The European Agency for the Evaluation of Medicinal Products
Evaluation of Medicines for Human Use

London, 19 September 2002
CPMP/EWP/908/99

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP)

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

DISCUSSION IN THE EFFICACY WORKING PARTY	January 2000
TRANSMISSION TO CPMP	July 2001
RELEASE FOR CONSULTATION	July 2001
DEADLINE FOR COMMENTS	October 2001
DISCUSSION IN THE EFFICACY WORKING PARTY	June 2002
TRANSMISSION TO CPMP	September 2002
ADOPTION BY CPMP	September 2002

7 Westferry Circus, Canary Wharf, London, E14 4HB, UK
Tel. (44-20) 74 18 84 00 Fax (44-20) 74 18 8613
E-mail: mail@emea.eu.int http://www.emea.eu.int

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

1. 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 effectiveness 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 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 family-wise type I error in the strong sense, i.e., there is control on the probability to reject at least one true null hypothesis, regardless which subset of null hypotheses happens to be true. The issue of setting an appropriate type I error level on a submission level when this includes the need for more than one confirmatory trial is discussed in a separate Points-to-Consider document (CPMP/2330/99 Points to Consider on Application with 1.) Meta-analyses and 2.) One Pivotal study).

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

- Adjustment of multiplicity – when is it necessary and when is it not?
- How to interpret significance with respect to multiple secondary variables and when can a claim be based on one of these?
- When can reliable conclusions be drawn from a subgroup analysis?
- When is it appropriate for CPMP to restrict licence to a subgroup?
- How should one interpret the analysis of "responders" in conjunction with the raw variables?
- How should composite endpoints be handled statistically with respect to regulatory claims?

There are further areas concerning multiplicity in clinical trials which, according 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 the claim of effectiveness) or stepwise designed studies (with the possibility for adaptive changes for the future steps). However, due to the importance of the problem and the amount of information specific to this issue it appears appropriate that a separate document may cover these aspects.

Interpretations of repeated evaluations of the primary efficacy variable at repeated visits 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 endpoint, primary evaluations are made at pre-specified visits. Therefore potential multiplicity issues concerning the analysis of repeated measurements are not considered in this document.

2. ADJUSTMENT FOR MULTIPLICITY – WHEN IS IT NECESSARY AND WHEN IS IT NOT?

A clinical study that requires no adjustment of the type I error is one that consists of two treatment groups, that 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 many situations where no multiplicity concern arises, for example, having predefined the primary variables and all secondary variables are declared supportive.

In the literature, methods to control the overall type I error α are sometimes called multiple-level-α-tests". Controlling type I error family-wise often (but not always) means that the accepted and pre-specified amount α of type I error has to be split, and that the various null hypotheses have to be tested at the resulting fraction of α. This is usually referred to as 'adjusting the type I error level'. The algorithms that define how to "spend" α in this way are of different complexity. Often, for the more complex procedures, clinical interpretation of the findings can become difficult. For example, for the purpose of estimation and for the appraisal of the precision of estimates, confidence intervals are of paramount importance but methods for their construction that are consistent with the tests are not available for many of the more complex multiple-level-α-tests (or more generally closed tests) aiming at controlling the type I error. When choosing an approach, it is recommended to consider whether the existing valid statistical procedures allow a satisfactory clinical interpretation.

Because alternative methods to deal correctly with multiplicity are often available which may lead to different conclusions, pre-definition of the preferred multiple-level-α-test is necessary. To avoid problems in interpretation, details of the procedure should be contained in the study protocol or the statistical analysis plan.

If a multiple test situation occurs which was not foreseen, a conservative approach will be necessary e.g. Bonferroni's or a related procedure. Inherently there will be a loss of power. Therefore if a multiple test situation is foreseen pre-specification of the method use to deal with this is recommended

This document discusses situations with relevance for multiple testing in clinical trials and commonly practised and acknowledged methods for controlling (or adjusting) type I error.

2.1 Multiple primary variables – when no formal adjustment is needed.

The ICH E9 guideline on biostatistical principles in 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 clinical 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 variables 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 because alternative hypotheses and limits for the power of the single primary variables have to be defined and balanced against each other to give the study a solid basis to meet its objectives.

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. Indeed the methods are members from the set of closed testing procedures that control the family-wise error rate.

2.1.1. Two or more primary variables are needed to describe clinically relevant treatment benefits

Statistical significance is needed for all primary variables. Therefore, no formal adjustment is necessary.

Here, interpretation of the results is most clear-cut because, in order to provide sufficient evidence of the clinically relevant treatment benefit, each null hypotheses on every primary variable has to be rejected at the same significance level (e.g. 0.05). For examples of this clinical situation, see CPMP Note for Guidance for the treatment of Alzheimer's disease, or CPMP Points to Consider on clinical investigation of medicinal products in the chronic treatment of patients with chronic obstructive pulmonary disease. In these situations, there is no intention or opportunity to select the most favourable result and, consequently, the individual type I error levels are set equal to the overall type I error level α, i.e. no reduction is necessary. This procedure inflates the relevant type II error (here: falsely accepting that at least one null hypothesis is true), which in the worst case scenario is the sum of the type II errors connected with the individual hypotheses. This inflation must be taken into account for a proper estimation of the sample size for the trial.

2.1.2. Two or more primary variables ranked according to clinical relevance

No formal adjustment is necessary. However, no confirmatory claims can be based on variables 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. Typical examples are (i) acute effects in depressive disorders followed by prevention of progression (ii) 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 seriousness of the considered variables) or may result from the particular interests of the investigator. Again, no reduction or splitting of α is necessary. The hierarchical order for testing null hypotheses, however, has to be pre-specified in the study protocol. The effect of such a procedure is that no confirmatory claims can be based on variables that have a rank lower than or equal to that variable whose null hypothesis was the first that could not be rejected. Confidence intervals that are consistent with this hierarchical test procedure can be derived. Evidently, type II errors are inflated for hypotheses that correspond to variables with lower ranks. Note that a similar procedure can be used for dealing with secondary variables (see 3.2).

In the literature it is possible to find many methods of dealing with multiple variables that are of value for situations which may, however, be rarely met in confirmatory clinical trials, and

which, therefore, are not discussed in this document. Before applying such methods regulatory dialogue is recommended.

2.2 Analysis sets

Multiple analyses may be performed on the same variable but with varying subsets of patient data. As is pointed out in the ICH E9 guideline on biostatistical principles for clinical trials, 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 analyses on varying subsets of subjects or with varying measurements for the purpose of investigating the sensitivity of the conclusions drawn from the primary analysis should not be subjected to adjustment for type I error. The main purpose of such analyses is to increase confidence in the results obtained from the primary analysis

2.3 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. Sometimes a two step procedure is 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. 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. There are situations, where selecting the final statistical model based on a formal Blind Review (see ICH E9) is exempted from such concerns. Opportunities for choice in such procedures are often subtle, when these procedures use comparative treatment information, and the influence on the overall type I error is difficult to assess. Finally, the need to change 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 cannot be recommended even when it appears that there is no element of choice.

2.4 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 variables, except for the situation that the observed effects show in the opposite direction and may raise a safety concern (see also 3.3). In the case of adverse effects p-values are of very limited value as substantial differences (expressed as relative risk or risk differences) will raise concern, depending on seriousness, severity or outcome, irrespective of the p-value observed.

In those cases where a large number of statistical test procedures is used 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 clear that in this situation there is no control over the type I error for a single hypothesis and the importance and plausibility of such results will depend on prior knowledge of the pharmacology of the drug.

2.5 Multiplicity concerns in studies with more than two treatment arms

As for studies with more than one primary variable, 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, only rarely have these more complex designs been applied in confirmatory clinical trials. 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 family-wise type I error in the strong sense (i.e. application of closed test procedures) is a minimal prerequisite for confirmatory claims. It should be remembered that the usual confidence intervals for the pairwise differences between treatment groups are – except for a few instances - not consistent with the closed testing procedures, and are usually too narrow.

2.5.1 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 three treatments: the reference drug, placebo and the investigational drug. Usually the aims of such a study are manifold: (1) to demonstrate superiority of the investigational drug over placebo (proof of efficacy); (2) to demonstrate superiority of the reference drug over placebo (proof of assay sensitivity, see ICH E10, section 2.5.1.1.1); and (3) to demonstrate that the investigational drug retains most of the efficacy of the reference drug as compared to placebo (proof of non-inferiority). If all of these are objectives , all three comparisons must show statistical significance at the required level, and no formal adjustment is necessary. A failure to show the investigational drug as superior to placebo could then be explained either as the investigational drug being not effective (when the reference drug showed superiority over placebo), or as lack of assay sensitivity (when test and reference drug failed to show superiority over placebo).

2.5.2 Proof of efficacy for a fixed combination

For fixed combination medicinal products the corresponding CPMP guideline (CPMP/EWP/240/95) 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 mono therapies and the combination therapy in a 3-arm study. Such a study is considered successful, if the combination is shown superior to both components. No formal adjustment of the overall significance level is necessary, because both pairwise comparisons must show statistically significant superiority.

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 useful dose combinations) for recommended use in the intended patient population. While (1) usually is achieved using global test strategies, appropriate closed test procedures have to be applied for the purpose of achieving (2).

2.5.3 Dose-response studies

For therapeutic dose response studies that aim at identifying one or several doses of an investigational drug for its recommended use in a specific patient population, the control of the family-wise type I error in the strong sense is mandatory. Due to the large variety of design features, assumptions and aims in such studies (e.g. assuming or not assuming monotonicity of the dose response with increasing dose; finding the minimally effective dose under the constrains of the used design; finding a dose that is equivalent (non-inferior) to the

recommended dose of a reference drug), specific recommendations are beyond the scope of this document. There are various methods published in the relevant literature on closed test procedures with relevance to multiple dose studies that can be adapted to the specific aims and that provide the necessary control on the type I error.

Sometimes a study is not powered sufficiently for the aim to identify and recommend a single effective and safe dose (or a dose range) but is successful only at demonstrating an overall positive correlation of the clinical effect with increasing dose. This is already a valuable achievement (see ICH E4, section 3.1). Estimates and confidence intervals from pairwise comparisons of single doses are then used in an exploratory manner for the planning of future studies. In this case, an adjustment of the type I error is not necessary.

3. HOW TO INTERPRET SIGNIFICANCE WITH RESPECT TO MULTIPLE SECONDARY VARIABLES AND WHEN CAN A CLAIM BE BASED ON ONE OF THESE?

Traditionally, in clinical trial protocols there will be a number of secondary variables for efficacy. Up to now there has been no common consent about the role and the weight of secondary endpoints in clinical trials.

3.1 Variables expressing supportive evidence

No claims are intended; confidence intervals and statistical tests are of exploratory 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 license or for an additional labelling claim. Here, the inclusion of secondary variables is intended to yield supportive evidence related to the primary objective, and no confirmatory conclusions are needed. Confidence intervals and statistical tests are of exploratory nature and no claims are intended.

3.2 Secondary variables which may become the basis for additional claims

Significant effects in these variables 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

More importantly, secondary variables may be related to secondary objectives that become the basis for an additional claim, once the primary objective has been established (see 2.1.2). A valid procedure, to deal with this kind of secondary variable is to proceed hierarchically. Once the null hypothesis concerning the primary objective is rejected (and the primary objective thus established), further confirmatory statistical tests on secondary variables can be performed using a further hierarchical order for the secondary variables themselves if there is more than one. In this case, primary and secondary variables differ just in their place in the hierarchy of hypotheses which, of course, reflects their relative importance in the study. It is of note to mention that changes in secondary variables that are considered a direct consequence of the respective changes in the primary variables cannot be part of the labelling claims. For example, symptoms of depression in schizophrenic patients disappear as patients get into remission from schizophrenia. In this situation, a separate labelling claim on an anti-depressive action of the treatment cannot be made.

3.3 Variables indicative of clinical benefit

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

Variables 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 variables 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 fell short of achieving its primary objective, this would be a typical situation where information from further studies would be needed which can be used in support of the observed beneficial effect.

If however, the same variable that may indicate a major clinical benefit exhibits treatment effects in the opposite direction this would give rise to concerns about the safety. A license may then well be refused, regardless of whether or not the variable was embedded in a confirmatory scheme.

4. RELIABLE CONCLUSIONS FROM A SUBGROUP ANALYSIS, AND RESTRICTION OF THE LICENSE TO A SUBGROUP

Reliable conclusions from subgroup analyses generally require pre-specification and appropriate statistical analysis strategies. A license 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, i.e. the demonstration of a significant overall clinical benefit. A specific claim of a beneficial effect in a particular subgroup requires pre-specification of the corresponding null hypothesis and an appropriate confirmatory analysis strategy. It is highly unlikely that claims based on subgroup analyses would be accepted in the absence of a significant effect for the overall study population. Considerations of power would be expected to be covered in the protocol, and randomisation would generally be stratified.

The evaluation of uniformity of treatment effects across subgroups is a general regulatory concern. Some factors are known to cause heterogeneity of treatment effects such as gender, age, region, severity of disease, ethnic origin, renal impairment, or differences in absorption or metabolism. Analyses of these important subgroups should be a regular part of the evaluation of a clinical study (when relevant), but should usually be considered exploratory, unless there is *a priori* suspicion that one or more of these factors may influence the size of effect. However, when a strong interaction is found that indicates an adverse effect of the treatment in one of the subgroups and no convincing explanation for this phenomenon is available or other information confirms the likelihood of an interaction then patients from the respective sub-population may be excluded from the license until additional clinical data are available. This may also apply when there are historical reasons for regulators to believe that a certain sub-population of patients will not benefit from the drug and the results do not strongly contradict this believe.

Restriction of a license to certain subgroups is also possible, if a large variety of sub-populations are investigated without proper plans to deal with this situation in the protocol. From the regulatory perspective an overall positive result (statistically and clinically) in the whole study population may not lead to valid claims for all sub-populations if there is a

reason to expect heterogeneity of the treatment effect in the respective sub-populations. If a meaningful definition of the overall study population is lacking, licensing may be limited to sub-populations which are adequately represented and in which statistically significant and clinically relevant results were observed.

5. 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 variable(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 epileptic disorders, it is difficult to interpret small but statistically significant improvements in the mean level of the primary variables. For this reason the term "responder" (and "non-responder") is used to express the clinical benefit of the treatment to individual patients. There may be a number of ways to define a "responder"/"nonresponder". The definitions should be pre-specified in the protocol and should be clinically convincing. In clinical 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 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. CPMP guideline on clinical investigation of medicinal products in the treatment of Parkinson's disease). 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 (e.g. CPMP Note for Guidance on clinical investigation of drugs used in weight control, or on the treatment of Alzheimer's disease). 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 otherwise disappointing results on the primary variables.

6. HOW SHOULD COMPOSITE VARIABLES BE HANDLED STATISTICALLY WITH RESPECT TO REGULATORY CLAIMS?

Usually, the composite variable 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. Treatment should beneficially affect all components, or at least should the clinically more important components not be affected negatively. Any effect of the treatment in one of the components that is to be reflected in the indication should be clearly supported by the data.

There are two types of composite variables. The first type, namely the rating scale, arises as a combination of multiple clinical measurements. With this type there is a longstanding

experience 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 variables therefore present a means to increase the percentage of patients that reach the clinical outcome, and hence the power of the study.

6.1 The composite variable as the primary endpoint.

When a composite variable is used to show efficacy it will usually be the primary endpoint. Therefore, 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 license. 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.

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

When defining a composite variable it is recommended to include only components for which it can be assumed that treatment will influence them similarly. The assumption of similarly directed treatment effects on all components should be based on past experience with studies of similar type. Adding a component that foreseeably is insensitive to treatment effects will lead to an increase in variability, even if it does not affect unbiasedness of the estimation of the treatment difference. A direct consequence would be a decrease in sensitivity for demonstrating superiority between different treatment arms. An increased variance is also a undesirable property in non-inferiority or equivalence studies. Non-inferiority studies will be hard to interpret if negative effects on some components are observed. For studies aiming to show superiority the more general component is preferred as primary endpoint as this is the most conservative analysis. For non-inferiority/equivalence studies the more specific component (e.g. disease related mortality) are preferred as primary endpoint for the same reason.

6.3 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 as censored who die before they reach the hospital. It is better practice to study a composite endpoint that includes all more 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 how much less then statistical significance in the wrong direction will generate suspicion of an adverse effect, a

way to create confidence in support of 'no adverse effect', once the data is observed, is to address this issue 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.

6.4 Any effect of the treatment on one of the components that is to be reflected in the indication 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 low occurrence. 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 for the indication, avoiding the implication of an effect on mortality.

7. CONCLUSION

In clinical studies it is often necessary to answer more than one question about the efficacy (or safety) of one or more treatments in a specific disease, because the success of a drug development program may depend on a positive answer to more than a single question. It is well known that the chance of a spurious 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 different 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 connected with a particular procedure may be a criterion for its selection.

Additional claims on statistically significant and clinically relevant findings based on secondary variables or on subgroups are possible only after the primary objective of the clinical trial has been achieved, and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy.