INTERNATIONAL SOCIETY FOR CLINICAL BIOSTATISTICS

Chairman, Sub-Committee on Statistics in Regulatory Affairs: A/Prof Jørgen Seldrup, PhD
Clinical Trials and Epidemiology Research Unit, Ministry of Health, 10 College Road, Singapore 169851
Telephone: +65 220 1463, Fax: +65 220 1475, e-mail: jorgen@cteru.gov.sg

---

Comments regarding CPMP/EWP/908/99 draft: "Points to consider on multiplicity issues in clinical trials", London, 26 July 2001.

General

This is a generally well-considered document providing helpful discussion and sensible recommendations regarding many issues.

However, it is noted that in the corresponding Concept Paper of 22 April 1999 the 'comparison of more than two treatment groups' is mentioned as a problem (section 2 of the second paragraph). The current PtC does not address this issue. It would be helpful to have some guidance on this.

In addition, a multiplicity issue that is rarely discussed in its entirety involves the use of performing two studies where a treatment is considered to be of benefit only if the null hypothesis of the primary efficacy variable is rejected at the 0.05 level of significance in both studies. While adjustment of the overall power, e.g., by setting the power to detect a pre-specified difference to 0.90 in each study so that the overall power is no worse than 0.80, is sometimes discussed, the effect of this procedure on the reduction of the Type I error to possibly well below 0.05 is often missing.

Specific Points

1. Section 2.1.1
   The case is discussed where hypotheses on more than one primary variable need to be rejected at the 0.05 level of significance in order to conclude that the treatment is of benefit. It is stated that, at worst, the type II error is then the sum of the individual Type I errors. However, it may also be noted that the overall type I error is smaller than 0.05.

2. Section 2.3
   The discussion in this section seems to suggest that a test based on the unblinded data to determine which statistical test should be used (e.g., t-test or Wilcoxon test) should not be performed because it could potentially allow the investigator to choose the test most favorable to the result that is desired. Does this mean that even if the assumptions of a statistical test clearly aren't met, the investigator still has to use that test? Would that argument still be used if an investigator used what turned out to be an inappropriate test but that it yielded a more favorable

result than the more appropriate test?

It would be appropriate to refer to ICH E9 which prescribes an a-priori specification for a testing strategy.

3. Section 3.3
   The last sentence in this section states *"An adjustment for multiplicity is counterproductive for considerations of safety."* Why? The same argument for multiplicity adjustment for efficacy variables can also be made for safety variables, i.e., even if there are no differences between treatments with regard to safety, looking at many different safety variables, is liable to turn up some that are statistically significant. A multiplicity argument might imply that the difference should be ignored. However, the importance and plausibility of such a safety result should be seriously considered.

4. Section 4
   The last sentence in this section raises the danger that clinical trials may take longer to perform, since adequate representation of subgroups would require over-sampling of minorities thus potentially leading to an extension of the time needed to complete patient recruitment.
   There can be an ethical conflict in providing proof that efficacy is present in all subgroups, since this may require continued recruitment of patients beyond a point at which many physicians would consider the general question to have been answered. This is a difficult issue and on the whole the document is cautious and sensible in its discussion.
   It is stated that if we find an interaction in a particular subgroup where the treatment effect is adverse, then this subgroup can be excluded from the license. There is a multiplicity issue here as well, i.e., even if the overall effect is that the treatment is beneficial, there are liable to be subgroups where this is not the case. Hence, this seems an unreasonable action to take without further discussion. Also, the phrase *"adverse effect of treatment"* needs to be clarified. Is this a statistically significant result that is actually in favour of placebo (or comparative treatment) or is statistical significance not implied?

5. Section 5
   The discussion of analysis of “responders” is disappointing. The notion of a “responder” is part of the medical folklore and without statistical respectability. Unless correlation between outcome and baseline measures is perfect, some patients will always appear to respond better than others even if they do not. Suppose that in a parallel group trial exactly the same effect has been seen for every patient. There is thus no question of separating patients into “responders” and “non-responders”. The fallacy of this is easy to see. Consider a histogram of outcome values of the placebo group and shift each point a fixed difference to the right, corresponding to every patient having benefited by the same amount (assuming high values are good). Then, depending on the degree of shift, a different proportion of patients now has higher values than the original group mean. However, each patient has benefited by exactly the same amount, so describing the proportion of “responders” is simply nonsense. Either all responded or none did. The ‘proportion of “responders”’ is thus an indirect and inferior measure of the mean treatment effect.
   Statistical theory shows that such separation of “responders” and “non-responders” is not possible unless a repeated measures design has been used. (Ideally multiple n of 1 trials.) Even in a simple cross-over trial this is not possible. It is well known, for example, that if one wishes to analyse individual bio-equivalence, an AB/BA design would be inadequate. Clinical trials are generally not capable of identifying the “benefit of the treatment to individual patients”

The simplest explanation of the Alzheimer's trials to which the document refers is that all patients benefit by a disappointing amount.
Finally, it also seems that the focus in this section is shifted to the choice of how to express the end-point (dichotomous Vs continues values) rather than on the multiple end-points issue per se. It goes without saying that from a pragmatic standpoint the two approaches are completely different.

6. Section 6
Although many of the considerations in section six seem reasonable, their practicality is difficult to quantify. There are also some philosophical weaknesses. For example, it is stated (in 6.3), "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". However, it is difficult to see how the regulator would be worse off in this respect than in a situation in which one of the positive component elements of the score had been designated as a single primary outcome, a situation the guideline appears to regard as unproblematic. With the composite score at least, the fact that one of the components is negative requires the others to be even more strongly positive to compensate. The litmus test should be whether a total composite score is relevant, in which case the total result should be accepted.

Of course, using a significant effect on a composite score as the single basis for the claim that the treatment benefits each element of it is false, and the discussion in 6.4 is very clear and sensible on this point.

Points of formatting

The format seems unnecessarily clumsy particularly the long titles of sections and sub-sections. Below we list the current and a suggested title for several sections:

2.1
Existing: The need for adjustment when there are multiple primary variables
Proposed: Multiple primary variables – when no adjustment is needed

2.1.1
Existing: Two or more *variables* primary variables are needed to describe clinically relevant treatment benefits (note variables appears twice)
Proposed: Two or more primary variables are needed to describe clinically relevant treatment benefits

2.1.2
Existing: Two or more primary variables with prioritisation
Proposed: Two or more primary variables ranked according to clinical relevance

2.2
Existing: Analysis sets
Proposed: Keep the same title, but make it 2.1.3

2.3
Existing: Alternative statistical methods

Proposed: Following the suggestion above this would become 2.2 and we suggest to add to the title " – multiplicity concerns

[This way 2.1 discusses situations where no adjustment is necessary whilst 2.2 addresses situations where there is concern about adjustment and thus more clearly fitting the overall title of section 2]

3

Existing: How to interpret significance with respect to multiple secondary variables and when can a claim be based on one of these?

Proposed: Multiple secondary variables and claims

3.1

Existing: Variables yielding supportive evidence

Proposed: Variables expressing supportive evidence

3.2

Existing: Secondary variables which may become the basis for additional claims

Proposed Variables expressing potential for additional claims

3.3

Existing: Indicative variables

Proposed: Variables indicative of clinical benefit

4

Existing: When can reliable conclusions be drawn from a subgroup analysis, and when is it appropriate for CPMP to restrict the *licence* to subgroup? [note the spelling of license]

Proposed: Subgroup analysis, reliable conclusions and regulatory claims

5

Existing: How should one interpret the analysis of "reponders" in conjuncion with the raw variables?

Proposed: Interpretation of a "responder" analysis

6.

Existing: How should composite variables be handled statistically with respect to regulatory claims

Proposed: Composite variables and regulatory claims

6.1

Existing: The composite endpoint is the primary endpoint [note inconsistent use of 'composite endpoint' in connection with a composite variable]

Proposed: A composite variable as the primary endpoint

6.2

Existing: Treatment should be expected to affect all components beneficially

Proposed: Components of composite variables not affected in similar ways by treatment

6.3

Existing: The clinically more important components should at least not be affected negatively
Proposed: Clinically important components not affected negatively [and make it 6.2.1]

6.4
Existing: 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
Proposed: Regulatory claims and the effect of treatment on a component [and make it 6.2.2]

The need for clarification

i) Section 2 (line 12): ".....confidence intervals are of paramount importance but are not available for many of the more complex multiple-level-α-tests aiming at controlling the type I error". It seems somewhat loose to say that a 'confidence interval is not available for a test'

ii) Section 2.1 (line 7): ".....In this situation.....". It is unclear if "this" goes on the entire previous sentence or just on the second part of the sentence

iii) Section 2.1.2 (lines 2-3): ".....convincing results in others would clearly add to the value of the treatment". Suggestion: 'convincing results related to others would support the benefit of the treatment'.

iv) Section 3.2 (line 12): *"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 claims"*. The meaning here of the term "claim" is not clear. Is the meaning 'independent claim' or 'within the framework of the disease under study'?

Typographical changes to improve the text

a) Section 2 "Adjustment of multiplicity....." to 'Adjustment **for** multiplicity'
b) Section 3.3 "Variables that are indicative for a major clinical benefit....." to 'Variables that are indicative **of** a major clinical benefit'
c) Section 6 Ensure consistent use of "composite variable"