EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

15 December 2016
EMA/CHMP/44762/2017
Committee for Human Medicinal Products (CHMP)

Guideline on multiplicity issues in clinical trials

Draft

Draft agreed by Biostatistics Working Party (BSWP)	November 2016
Adopted by CHMP for release for consultation	15 December 2016
Start of public consultation	01 April 2017
End of consultation (deadline for comments)	30 June 2017

This guideline replaces the 'Points to consider on multiplicity issues in clinical trials' (CPMP/EWP/908/99).

Comments should be provided using this template. The completed comments form should be sent to Multiplicity_GL@ema.europa.eu.

Keywords	***Multiplicity, hypothesis test, type I error, subgroup, responder, estimation, confidence interval***

30 Churchill Place ● Canary Wharf ● London E14 5EU ● United Kingdom
Telephone +44 (0)20 3660 6000 **Facsimile** +44 (0)20 3660 5555
Send a question via our website www.ema.europa.eu/contact

Guideline on multiplicity issues in clinical trials

Table of contents

1. Executive summary

This guideline is intended to provide guidance on how to deal with multiple comparison and control of type I error in the planning and statistical analysis of clinical trials.

In 2002 the EMA Points to Consider on Multiplicity issues in clinical trials (EMA/286914/2012) was adopted. Following the EMA Concept paper on the need for a guideline on multiplicity issues in clinical trials which was published in 2012, this guideline was developed as an update of the above mentioned Points to Consider considering new regulatory advisements, including a new section on multiplicity in estimation, accounting for new approaches in dose finding and clarifying specific issues and applications.

The present document should be considered as a general guidance. The main considerations for multiplicity issues encountered in clinical trials are described. Specific issues, including adjustment of elementary hypothesis tests for multiplicity, multiple primary endpoints, analysis sets and alternative statistical methods are addressed.

The main scope is to provide guidance on the confirmatory conclusions which are usually based on the results from pivotal Phase III trials and, to a lesser extent, on Phase II studies. The guideline mainly discusses issues in decision making for a formal proof of efficacy.

In clinical studies it is often necessary to answer more than one question about the efficacy (or safety) of the experimental treatment in a specific disease, because the success of a drug development programme may depend on a positive answer to more than a single question. It is well known that the likelihood of a positive chance finding increases with the number of questions posed, if no actions are taken to protect against the inflation of false positive findings from multiple statistical tests. In this context, concern is focused on the opportunity to choose favourable results from multiple analyses. It is therefore necessary that the statistical procedures planned to deal with, or to avoid, multiplicity are fully detailed in the study protocol or in the statistical analysis plan to allow an assessment of their suitability and appropriateness.

Various methods have been developed to control the rate of false positive findings. Not all of these methods, however, are equally successful at providing clinically interpretable results and this aspect of the procedure should always be considered. Since estimation of treatment effects is usually an important issue, the availability of confidence intervals with correct coverage that allow for consistent decision making with the primary hypothesis testing strategy may be a criterion for the selection of the corresponding multiple testing procedure.

Additional claims on statistically significant and clinically relevant findings based on secondary endpoints or on subgroups are formally possible only after the primary objective of the clinical trial has been achieved (‘claim’ is used as shorthand for a confirmatory conclusion which is then prioritised in trial reporting and used as primary basis for asserting that efficacy or safety has been established), and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy.

This document should be read in conjunction with other applicable EU and ICH guidelines (see Section 4).

2. Introduction

Multiplicity of inferences is present in virtually all clinical trials. The usual concern with multiplicity is that, if it is not properly handled, unsubstantiated claims for the efficacy of a drug may be made as a consequence of an inflated rate of false positive conclusions. For example, if statistical tests are performed on five subgroups, 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 statistically significant test increases to approximately 12%.

This example shows that multiplicity can have a substantial influence on the rate of false positive conclusions which may affect approval and labelling of an investigational drug whenever there is an opportunity to choose the most favourable result from two or more analyses. If, however, there is no such choice, then there can be no influence. Examples of both situations will be discussed later. Control of the study-wise rate of false positive conclusions at an acceptable level α is an important principle and is often of great value in the assessment of the results of confirmatory clinical trials.

A number of methods are available for controlling the rate of false positive conclusions, the method of choice depending on the circumstances. Throughout this document the term 'control of type I error' rate will be used as an abbreviation for the control of the study-wise type I error in the strong sense, *i.e.* there is control on the probability to reject at least one out of several true null hypotheses, regardless of which subset of null hypotheses happens to be true.

3. Scope

The scope of this guideline is to provide guidance on the confirmatory conclusions which are usually based on the results from pivotal Phase III trials and, to a lesser extent, on Phase II studies. The guideline mainly discusses issues in decision making for a formal proof of efficacy. Due to the precautionary principle in safety evaluations, reducing the rate of false negative conclusions on harm is usually more important than controlling the number of false positive conclusions and rigorous multiplicity adjustments could mask relevant safety signals.

The principles discussed in this guideline follow the frequentist approach in statistical decision theory, where the validity of a confirmatory conclusion is defined by limiting the probability of a false positive conclusion relating to data sampling and pre-defined statistical procedures of a specific study at a pre-specified level α. 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 where more than one confirmatory trial is included in a submission.

This document does not attempt to address all aspects of multiplicity but mainly considers issues that have been found to be of importance in European marketing authorisation applications. These are:

- Adjustment of multiplicity – when is it necessary and when is it not?
- How to interpret significance with respect to multiple secondary endpoints and when can a regulatory claim be based on one of these?
- When can confirmatory conclusions be drawn from a subgroup analysis?
- How should one interpret the analysis of 'responders' in conjunction with the analysis of raw variables and how should composite endpoints be handled statistically with respect to regulatory claims?
- How should multiplicity issues be addressed in estimation?

There are further areas concerning multiplicity in clinical trials which, according to the above list of issues, are not the focus of this document. 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. However, due to the importance of the problem and the amount of information specific to this issue these aspects are discussed in the CHMP Reflection Paper on Methodological issues in Confirmatory Clinical Trials planned with an Adaptive Design (CHMP/EWP/2459/02).

Interpretations of evaluations of the primary efficacy variable at repeated visits per patient usually do not cause multiplicity problems, because in the majority of situations either an appropriate summary measure has been pre-specified or according to the requirements on the duration of treatment, primary evaluations are made at a pre-specified visit. Therefore potential multiplicity issues concerning the analysis of repeated measurements are not considered in this document.

4. Legal basis and other relevant guidance documents

This guideline has to be read in conjunction with Directive 2001/83 as amended and other applicable EU and ICH guidance documents, especially:

Note for Guidance on Dose-Response Information to Support Drug Registration - CPMP/ICH/378/95 (ICH E4)

Note for Guidance on Statistical Principles for Clinical Trials - CPMP/ICH/363/96 (ICH E9)

Guideline on the choice of the non-inferiority margin - CPMP/EWP/2158/99

Guideline on the Investigation of subgroups in confirmatory clinical trials - EMA/CHMP/539146/2013

Guideline on Clinical Development of Fixed Combination Medicinal Products – EMA/CHMP/281825/2015

Points to Consider on Application with 1. Meta-analyses and 2. One Pivotal study - CPMP/2330/99

Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design - CHMP/EWP/2459/02

5. Adjustment of elementary hypothesis tests for multiplicity – when is it necessary and when is it not?

A clinical study that requires no adjustment of the significance level of elementary hypothesis tests (*i.e.* single statistical tests on one parameter only) is one that consists of two treatment groups, which uses a single primary variable, and has a confirmatory statistical strategy that pre-specifies just one single null hypothesis relating to the primary variable and no interim analysis. Although all other situations require attention to the potential effects of multiplicity, 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, and all other endpoints are declared supportive. The assessor should expect to find in the protocol and analysis plan a discussion on the aspects of trial design, conduct and analysis that give rise to multiple testing and the proposed strategy for controlling the study-wise rate of false positive confirmatory conclusions.

Methods to control the overall type I error rate α are sometimes called multiple-level-α tests. Controlling the type I error rate study-wise is frequently done by splitting the accepted and pre-

specified type I error rate α and by then testing the various null hypotheses at fractions of α. This is usually referred to as 'adjusting the local significance level' (*i.e.* adjusting the significance level of each test). Other test procedures are available, that can be more powerful if the correlation between the test statistics are taken into account, *e.g.* the Dunnett's test on multiple comparisons to a single control. The algorithms that define how to 'spend' α are of different complexity.

In general, more than one approach is available to correctly deal with multiplicity issues. These different methods may lead to different conclusions and for this reason the details of the chosen multiplicity procedure should be part of the study protocol and should be written up without room for choice.

5.1. Multiple primary endpoints – when no formal adjustment of the significance level is needed

The ICH E9 guideline on statistical principles for clinical trials recommends that generally clinical trials have one primary variable. A single primary variable is sufficient, if there is a general agreement that a treatment induced change in this variable demonstrates a clinically relevant treatment effect on its own. If, however, a single variable is not sufficient to capture the range of clinically relevant treatment benefits, the use of more than one primary variable may become necessary. Sometimes a series of related objectives is pursued in the same trial, each with its own primary variable, and in other cases, a number of primary endpoints are investigated with the aim of providing convincing evidence of beneficial effects on some, or all of them. In these situations planning of the sample size becomes more complex due to the different alternative hypotheses related to the different endpoints and due to the assumed correlation between endpoints.

If more than one primary endpoint is used to define study success, this success could be defined by a positive outcome in all endpoints or it may be considered sufficient, if one out of a number of endpoints has a positive outcome. Whereas in the first definition the primary endpoints are designated as co-primary endpoints, the latter case is different and would require appropriate adjustment for multiplicity. More generally, in case of more than two 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.

For trials with more than one primary variable the situations described in the following subsections can be distinguished. The methods described allow clinical interpretation, deal satisfactorily with the issue of multiplicity but avoid the need for any formal adjustment of type I error rates. Other methods of dealing with multiple variables, that are more complex, are possible and can be found in the literature. In general, regulatory dialogue is recommended before applying these methods.

5.1.1. Two or more primary endpoints are needed to describe clinically relevant treatment benefits

Statistical significance is needed for all primary endpoints. Therefore, no formal adjustment of the significance level of the elementary hypothesis tests is necessary.

Here, interpretation of the results is most clear-cut because, in order to provide sufficient evidence of the clinically relevant efficacy, each null hypothesis on every primary variable has to be rejected at the same significance level (*e.g.* 0.05). For example, according to the CHMP Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (EMA/CHMP/483572/2012), lung function would be insufficient as a single primary endpoint and should be accompanied by an additional co-primary endpoint, which should either be a symptom-based endpoint or a patient-related endpoint.

In these situations, there is no intention or opportunity to select the most favourable result and, consequently, the individual significance levels are set equal to the overall significance level α, *i.e.* no adjustment is necessary. Even though in this situation all hypotheses can be assessed at the same type I error level, the need for a significant result for more than one primary hypothesis will reduce the power of the statistical procedure or increase the sample size that is needed for a given power. This inflation must be taken into account for a proper estimation of the sample size for the trial.

5.1.2. Two or more endpoints ranked according to clinical relevance

No numerical adjustment of each single hypothesis test is necessary. However, 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.

Sometimes a series of related objectives is pursued in the same trial, where one objective is of greatest importance but convincing results in others would clearly add to the value of the treatment. A typical example is the reduction of mortality in acute myocardial infarction followed by prevention of other serious events. In such cases the hypotheses may be tested (and confidence intervals may be provided) according to a hierarchical strategy. The hierarchical order may be a natural one (*e.g.* hypotheses are ordered in time or with respect to the importance of the considered endpoints) or may result from the particular interests of the investigator. Hierarchical testing can be considered as a specific multiplicity procedure. Although such a procedure may be considered as a particular adjustment, no reduction or splitting of the single α levels is necessary since the pre-defined ordering avoids any choice in the assessment. The hierarchical order for testing null hypotheses, however, has to be pre-specified in the study protocol, including a clear specification of the set of hypotheses that need to be significant before the trial is claimed successful. 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. Evidently, type II errors are inflated for hypotheses that correspond to endpoints with lower ranks. Note that a similar procedure can be used for dealing with secondary endpoints (see Section 6.2).

5.2. Analysis sets

Multiple analyses may be performed on the same variable but with varying subsets of patient data. As is pointed out in ICH E9, the set of subjects whose data are to be included in the main analyses should be defined in the statistical section of the study protocol. From these sets of subjects one (usually the full set) is selected for the primary analysis.

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 type I error (in contrast, however, to the confirmatory subgroup analyses described in Section 7, see also CHMP Guideline on the Investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013)). The main purpose of such analyses is to increase confidence in the results obtained from the primary analysis.

5.3. Alternative statistical methods – multiplicity concerns

Different statistical models or statistical techniques (*e.g.* parametric vs. non-parametric or Wilcoxon test versus log-rank test) are sometimes tried on the same set of data. A two-step procedure may be applied with the purpose of selecting a particular statistical technique for the main treatment comparison based on the outcome of the first statistical (pre-)test, the first one of the two steps. Multiplicity concerns would immediately arise, if such procedures offered obvious opportunities for

selecting a favourable analysis strategy based on knowledge of the patients' assignment to treatments. In other words, the correct type I error rate refers to the overall procedure that includes the pre-test and the selected test, and therefore such a two-test procedure does not usually control the type I error. Opportunities for choice in such procedures are often subtle, especially when these procedures use comparative treatment information, and the influence on the overall type I error is difficult to assess. Applying the same line of thought, type I error control for analyses that include model selection procedures should be based on the overall procedure. Type I error control on the basis of the finally selected model only is usually not sufficient. In addition, any *post hoc* selection of the model is not considered appropriate for a confirmatory Phase III trial.

In some situations the selected statistical model is based on a formal blind review, *i.e.* on the basis of the pooled data set from the different treatment groups hiding the information on the allocated treatment. It is also important in this case that there is no inflation in the type I error. Therefore, the selection of the statistical model according to the results of a blinded analysis should be properly justified with respect to type I error control and its potential impact on the treatment effect estimate as regards bias.

In summary, the need to change or define important key features of a study on a *post hoc* basis may question the credibility of the study and the robustness of the results with the possible consequence that a further study will be necessary. Therefore, such procedures are not recommended. Confirmatory analyses should be fully and precisely pre-defined to exclude the possibility of performing different analyses *post hoc*.

5.4. Multiplicity in safety variables

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, except for the situation that the observed effects go in the opposite direction and may raise a safety concern (see also Section 9.3).

In the case of adverse effects, p-values are of very limited value as substantial differences (expressed as relative risk or risk differences) require careful assessment and will in addition raise concern, depending on seriousness, severity or outcome, irrespective of the p-value observed. A non-significant difference between treatments will not allow for a conclusion on the absence of a difference in safety. In other words, in line with general principles, a non-significant test result should not be confused with the demonstration of equivalence.

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 counterproductive for considerations of safety. It is likewise clear that in this situation there is no control of the type I error for a single hypothesis and the importance and plausibility of 'significant findings' will depend on prior knowledge of the pharmacology of the drug, and sometimes further investigations may be required.

5.5. Multiplicity concerns in studies with more than two treatment arms

As for studies with more than one primary endpoint, the proper evaluation and interpretation of a study with more than two treatment arms can become quite complex. This document is not intended to provide an exhaustive discussion of every issue relating to studies with multiple treatment arms. Therefore, the following discussion is limited to the more common and simple designs. As a general rule it can be stated that control of the study-wise type I error is a minimal prerequisite for confirmatory claims.

5.5.1. The three arm 'gold standard' design

For a disease, where a commonly acknowledged reference drug therapy exists, it is often recommended (when this can be justified on ethical grounds) to demonstrate the efficacy and safety of a new substance in a three-arm study with the reference drug, placebo and the investigational drug. Ideally, though not exclusively, the aims of such a study are to demonstrate superiority of the investigational drug over placebo (proof of efficacy) and to show that the investigational drug retains, at least, most of the efficacy of the reference drug as compared to placebo (proof of non-inferiority). If study success is defined by non-inferiority to the reference product combined with superiority to placebo both comparisons must show statistical significance at the required level and no formal adjustment of the significance level for the single hypotheses tests is necessary. In some settings, however, superiority to placebo is the main criterion for approval, and the comparison to the reference is not considered to be primary. In this case study success could be based on a significant superiority to placebo only, but any additional confirmatory conclusion on non-inferiority to the reference would require a pre-specified multiplicity procedure, e.g. a hierarchical procedure testing superiority to placebo first followed by a test on non-inferiority to the reference.

5.5.2. Proof of efficacy for a fixed combination

For fixed combination medicinal products the corresponding CPMP guideline (CPMP/EWP/240/95 Rev. 1) requires that "each substance of a fixed combination must have documented contribution within the combination". For a combination with two (mono) components, this requirement has often been interpreted as the need to conduct a study with the two components as monotherapies and the combination therapy in a three-arm study (or a four-arm study including placebo in some settings). 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.

Multiple-dose factorial designs are employed for the assessment of combination drugs for the purpose (1) to provide confirmatory evidence that the combination is more effective than either component drug alone (see ICH E4 Note for Guidance on Dose Response Information to support Drug Registration (CPMP/ICH/378/95)), and (2) to identify an effective and safe dose combination (or a range of dose combinations) for recommended use in the intended patient population. While (1) usually is achieved using global test strategies, multiplicity has to be addressed for the purpose of achieving (2).

5.5.3. Dose-response studies

Phase II dose-finding studies are usually designed to estimate the dose-response relationship, *e.g.* with an appropriate regression model, that could be used to reasonably estimate an appropriate dose. Usually the statistical inference should focus on estimation rather than on testing, and a procedure that selects the lowest dose that shows a statistically significant difference to placebo is often of limited value and can be misleading. 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. A valuable achievement in such a trial is the demonstration of an overall positive correlation of the clinical effect with increasing dose (see ICH E4, Section 3.1). 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. ICH E4 also mentions under which circumstances a dose-response study can be part of the confirmatory package and in this instance a pre-specified plan to control the type I error is of importance.

However, for pivotal Phase III studies that use several dose groups and aim at selecting and confirming one or several doses of an investigational drug for its recommended use in a specific patient population, control of the study-wise type I error is mandatory. Due to the large variety of design features, assumptions and aims in such studies, specific recommendations are beyond the scope of this document. 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.

6. How to interpret significance with respect to multiple secondary endpoints and when can a regulatory claim be based on one of these?

Multiple secondary endpoints are included in virtually all clinical trials. These secondary endpoints will usually be included with the objective of adding weight in support of the primary efficacy claim (see Section 6.1). On occasion the secondary endpoints will be included to support a second efficacy claim (see Section 6.2). For example a symptomatic effect may be a different claim from a disease-modifying effect, and treatment and maintenance of effect may be thought of as different claims. For the purpose of this document, and distinguishing between the two sub-sections below, a claim can be thought of as a confirmatory conclusion of therapeutic efficacy or safety in a particular treatment context. The reader should not directly relate use of the word claim with the possibility to make statements or present data in the Summary of Product Characteristics, which is governed by a separate regulatory guidance document. Instead, 'claim' is used as shorthand for a confirmatory conclusion which is then prioritised in a clinical study report, clinical overview or clinical summary, and is used as a primary basis for asserting that efficacy or safety has been established.

6.1. Secondary endpoints expressing supportive evidence

No claims are intended; confidence intervals and statistical tests are of descriptive nature.

Secondary endpoints may provide additional clinical characterisation of treatment effects but are, by themselves, not sufficiently convincing to establish the main evidence in an application for a licence or for an additional labelling claim. Here, the inclusion of secondary endpoints is intended to yield supportive evidence related to the primary objective, and no confirmatory conclusions are needed. Confidence intervals and statistical tests are of descriptive nature and no claims are intended.

Including secondary endpoints in a multiple testing procedure (*e.g.* a 'hierarchy') is therefore not mandated, but permits a quantification of the risk of a type I error regarding these endpoints, which may lend support that an individual result is sufficiently reliable when included in the Summary of Product Characteristics.

The ranking of endpoints in a hierarchy can be a source of controversy. In principle, the planning and assessment of a clinical trial should prioritise those endpoints of greatest interest from a clinical perspective, but it has become common practice to rank endpoints based on the likelihood that the individual null hypothesis can be rejected. Ideally the clinical assessment should focus on those endpoints of greater clinical importance and the sponsor runs a risk of type II error if the more clinically important endpoint is set below another endpoint in the hierarchy for which the individual null hypothesis is not rejected.

In the event that no formal multiple testing procedure is utilised, it can still be advantageous to specify a few key secondary endpoints in the protocol that are of greater importance for assessment since selection of positive results from an unstructured list of secondary endpoints would not generally be

considered to provide data that are reliable for inference or for presentation in the Summary of Product Characteristics.

6.2. Secondary endpoints which may become the basis for additional claims

Significant effects in these endpoints can be considered for an additional claim only after the primary objective of the clinical trial has been achieved, and if they were part of the confirmatory strategy.

Secondary endpoints may be related to secondary objectives that become the basis for an additional claim, once the primary objective has been established (see Section 5.1.2). A possible simple procedure to deal with this kind of secondary endpoint is to proceed hierarchically; other procedures are also available. Once the null hypothesis concerning the primary objective is rejected (and the primary objective is thus established), further confirmatory statistical tests on secondary endpoints can be performed using a hierarchical order for the secondary endpoints if there is more than one. In this case, primary and secondary endpoints differ just in their place in the hierarchy of hypotheses which, of course, reflects their relative importance in the study. 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.

6.3. Secondary endpoints indicative of clinical benefit

If not defined as primary endpoints, clinically very important endpoints (e.g. mortality) need further study when significant benefits are observed, but the primary objective has not been achieved.

Endpoints that have the potential of being indicative of a major clinical benefit or may in a different situation present an important safety issue (*e.g.* mortality) may be relegated to secondary endpoints because there is an *a priori* belief that the size of the planned trial is too small (and thus the power too low) to show a benefit. If, however, the observed beneficial effect is much higher than expected but the study falls short of achieving its primary objective, this would be a typical situation where information from further studies would be needed to support the observed beneficial effect.

If, however, the same endpoint that may indicate a major clinical benefit exhibits a treatment effect in the opposite direction, this would give rise to safety concerns (in the example of increased mortality). A Marketing Authorisation may not be granted, regardless of whether or not this endpoint was embedded in a confirmatory scheme.

7. Reliable conclusions from a subgroup analysis, and restriction of the licence to a subgroup

Reliable conclusions from subgroup analyses generally require pre-specification and appropriate statistical analysis strategies. A licence may be restricted if unexplained strong heterogeneity is found in important sub-populations, or if heterogeneity of the treatment effect can reasonably be assumed but cannot be sufficiently evaluated for important sub-populations.

In clinical trials there are many reasons for examining treatment effects in subgroups. In many studies, subgroup analyses have a supportive or exploratory role after the primary objective has been accomplished. A specific claim of a beneficial effect in a particular subgroup requires pre-specification of the corresponding null hypothesis (including the precise definition of the subgroup) and an appropriate confirmatory analysis strategy. Multiplicity issues arise if study success is defined by the demonstration of a beneficial effect of the treatment in the whole study population or in a pre-defined

subgroup (or in one of several subgroups). An appropriate pre-planned multiplicity adjustment is needed for an unambiguous confirmatory conclusion. The complexity of the multiplicity procedure is increased if decision making is possible at an interim time point or after the final analysis. The number of subgroups should be small, in order to efficiently apply an appropriate multiplicity procedure.

Considerations of power are expected to be covered in the protocol, and randomisation would generally be stratified by the most important explanatory covariates. Decision making based on subgroup analyses in general are dealt with in the CHMP guideline on the Investigation of Subgroups in Confirmatory Clinical Trials (EMA/CHMP/539146/2013).

8. How should one interpret the analysis of 'responders' in conjunction with the raw variables?

If the 'responder' analysis is not the primary analysis it may be used after statistical significance has been established on the mean level of the required primary endpoint(s), to establish the clinical relevance of the observed differences in the proportion of 'responders'. When used in this manner, the test of the null hypothesis of no treatment effect is better carried out on the original primary variable than on the proportion of responders.

In a number of applications, for example those concerned with Alzheimer's disease or depressive disorders, it may be difficult to interpret small but statistically significant improvements in the mean level of the primary endpoint. For this reason the term 'responder' (and 'non-responder') is used to express the clinical benefit of the treatment in terms of effects seen in individual patients. There may be a number of ways to define a 'responder'/'non-responder'. The definitions should be pre-specified in the protocol and should be clinically convincing. In clinical regulatory guidelines, it is stated that the 'responder' analysis should be used in establishing the clinical relevance of the observed effect as an aid to assess efficacy and clinical safety. It should be noted that in instances there is some loss of information (and hence loss of statistical power) connected with breaking down the information contained in the original variables into 'responder' and 'non-responder'.

In some situations, the 'responder' criterion may be the primary endpoint (*e.g.* CHMP guideline on clinical investigation of medicinal products in the treatment of Parkinson's disease (EMA/CHMP/330418/2012 rev. 2)). In this case it should be used to provide the main test of the null hypothesis. However, the situation that is primarily addressed here is when the 'responder' analysis is used to allow a judgement on clinical relevance, once a statistically significant treatment effect on the mean level of the primary variable(s) has been established. In this case, the results of the 'responder' analysis need not be statistically significant but the difference in the proportions of responders should support a statement that the investigated treatment induces clinically relevant effects.

It should be noted that a 'responder' analysis cannot rescue the negative results on the primary endpoint(s).

9. How should composite endpoints be handled statistically with respect to regulatory claims?

Usually, the composite endpoint is primary. All components should be analysed separately. If claims are based on subgroups of components, this needs to be pre-specified and embedded in a valid confirmatory analysis strategy. In the event that treatment does not beneficially affect all components, in particular where the clinically more important components are affected negatively, interpretation will be very difficult. Any effect of the treatment in one of the components that is proposed to be reflected in the product information should be clearly supported by the data.

There are two types of composite endpoints. The first type, namely the rating scale, arises as a combination of multiple clinical measurements. With this type there is a longstanding experience and/or validation of its use in certain indications (*e.g.* psychiatric or neurological disorders). This type of composite variable is not discussed further in this guideline.

The other type of a composite variable arises in the context of survival analysis. Several events are combined to define a composite outcome. A patient is said to have the clinical outcome if s/he suffers from one or more events in a pre-specified list of components (*e.g.* death, myocardial infarction or disabling stroke). The time to outcome is measured as the time from randomisation of the patient to the first occurrence of any of the events in the list. Usually, the components represent relatively rare events, and to study each component separately would require unmanageably large sample sizes. Composite endpoints therefore often present a means to increase the percentage of patients that reach the clinical outcome, and hence increase the power of the study.

9.1. The composite endpoint as the primary endpoint

When a composite endpoint is used to show efficacy it will often be the primary endpoint. In this case, it must meet the requirements for a single primary endpoint, namely that it is capable of providing the key evidence of efficacy that is needed for a licence. It is recommended to analyse in addition the single components and clinically relevant groups of components separately, to provide supportive information. There is, however, no need for an adjustment for multiplicity provided significance of the primary endpoint is achieved. If claims are to be based on (subgroups of) components, this needs to be pre-specified and embedded in a valid confirmatory analysis strategy.

9.2. Treatment should be expected to affect all components in a similar way

A composite endpoint must make sense from a clinical perspective. For any component that is included in the composite, it is usually appropriate that any additional component reflecting a worse clinical event is also included. For example, if it is agreed that hospitalisation is an acceptable component in a composite endpoint, it would be usual to also include components for more adverse clinical outcomes that are relevant to the clinical setting (*e.g.* non-fatal myocardial infarction and stroke) and death. Excluding such events, with an argument that no beneficial effect can be expected or that these will be captured in the safety assessment, or focussing on specific types of events (for example disease-related mortality in preference to all-cause mortality) introduces difficulties for analysis and interpretation that should be approached carefully. In this event, the primary composite should always be presented and interpreted alongside a secondary analysis in which no important clinical outcomes are excluded.

In the event that treatment does not beneficially affect all components of a composite endpoint, in particular where the clinically more important components are affected negatively, interpretation will be complicated and the choice of composite as the primary variable should be carefully considered. An assumption of similarly directed treatment effects on all components should be based on past experience with studies of similar type. Whilst it may often be reasonable, *a priori*, to assume that no component of a composite relating to efficacy will be adversely affected, 'net clinical benefit' endpoints are employed to investigate whether beneficial effects are offset by increased detrimental effects. Because of the assumptions made in 'weighting' the components and in the overall interpretation, such composites will not usually be appropriate primary endpoints.

Composite endpoints also pose particular issues in the non-inferiority or equivalence setting, and analogously in relation to assessment of safety. Adding a component that foreseeably is insensitive to treatment effects tends to decrease sensitivity of the comparison, even if it does not affect

unbiasedness of the estimation of the treatment difference. An increased variance is an undesirable property in non-inferiority or equivalence studies. For non-inferiority or equivalence studies the more specific component (*e.g.* disease related mortality) can be preferred as primary endpoint for this reason, though again both this and the more general composite including all relevant events should be considered together.

9.3. The clinically more important components should at least not be affected negatively

If time to hospitalisation is an endpoint in a clinical study 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 as components, including death in this example. One concern with composite outcome measures from a regulatory point of view is, however, the possibility that some of the treatments under study may have an adverse effect on one or more of the components, and that this adverse effect is masked by the composite outcome, *e.g.* by a large beneficial effect on some of the remaining components. This concern is particularly relevant if the components relate to different degrees of disease severity or clinical importance. For example, if all-cause mortality is a component, a separate analysis of all-cause mortality should be provided to ensure that there is no adverse effect on this endpoint. Since there is no general agreement on how much evidence is needed to generate suspicion of an adverse effect, it is recommended that this issue is addressed at the planning stage. For example, the study plan could address the size of the risk of an adverse effect on the more serious components that can be excluded (assuming no treatment difference under the null hypothesis) with a sufficiently high probability given the planned sample size, and the study report should contain the respective comparative estimates and confidence intervals.

Non-inferiority studies will also be particularly hard to interpret if negative effects on some components are observed for the experimental drug and are outbalanced by other components of the composite.

9.4. Any effect of the treatment on one of the components that is intended to be reflected in the product information should be clearly supported by the data

An important issue for consideration is the claim that can legitimately be made based on a successful primary analysis of a composite endpoint. Difficulties arise if the claims do not properly reflect the fact that a composite endpoint was used, *e.g.* if a claim is made that explicitly involves a component with the lowest frequency amongst all components. For example, if the composite outcome is death or liver transplantation and there are only a few deaths, a claim to reduce mortality and the necessity for liver transplantation would not be satisfactory, because in this context the effect on mortality will have a weak basis. This does not mean that one should drop the component death from the composite outcome, because the outcome liver transplantation would be incomplete without simultaneously considering all disease-related outcomes that are at least as serious as liver transplantation. However, it does mean that different wording should be adopted in the product information, avoiding the implication of a demonstrated effect on mortality.

10. Multiplicity issues in estimation

Often, for the more complex procedures, clinical interpretation of the findings can become difficult. For the purpose of estimation and for the appraisal of the precision of estimates, confidence intervals are of paramount importance. Multiple confidence intervals with an adjusted confidence level or multidimensional confidence regions (covering more than one unknown parameter with a given probability for the simultaneous assessment of multiple parameters) are typically used for multiple

comparisons but methods for their construction that are consistent with the tests are not available or not useful for many of the complex multiple testing procedures used to control the type I error. Nevertheless, a valid statistical procedure is useful only if it allows for a meaningful and informative clinical interpretation. 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.

10.1. Selection bias

Multiple comparisons may lead to a bias in estimation which is defined by the difference between the mean estimation and the parameter to be estimated. For example, in a situation where several treatment groups are compared to placebo the strategy that chooses the treatment with the largest difference to placebo as the treatment that should be marketed will, on average, lead to an overestimation of the corresponding treatment effect. If selection is made not on the basis of the treatment effect it may still be based on an endpoint that is correlated with efficacy.

Whereas the term selection bias often relates to the bias resulting from a specific patient or subgroup selection, selection bias in the context of multiple comparisons refers to a biased estimation resulting from selecting a specific treatment (*e.g.* a specific dosage) based on the data that are subsequently used for estimation.

Selection bias is usually lower (but still present) if the selection is performed at an interim analysis. Selection at an earlier interim analysis leads to a lower bias, although it is less informative. However, methods are available to reduce selection bias, such as shrinkage estimation or specific model based analyses. Maximum bias should be gauged in order to account for it in the risk benefit assessment.

10.2. Confidence intervals

As can occur with multiple testing, multiple confidence intervals may also increase the chance of false decisions since the probability that a set of multiple non-adjusted confidence intervals cover correctly all parameters to be estimated would usually be less than the pre-specified nominal coverage probability related to the single confidence intervals.

Informative confidence regions that correspond to multiplicity procedures may, however, not always be available or may be difficult to derive. If the confidence regions do not correspond to the hypothesis testing procedure, different conclusions are possible, *e.g.* a confidence interval excluding the null hypothesis combined with a non-significant testing result or *vice versa*. The decision should, however, be based on the hypothesis test. In that case it is advised to use simple but conservative confidence interval methods, such as Bonferroni-corrected intervals, ensuring that the uncertainty about the beneficial effects is properly understood.