity but only provides limited details and/or examples of methodologies that could be used to control for multiplicity. Including methodologies and/or examples would help to illuminate the key principles. These methodologies and/or examples should not be construed as a regulatory recommendation and should not be introduced at any technical level in order to leave room for further methodological developments.  Possible parts of the document that could benefit are:   * Lines 180-183: provide examples of different methods (e.g. Holm, Hochberg, etc.) * Section 5.1.2: examples of defining hierarchies * Lines 406-408: provide examples of situations in which hierarchical methods for primary and secondary endpoints may not be appropriate   See also the specific comments below on Lines 113-146, 174-175 and 406-409, among others. |  |
|  | **Comment:**  We note that there is no mentioning on how to address multiplicity in terms of estimands (if more than one estimand has been defined). We would welcome the inclusion of the estimand concept in the guideline as it could potentially simplify and clarify some of the sections of the document (e.g. those on Analysis Sets or Composite Endpoints). As an example, in Section 5.2 performing robustness assessments on different subsets of patients is discussed and advised. In the anticipated language of the estimand amendment to the ICH E9 guideline, this amounts to testing different estimands, rather than robustness testing of the same estimand. The ICH E9 working group members advertised it as one of the advantages of the estimand amendment to ICH E9 that the introduction of the estimand concept would focus and highly likely reduce the need for wider sets of (sensitivity) analyses with unclear contribution to robustify the primary analysis. |  |
|  | **Comment:**  Throughout the draft guideline, the terms “variable/endpoint”, “trial/study”, and “subject/patient” seem to be used in an exchangeable way. It might be helpful to agree on one term in each case and use it consistently. |  |
|  | **Comment:**  The agency may consider addressing additional topics related to multiplicity issues in clinical trials that are currently not covered in the draft guidance. Possible topics include:   1. Endpoint analysis based on data pooled from two pivotal trials. 2. Different regional requirements on the choice of the primary endpoint in global clinical trials. For example, will an “agency-wise” error rate control be sufficient in EMA’s perspective? 3. Multiplicity considerations in combined non-inferiority / superiority trials, possibly for more than one endpoint or dose |  |

1. 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 70-78 |  | **Comment:**  In addition to answering more than one question in a single clinical trial, multiplicity issues may also arise when multiple doses and/or treatments are being studied in a single study or a study may have multiple looks at the data (i.e. interim analyses). While these subjects are covered in the document there is no mention of them in the executive summary.  **Proposed change:**  Please add text to this effect. Possible text could be:  “*Similarly, statistical methods may be needed to deal with, or to avoid, multiplicity issues arising from studies with multiple doses and/or treatment arms or studies with one or more interim analyses.*” |  |
| Lines 77, 170 |  | **Comment:**  There is inconsistency regarding the place where methods for multiplicity adjustments should be detailed. Line 77 states “… in the study protocol or in the statistical analysis plan ….” Line 170 states “… in the protocol and analysis plan ….” |  |
| Line 83 |  | **Comment:**  Section 10 seems to imply that simple conservative confidence intervals should be provided for treatment effect estimates when complex multiplicity adjustment methods are used for which no satisfactory equivalent confidence regions are available. In contrast, this section appears to suggest that confidence intervals and acceptance regions of multiple tests have to match.  See also the specific comment on Lines 593-599.  **Proposed change:**  Please add appropriate clarification to this effect to state whether either of these approaches is acceptable. |  |
| Lines 97-100 |  | **Comment:**  In obtaining the value 12%, it would be helpful to explicitly state that it is assumed there is no true effect in any subgroup. In addition, it would be clearer to change the wording of “at least one false positive statistically significant test” to “at least one false positive conclusion with statistical significance”.  **Proposed change:**  “*For example, if statistical tests are performed on five subgroups when there is no true effect in any subgroup, independently of each other and each at a significance level of 2.5% (one-sided directional hypotheses), the chance of finding at least one false positive conclusion with statistical significance increases to approximately 12%.*” |  |
| Line 105 |  | **Comment:**  The term “study-wise type I error” is not commonly used compared to the term “familywise type I error”. A formal definition of the term “study-wise type I error” is missing despite its importance for this document. Without a formal definition, confusion could arise as to whether “study-wise type I error” is exchangeable with “familywise type I error” or whether all hypotheses in a study have to be included in the multiplicity adjustment.  **Proposed change:**  Please add appropriate clarification to this effect |  |
| Lines 113-146 |  | **Comment:**  There are Bayesian methods that can be used to address multiplicity issues.  **Proposed change:**  Please add text to this effect. |  |
| Lines 116-118 |  | **Comment:**  For safety evaluations, false positive conclusions are also important and may harm the benefit-risk profile of the drug and further impact the possibility of the approvability of the drug. It thus seems too strong a statement that (rigorous) multiplicity adjustment is counterproductive in flagging safety signals and thus inappropriate for safety evaluations. While study-wise error rate control could be too strict, in practice we often use multiplicity adjustments to control the false discovery rate for safety evaluations. It would be helpful to allow appropriate forms of multiplicity adjustments for safety evaluation.  **Proposed change:**  “*Following the precaution principle in safety evaluations, it is usually more important to reduce the rate of false negative conclusions on harm than to control the number of false positive conclusions. Hence, strictly controlling the study-wise error rate of false positive conclusions can mask safety risks and will often be inappropriate. However, appropriately controlling the false discovery rate for safety evaluations could balance false negative and false positive conclusions on harm.*”  See also the specific comment on Lines 293-295. |  |
| Lines 122-124 |  | **Comment:**  Confusing wording, suggest deleting part of the sentence.  **Proposed change:**  “*The CHMP Points to Consider on Application with 1. Meta-analyses and 2. One Pivotal Study (CPMP/2330/99) covers the situation when the type I error needs to be controlled at the submission level for these cases. ~~where more than one confirmatory trial is included in a submission~~.*” |  |
| Lines 142-146 |  | **Comment:**  There are situations in long-term trials of well-established indications in chronic diseases where more than one time point may be of interest: in addition to a pre-defined time point for short-term efficacy (minimum requirement for approval), long-term efficacy or speed of onset may also be important for confirmatory conclusions (as opposed to capturing repeated visits per patient through e.g. a summary measure.  **Proposed change:**  We propose to add the following sentence to the end of Line 146:  “*In situations where analyses at multiple time points are needed for confirmatory conclusions (e.g., short-term efficacy and long-term efficacy), regulatory dialogue is recommended to decide on the need for a multiplicity adjustment.*” |  |
| Lines 173-179 |  | **Comment:**  The term “multi-level-alpha tests” is used only here, whereas the term “multiple testing procedures” is used elsewhere in the document. We propose to use the more commonly accepted term “multiple testing procedures”. Furthermore, the paragraph would benefit from being a bit more specific. We propose to replace the paragraph by the following.  **Proposed change:**  “*Controlling the type I error rate study-wise is frequently done by testing the individual hypotheses at adjusted local significance levels that are fractions of the pre-specified overall significance level α. A valid and simple procedure is the Bonferroni test. More powerful multiple testing procedures are available if the correlations between the test statistics are taken into account (e.g. the Dunnett’s test for multiple comparisons of several treatments to a single common control). Further improvements are possible by stepwise procedures that allow propagating the levels of rejected hypotheses to not yet rejected ones (e.g. Bonferroni-Holm or graphical procedures).*” |  |
| Lines 174-175 |  | **Comment:**  The word “splitting” might suggest that here Bonferroni method is meant. This does not seem adequate since there are generally more powerful methods, such as Bonferroni Holm. To avoid confusion, a reference to an appropriate method, such as Bonferroni Holm should be inserted.  **Proposed change:**  Insert a reference to an appropriate method such as Bonferroni Holm (Holm, S. A Simple Sequentially Rejective Multiple Test Procedure Scandinavian Journal of Statistics, 1979, 6, 65-70) |  |
| Line 184 |  | **Comment:**  The title of Section 5.1 may be misunderstood as suggesting that there is no need for adjustment for multiple primary endpoints. In particular, the hierarchical procedure described within this section is generally considered to be a type of formal adjustment. We propose to shorten the title and to clarify the meaning of “formal adjustment” in the text.  **Proposed change:**  “*5.1. Multiple primary endpoints ~~– when no formal adjustment of the significance level is needed~~*” |  |
| Lines 184-207 |  | **Comment:**  It should be acknowledged that different primary endpoints are likely to require different sample sizes to achieve adequate power to detect the treatment effect of interests. Thus, for one endpoint the study may be grossly over-powered so that for the endpoint in question there may be high probability of statistical significance even for a clinically irrelevant difference. This is not to suggest downsizing current trials, but to emphasize the need for appropriate sample size calculations and understanding operating characteristics of a given trial design through tailored clinical scenario evaluations.  **Proposed change:**  Please add text to this effect. |  |
| Lines 196-200 |  | **Comment:**  The guideline proposes to reserve the term “co-primary” for the first scenario. We appreciate that this guideline clearly defines the scope of the term “co-primary” as in practice it is often used in a broader sense, thus leading to confusion.  It would be helpful if this guideline could in addition also introduce clear terminology for the second scenario, to distinguish the two different scenarios.  **Proposed change:**  Please add text to this effect. |  |
| Lines 196-202 |  | **Comment:**  Can the principles discussed in this paragraph be generalised to situations in which there are multiple doses, treatments or populations?  **Proposed change:**  Please add appropriate text to this effect to state whether it is or is not. |  |
| Lines 200-202 |  | **Comment:**  The sentence is not clear and the case of two primary endpoints should also be included.  **Proposed change:**  “*More generally, in case of ~~more than~~ two or more primary endpoints, adjustment is needed if not all endpoints need to be significant to define study success. ~~and the inability to exclude deteriorations in other primary endpoints would have to be considered in the overall benefit/risk assessment.~~ Moreover, declaring study success may depend on additional requirements regarding the absence of harm in other primary endpoints.*” |  |
| Line 205 |  | **Comment:**  It would be helpful to clearly differentiate between “significance level”, “type I error” and “type I error rate” throughout the document. In Line 205, for example, the statement “*adjustment of type I error rates*” needs to be corrected – one would adjust significance levels or p-values, but not error rates. |  |
| Lines 212-214 |  | **Comment:**  Interpretation is not necessarily clear-cut as statistical significance does not mean clinical relevance and the two endpoints could have very different powers. Trials with multiple endpoints are likely to have been powered on the endpoint requiring the larger sample size (see earlier comment on Lines 184-207).  **Proposed change:**  Please add text to this effect |  |
| Line 214 |  | **Comment:**  Lines 98-99 use a one-sided significance level of 2.5%. We suggest being consistent throughout the document.  **Proposed change:**  … same significance level (e.g. ~~0.05~~2.5% one-sided) |  |
| Line 214-218 |  | **Comment:** The selection between either a symptom-based endpoint or a patient-related endpoint as the other co-primary endpoint with lung function may add another layer of multiplicity. Please clarify if it is acceptable that the selection is data driven with proper multiplicity adjustment (e.g. Bonferroni adjustment between the symptom-based and patient-related endpoint).  In addition, symptom-based and patient-related endpoints are both patient reported endpoints. It’s not clear why they are indicated differently. |  |
| Lines 217-218 |  | **Comment:**  All endpoints are patient-related, the phrase below is how it should relate to a patient reported endpoint.  **Proposed change:**  “*… a patient ~~related~~ reported endpoint.*” |  |
| Line 225 |  | **Comment:**  This subsection appears to be somewhat confusing as the title of this subsection states that the rank is according to clinical relevance but Lines 233-235 provide other rationales that can legitimately be used for ranking. It would be helpful to remove “according to clinical relevance” in the title of the subsection.  **Proposed change:**  “5.1.2. Two or more endpoints ~~ranked according to clinical relevance~~ *ordered hierarchically*” |  |
| Lines 226, 241 |  | **Comment:**  In a hierarchical test, there are no two endpoints with equal rank. Otherwise, numerical adjustment of each hypothesis may be needed. Thus the expression “or equal to” could be misleading.  **Proposed change:**  “*However, no confirmatory claims can be based on endpoints that have a rank lower than that ~~or equal to~~ variable whose null hypothesis was the first that could not be rejected.*” (Line 226)  “*The effect of such a procedure is that no confirmatory claims can be based on endpoints that have a rank lower than ~~or equal to~~ that variable whose null hypothesis was the first that could not be rejected.*” (Line 241) |  |
| Lines 232-235 |  | **Comment:**  Often hierarchy is also based on probability of success of demonstrating the effect. That is, an endpoint may be lower in the hierarchy due to sample size required to demonstrate effect rather than clinical importance (i.e. MACE vs mortality).  **Proposed change:**  Can the agency please elaborate on this point and perhaps provide an example to illuminate.  Can the agency also elaborate in the setting in which clinical significance for endpoint higher in the hierarchy is not met, but is met for endpoint which has greater clinical relevance but is lower in the hierarchy? Example: study misses on 6MWT (primary) in heart failure, but hits on mortality (secondary). This is also relevant for Section 6.3.; the discussions at both places should be aligned. |  |
| Line 234 |  | **Comment:**  It would be helpful to clarify the meaning of “*hypotheses are ordered in time*”. |  |
| Lines 250-254 |  | **Comment:**  It states that no adjustment is necessary if subsets of subjects are analysed. This is in contrast to Section 7, but similar to the cited CHMP Guideline on the Investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013), which stated “It might be questioned whether the multiplicity associated with subgroup analyses and interaction tests should be addressed through changes to nominal significance levels for tests or presentation of confidence intervals. However, since these investigations serve as an indicator for further exploration, adjustment would be counter-intuitive and is not recommended”  **Proposed change:**  “*In general, multiple additional analyses on varying subsets of subjects or with varying measurements for the purpose of investigating the robustness of the conclusions drawn from the primary analysis should not be subjected to adjustment for multiplicity (see also CHMP Guideline on the Investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013)). This is in contrast, however, to the confirmatory subgroup analyses described in Section 7.*” |  |
| Lines 273-276 |  | **Comment:**  The sentences in these lines seem to be unclear. For example, “*It is also important in this case that there is no inflation in the type I error*” might be misunderstood that there per se is no type I error inflation. If there is no inflation in the type I error in blinded data analysis, then there is no justification for the requirement for type I error control for statistical model selection based on blinded data analysis.  **Proposed changes:**  Can the agency please clarify these sentences to this effect or delete the paragraph from Line 271 to Line 276 to eliminate confusion. |  |
| Line 274 |  | **Comment:**  Please give examples of what information a sponsor would need to provide to demonstrate that a blinded review of data did not lead to type 1 error inflation.  **Proposed change:**  Can the agency please add an example to this effect. |  |
| Line 276 |  | **Comment:**  An example would help to better understand when a bias could occur.  **Proposed change:**  Can the agency please add an example to this effect. |  |
| Lines 277-281 |  | **Comment:**  While ideally the principle highlighted in this paragraph should be followed, there may be situations where this is not possible. An example is in the area of rare diseases, where there may be limited to no historical data to guide the choice of endpoints and/or statistical methods. In these cases, quite often the sponsors are forced to choose one from many equally important candidate endpoints as primary based on very limited clinical experience. Often once the trial is completed, the data may indicate that other endpoints can inform efficacy or alternative statistical methods are more appropriate.  **Proposed change:**  Can the agency elaborate on this situation. |  |
| 283-286 |  | **Comment**:  Benefit-risk assessment is often conducted for a regulatory submission of a new therapy. Safety variables are assessed simultaneously with efficacy variables. Both safety and efficacy results are typically considered for a formal regulatory approval based on benefit-risk assessment. Safety and efficacy assessments are two different attributes of a therapy for a disease indication. If a safety endpoint must be met for the regulatory success, should it be considered as co-primary endpoint with the primary efficacy endpoint, i.e., no multiplicity adjustment is needed.  **Proposed change**:  *“When a safety variable is part of the confirmatory strategy of a study and thus has a role in the approval or labelling claims, it should not be treated differently from the primary efficacy endpoints. For example, it can be treated as a co-primary endpoint with the primary efficacy endpoint, i.e., no multiplicity adjustment is needed. The exception for the situation ~~that~~ is when the observed effects go in the opposite direction and may raise a safety concern (see also Section 9.3).”* |  |
| Lines 293-295 |  | **Comment:**  For safety evaluations, false positive conclusions are also important and may harm the benefit-risk profile of the drug and further impact the possibility of the approvability of the drug. It thus seems too strong a statement that (rigorous) multiplicity adjustment is counterproductive in flagging safety signals and thus inappropriate for safety evaluations. While study-wise error rate control could be too strict, in practice we often use multiplicity adjustments to control the false discovery rate for safety evaluations. It would be helpful to allow appropriate forms of multiplicity adjustments for safety evaluation.  **Proposed change:**  “*In those cases where a large number of statistical test procedures are performed to serve as a flagging device to signal a potential risk caused by the investigational drug it can generally be stated that ~~an adjustment for multiplicity is~~ rigorous multiplicity adjustments controlling the study-wise type I error rate are counterproductive for considerations of safety. Instead, multiplicity adjustments to control the false discovery rate may be used for safety evaluations.*”  See also the specific comment on Lines 116-118. |  |
| Line 329 |  | **Comment:**  The justification “because there is obviously no alternative” is not clear. Formally speaking, this is the intersection-union test and thus adjustments to the significance level are not needed.  **Proposed change:**  “*In case the intended contribution of the fixed combination is to improve efficacy, such a study is considered successful if the combination is shown superior to both components; no formal adjustment of the significance level for the single hypothesis tests is necessary. ~~because there is obviously no alternative.~~*” |  |
| Lines 331-336 |  | **Comment:**  In multiple-dose factorial designs, the source of multiplicity is largely due to the multiple doses, which is addressed in the following section. Thus it would be better if this paragraph were moved to Section 5.5.3 as a case which may require multiplicity adjustments.  **Proposed change:**  We propose to move this paragraph to Section 5.5.3. |  |
| Line 336 |  | **Comment:**  It would be helpful to clarify the meaning of “*global test strategies*”. |  |
| Lines 342-343 |  | **Comment:**  The consequence of “Therefore, the multiplicity adjustment of the different comparisons between groups in order to control the study-wise type I error may not be required in a Phase II trial” may not match the intent. Rather, one should discourage the use of pairwise comparisons in dose-response studies (which is not what the highlighted sentence currently says).  **Proposed change:**  “*~~Therefore, the multiplicity adjustment of the different comparisons between groups in order to control the study-wise type I error may not be required in a Phase II trial.~~*” |  |
| Line 344 |  | **Comment:**  It would be more accurate to use “association” than "correlation". Also, this statement assumes a monotonic drug-response relationship. In other cases (such as for umbrella curves) this statement may not hold.  **Proposed change:**  “*Assuming a monotonic drug-response relationship, a valuable achievement in such a trial is the demonstration of an overall positive ~~correlation~~ association of the clinical effect with increasing dose …*” |  |
| Lines 345-347 |  | **Comment:**  The reference to “regression models” earlier in this paragraph (line 339) may not be clear to the casual reader. Also, multiplicity considerations with respect to Type I error rate control and selection bias may still be applicable if multiple regression models are used, which should be acknowledged in this guideline on multiplicity issues in clinical trials.  **Proposed change:**  Expand  “*Estimates and confidence intervals of the relevant parameters in the regression models are used for an appropriate interpretation of the dose response and may be used for the planning of future studies.*”  to  “*Estimates and confidence intervals for the relevant parameters of suitable dose response regression models are used for an appropriate interpretation of the dose response and may be used for the planning of future studies. Moreover, if multiple regression models are used, multiplicity considerations with respect to type I error rate control and selection bias may be applicable (see EMA/CHMP/SAWP/757052/2013).*”  Reference: EMA/CHMP/SAWP/757052/2013 - Qualification Opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty |  |
| Lines 347-349 |  | **Comment:**  It would be helpful to give more detailed reference to ICH E4 regarding the meaning of “*the confirmatory package*” and “*the type I error*” in dose-response studies. |  |
| Lines 365-371 |  | **Comment:**  The word “claim” is not used consistently throughout the document and confusing. For example, the description in Lines 369-371 indicates “claim” is any confirmatory conclusion in study report, or clinical overview etc. while the descriptions in Lines 373 and 376 indicate that “claim” means “labelling claim”.  See also the specific comments on Lines 379-382 and 395-409.  **Proposed change:**  Please ensure the consistency of the use of the word “claim” throughout the document |  |
| Line 369 |  | **Comment:**  The term “*separate regulatory guidance document*” will not be informative to many readers; a specific reference should be included.  **Proposed change:**  Please provide a reference to guidance document |  |
| Lines 379-382 |  | **Comment:**  From the wording in this section it seems to be a recommendation that a multiple testing procedure is applied to secondary endpoints irrespective of whether they are for “claim” or for inclusion in the SMPC.  **Proposed change:**  Can the agency please clarify whether this is the intended guidance or is not. |  |
| Lines 383-389 |  | **Comment:**  The paragraph discusses the ranking of endpoints in a hierarchical test, which is a form of multiplicity adjustment. But this subsection mainly discusses secondary endpoints that do not provide confirmatory conclusions and thus don’t require multiplicity adjustments. Given the dedicated discussion on hierarchical tests in Section 5.1.2, it would be helpful to merge this paragraph into Section 5.1.2.  **Proposed change:**  We proposed to merge this paragraph with Section 5.1.2. |  |
| Lines 390-394 |  | **Comment:**  In current practice, key secondary endpoints are often included in the multiple testing procedures to seek confirmatory label claims. In this paragraph, however, it seems to be suggested that key secondary endpoints can be specified without multiplicity adjustment, yet positive results from them may be considered as reliable for inference or support information in the Summary of Product Characteristics. It would be helpful to clarify the definition of key secondary endpoints, their regulatory role, and their relationship with the primary and secondary endpoints that require multiplicity adjustments. |  |
| Lines 395-409 |  | **Comment:**  Please provide clarification in Section 6.2 that a hierarchical testing procedure for such secondary endpoints will support that results are sufficiently reliable for inclusion in the Summary of Product Characteristics, in order to be consistent with guidance in Section 6.1. |  |
| Lines 406-409 |  | **Comment:**  While the hierarchical test may be suitable when the order of hypotheses is easy to establish, it may be too strict for other situations, where more flexible procedures have clear advantages (e.g., Bonferroni-Holm, gatekeeping and graphical procedures). Thus it would be helpful to mention some well-established procedures which are also appropriate for multiplicity adjustments for primary and secondary endpoints.  **Proposed change:**   1. Delete the sentences in Lines 406-409:   “*~~However, more complex methods exist to control type I error over both primary and secondary endpoints, and these could be more useful in some circumstances. Depending on the degree of complexity, regulatory dialogue is recommended to assure that the outcome of the procedure can be interpreted in clinical terms.~~*”   1. Insert a new paragraph as follows after “*…, reflects their relative importance in the study.*”:   “*There are situations in practice where the hierarchical order is less clear among the secondary endpoints. In these cases, a hierarchical test may be too strict. Thus more flexible methods that do not impose a strict hierarchy may be preferred. These procedures are well established and usually have power advantages to reject at least one secondary hypothesis.*” |  |
| Lines 413-418 |  | **Comment:**  Please consider providing guidance and examples on assigning the clinically very important endpoint also as primary endpoint and adjust for multiplicity with the consideration of consistency between the endpoints, such that this clinically very important endpoint alone might be sufficient for a claim. |  |
| Line 423 |  | **Comment:**  To improve the flow of the presented material, Section 7 could be divided into two subsections to improve clarity: (1) Subgroup analyses as part of sensitivity analyses, (2) Subgroup analyses as part of the confirmatory strategy of a study.  In case (1), subgroup analyses are performed to explore whether the observed overall treatment effect is homogenous or have a similar trend among the subgroups. No multiplicity adjustment is needed. This is in line with Section 5.2 and the cited guideline EMA/CHMP/539146/2013.  In case (2), the aim of the subgroup analyses is confirmatory and the analysis of a subgroup is part of the main objective of the trial. In this case multiplicity adjustments are necessary. The principles of “pre-specification” and “appropriate statistical analysis strategies” are stated. However, the conclusion on “restricted licence” is incomplete. Please provide clear guidance on what to expect if the full population is not significant and the subgroup is. |  |
| Line 425 |  | **Comment:**  It would be helpful to clarify the scope of “*generally*”. |  |
| Lines 427-428 |  | **Comment:**  It is not clear which party can assume heterogeneity of the treatment effect and whether this is done a priori or after results have been seen. If heterogeneity can be reasonably assumed for subgroups of sizes insufficient to evaluate, this would raise questions about the soundness of the design. Clarification would be helpful. |  |
| Line 440 |  | **Comment:**  Stratifying for important explanatory variables is important for all studies, irrespective of whether sub-group testing is planned. Perhaps more relevant advice is to stratify according to the sub-group definition(s). |  |
| Lines 450-459 |  | **Comment:**  The agency is encouraged to better contextualize its considerations on responder analysis. For example, it is not clear in this case whether the responder analysis means a dichotomization of the original continuous primary endpoint and whether it corresponds to a secondary or an exploratory endpoint. It is also not clear whether it can be included in the multiplicity adjustment, if needed. A more concrete example would be helpful to understand the definition and the role of responder analyses in the context of the multiplicity adjustment. Please also note that a responder analysis does not establish clinical relevance according to Snapinn and Jiang (2014)  Reference: Snapinn and Jiang (2014) Incorporation of Clinical Meaningfulness Into the Analysis of a Continuous Variable: A More Powerful Alternative to the Responder Analysis, Statistics in Biopharmaceutical Research, 6:4, 349-355. |  |
| Lines 472; 493-494 |  | **Comment:**  It seems too strong a statement to mandate that all components of a composite “*should be analysed separately*”. As stated in Lines 486-487, some components may be quite rare, so that formal analyses would not be very meaningful. Also, individual components will often be subject to serious competing risks issues, with no fully satisfactory analytical solution (e.g., analysing a mild component when components that are more serious manifestations of the same condition are present).  The document might clarify that further analyses of certain individual components or groups of components are generally indicated and should be pre-defined appropriately, depending on frequency, clinical rationale, and interpretability. |  |
| Line 478 |  | **Comment:**  The guidance states that “*there are two types of composite endpoints*”. However, there might other composite endpoints, such as ranked composite endpoints where we can combine different types of events: for example, survival and quantitative endpoints.  **Proposed change:**  It might be helpful if the agency could change the wording in that effect. |  |
| Lines 486-489 |  | **Comment:**  This wording suggests that composites are mainly used to increase power or decrease sample size because individual components are too infrequent. While this is certainly true in many cases, the other typical reason for composites is based on clinical rationale, for example, including more serious manifestations of the same disease condition. This aspect is discussed however in the first paragraph of Section 9.2. Because this aspect of composites applies to all subsections, we propose to move the first paragraph in Section 9.2 before introducing Section 9.1.  **Proposed change:**  “*The components may represent relatively rare events, and to study each component separately would require unmanageably large sample sizes. In this regard, composite endpoints often present a means to increase the percentage of patients that reach the clinical outcome, and hence increase the power of the study.*” (Lines 486-489)  In addition, move Lines 500-510 right after Line 489:  “*Additionally, a composite endpoint must make sense from a clinical perspective. For any component that is included in the composite …*” |  |
| Lines 495-497 |  | **Comment:**  Does the sentence say that the sponsor does not need to adjust for multiplicity for the individual components, but merely pre-specify that it will want to make claims (of benefit) on them? A clarification of this point would be appreciated.   * Related to this question: Only the composite endpoint is usually properly powered and analyses of component endpoints, especially rare components, will often lack power. How should it be evaluated in this setting whether treatments beneficially affect all components? Any guidance would be much appreciated. |  |
| Lines 495-497 |  | **Comment:**  Clarification of the interpretation of component endpoints would be appreciated, given the issue of competing risks. Some subjects may experience multiple components of the composite endpoint (e.g. myocardial infarction, then stroke, and then death). Traditional analyses focused on the type of the patient's first event only but this has been criticized for potentially neglecting later potentially more severe events. |  |
| Line 497 |  | **Comment:**  It would be helpful to clarify what is meant by “*valid confirmatory analysis strategy*”. |  |
| Lines 498-499 |  | **Comment:**  The title of Section 9.2 could suggest to a reader that the magnitude of effect on each component should be similar (e.g., similar hazard ratios), which may not be generally a requirement or expectation. It is frequently the case that there is a larger impact on less serious components than on the more severe ones, and the reverse may also be true. It would seem that the most typical requirement is that components represent different manifestations of the same disease, and an effective treatment would hopefully have a positive effect on all (as more clearly suggested in Line 514).  **Proposed change:**  “*9.2. Treatment should be expected to affect all components in the beneficial direction*” |  |
| Lines 511-513 |  | **Comment:**  In Line 511, “*does not beneficially affect all components*” sounds as if this judgement is made based on the data obtained in the study; but then Line 513 is confusing, because it suggests that this can affect the choice of the primary variable, which has to be made in advance. Please clarify. A similar concern applies to Lines 474-476.  **Proposed change:**  “*If scientific rationale or past experience suggests that treatment does not beneficially affect all components of a composite endpoint, …*”  We also propose to make the same change to Lines 474-476. |  |
| Line 513 |  | **Comment:**  The sentence “*and the choice of composite as the primary variable should be carefully considered*.” might be misunderstood to allow data dredging after completing a study. At this time in the study it seems too late to make these considerations.  **Proposed change :**  “*… and the choice of composite as the primary variable should be carefully considered at the protocol design stage.*” |  |
| Lines 528-546 |  | **Comment:**  Wording should be consistent to avoid confusion: “*adverse effect*” (Line 534) and “*negative effect*” (Line 545; used in the context of non-inferiority studies) seem to have the same meaning. One of the two terms only should be used. |  |
| Line 561 |  | **Comment:**  We agree with the importance of unbiased estimation and simultaneous confidence regions, and the fact that they are difficult to derive for many multiple testing procedures. However, it is not clear what Section 10 currently recommends in this regard. For example, is it requested to provide simultaneous confidence intervals? If so, should they be “informative” (Line 593), and what is meant by this term? If not, what are the recommendations in Section 10.2? Likewise, clearer recommendations and regulatory guidance on how to address selection bias in practice (Section 10.1) would be helpful.  See also the specific comment on Lines 588. |  |
| Lines 584-585 |  | **Comment:**  This sentence could be misleading because an early selection leads to a lower bias in estimation for the selected dose but a higher chance to select the wrong dose.  **Proposed change:**  “*Selection bias is usually lower (but still present) if, for example, the selection of one of several doses is performed at an interim analysis and the effect of the selected dose is then estimated including additional patients none of whom are recruited into the discontinued dose arms. However, selection at an earlier interim analysis ~~leads to a lower bias, although it is less informative.~~ is also less informative with a higher chance to select the wrong dose.*” |  |
| Lines 586-587 |  | **Comment:**  Selection bias in an early interim analysis could be huge if the selection of one dose is in the wrong direction due to limited information at the early interim analyses. The guideline proposes to use shrinkage estimation or model based analysis to reduce selection bias. This statement is based on correct model assumptions for the estimation. Can the agency elaborate on this situation? |  |
| Line 588 |  | **Comment:**  The guideline provides a recommendation on selection of multiple testing procedures for the type I error control from the estimation aspect, i.e., corresponding methods for constructing consistent multiplicity adjusted confidence intervals/regions should be available for multiplicity control procedures and methods. As pointed out in the Guideline, methods for constructing adjusted confidence regions/intervals may not be available due to the complexity of multiplicity adjustment in the procedure. Examples include complex gatekeeping procedures and comprehensive graphic approaches for multiplicity control across multiple families of endpoints which could be a mixture of continuous, binary, and survival endpoints. These complex procedures do have a strong control of the study-wise family error rate but methods for constructing multidimensional confidence regions might not be available. These complex procedures are valid statistical procedures and endorsed by other regulatory agencies for multiplicity control in clinical trials designed for regulatory approvals.  EMA could provide further clarification in terms of selection of complex multiplicity adjustment procedures/methods in this guideline or provide additional recommendations on estimation methods that can be used to meet the clinical relevance requirement for these complex procedures. For example, would it be acceptable to suggest that unadjusted confidence intervals (and p-values) should also be presented as supplementary descriptive information regardless of the multiplicity procedure? This is particularly relevant in the situation where meaningful confidence regions corresponding to the multiplicity procedure do not exist as overly conservative intervals may not be very informative. |  |
| Lines 593-599 |  | **Comment:**  This paragraph states that, when the confidence regions do not correspond to the hypothesis testing procedure, decision should be based on the testing procedure and a simple but conservative confidence interval can be used. However Lines 81-84 and Lines 569-570 seem to suggest that the testing procedure may not be valid or chosen in this situation. It would be helpful to modify Lines 81-84 and Lines 569-570 to ensure consistency. |  |