Comments on CPMP Points to Consider on the Choice of Non-inferiority Margin (CPMP/EWP/2158/99, Draft, 26 February 2004) German Region of the International Biometric Society

General comments:

1. Approach of selecting delta as a proportion of difference between active comparator and placebo (p. 2 Introduction 2nd paragraph and p. 8 IV. 4th paragraph)

The PtC document strongly discourages choosing the margin as a percentage of the expected difference between active comparator and placebo. This is in contrast to the FDA that appears to favour the idea of percentage retention for the margin specification. It might cause problems for a global development, if the x% retained effect as usually accepted by the FDA would not be of help here as well. Therefore, the statement in the PtC document should be weakened by pointing out that specifying a margin based on x% retention is not considered acceptable as the only justification for the choice but one still needs to provide evidence why an x% retention of the active control effect is clinically meaningful based on the historical data.

2. Use of two-sided 95% CI and two-sided testing at level 5% in non-inferiority trials (p. 3 I.1)

Insisting on two-sided 95% confidence intervals (and analogously two-sided testing at 5% level) in non-inferiority trials should be reconsidered.

- In the context of non-inferiority testing, insistence on using confidence bounds of one-sided level 1-α/2 rather than 1-α, or equivalently, on performing one-sided tests at level α/2 is as arbitrary and dogmatic as for superiority trials involving comparisons with a negative control. In fact, one-sidedness of the null and the alternative hypothesis between which a statistical decision has to be made is the crucial feature distinguishing non-inferiority trials from studies aiming at establishing equivalence in the strict sense. As stated on page 4, line 16ff of the PtC document, if the non-inferiority margin is chosen appropriately non-inferiority is demonstrated if the confidence interval lies entirely between the margin and zero. Hence, the upper boundary of the confidence interval appears to play no role for decision making. Insisting on the application of tests for non-inferiority at level $\alpha/2$ would therefore only make sense if post-hoc redefinition of the treatment initially chosen as the active control to the experimental treatment would be an option. But this is clearly an unrealistic assumption, due to the fact that it is hardly in anybody's interest to establish non-inferiority of a standard therapy which has been in widespread use for many years, to an experimental treatment not yet approved to the market.

- Placing the rejection region in one-sided testing entirely on one side of the distribution reflects the fact that one-sided hypotheses have been derived and thus the underlying scientific theory has permitted more specific predictions than in the two-sided case. It is perfectly appropriate that more specific predictions result in a gain in power. Such a gain in power is highly welcome from an ethical as well as from an economic perspective as fewer patients are needed to demonstrate an effect with the same type II error probability.
- By using asymmetrical two-sided confidence intervals the risk of false positive decisions on non-inferiority can be considerably larger than $\alpha/2$. If it is the aim of the guideline to reduce the risk of false positive results in one-sided tests from α to $\alpha/2$, this should be stated explicitly.

3. Possibility of a non-inferiority trial without inclusion of a placebo arm (p. 4 II.)

It is stated that non-inferiority trials without inclusion of a placebo arm are not possible when "established effective agents do not consistently demonstrate superiority in placebo controlled trials (e.g. depression, allergic rhinitis)." It should be discussed what to do in indications such as severe depression where this condition holds but inclusion of a placebo arm is not possible due to ethical reasons. Will regulatory advice be possible in such situations?

4. Role of putative placebo comparison and definition of delta using indirect CI approach (p. 6 III.1.3.)

- The role of the putative placebo comparison and the assessment of non-inferiority using the specified margin needs to be clarified. It is not clear whether or not it is sufficient to chose the non-inferiority margin based on historical data such that the proof of non-inferiority can be used as a vehicle to assure that the indirect confidence interval of T vs. P excludes zero. Or whether - in addition to the comparison of T vs. R based on the pre-specified non-inferiority margin - a putative placebo comparison is also expected as part of the primary analysis of a non-inferiority trial. Note that in this case a potential outcome of a non-inferiority trial would be that non-inferiority of T vs. R cannot be proven, however the resulting putative placebo comparison still results in an indirect confidence interval excluding zero! Which of the two approaches is the primary one? If both parts are expected, the issue of multiplicity, if any, should be addressed.
- Could it be made more clear how the indirect confidence interval can provide guidance for choosing the non-inferiority margin. It seems to contradict the statement made in the PtC that “the choice of margin should be independent of considerations of power” (p. 4 line 13).
- Interestingly, the terms “systematic review" and “meta-analysis” are avoided in this section with regard to how to get an (unbiased) estimate for the reference-placebo difference. Instead the non-well defined term "detailed literature search" is used. Why? Moreover, the problem of achieving different estimates by including different trials (e.g. depending on quality criteria) or by using different methods for combining the single study results (e.g. fixed or random effects model) is not mentioned here.

5. Extreme areas where it is difficult to justify any non-inferiority margin (p. 9ff V.)

- Restriction to test products that fall into category 2 (“small advantage”) is too restrictive in areas of life threatening diseases, e.g. cancer: Only for those products the proposed procedure (“superiority trial with relaxed significance level”) can achieve a power of more than 50%. Hence, there would be no more opportunity for the registration of products with equal efficacy or a very small efficacy disadvantage. However, this might also be worthwhile if they have other advantages as, e.g., prolonged progression-free survival. Without pointing to the alternatives mentioned in V.5, the current statement in this paragraph is contradictory to the statement in section V.5.
- Following the proposal of replacing non-inferiority trials by superiority trials with less confidence makes α to a quantity of feasibility. Instead, the decision criterion (“significance at level α”) should be kept constant for all types of trials and, if necessary, the value of the non-inferiority margin should be chosen such that the study is feasible. Consideration on feasibility of a study is per se not a negative attitude, in particular if discussed with regulatory bodies. In the planning stage, for any sample size calculated for a superiority trial at relaxed α there exists a corresponding non-inferiority margin that leads to the same power in a non-inferiority trial performed at the common value of α. Similarly, in the analysis a confidence interval with decreased coverage probability can be calculated and compared to zero or the assessment can be based on the 95% confidence interval – these considerations use the same data and why should the first approach give more evidence than the latter?
- In addition, the suggested approach means that a certain delta (i.e. the extra number of deaths which are accepted as well by this approach with regard to the conventional level α) is implicitly hidden in the relaxed α. Furthermore, delta will depend on the assumptions made for sample size considerations. Hence, the suggested approach does not yield a solution. In contrast, the interpretation becomes even more difficult and lacks transparency.

- The proposed approach would introduce a new type of superiority less stringent compared to what is currently meant by this term. How would a test treatment be labeled that is successful with respect to this criterion – superior or non-inferior?
- It should be clarified which hypotheses are to be tested in situations where it is advised to use two primary endpoints, one representing efficacy, the other safety (p. 10, V.5).

6. Choice of non-inferiority margin in studies with binary endpoints (p. 11 VI.)

The PtC document proposes to use a combined non-inferiority criterion where the margin is pre-specified as both a relative and an absolute difference. For the analysis, the margin should be chosen depending on the reference rate using the most conservative. However, alternative approaches exist. For example, a natural approach seems to be fixing the measurement scale (absolute, relative, odds ratio) and the relating test/method for calculation of the confidence interval, defining in the protocol non-inferiority margins depending on the observed reference rate, and selecting the margin to be applied in the analysis according to this rule. The definition of the non-inferiority margin over a range of overall rates should be based on medical and statistical considerations following principles as formulated, e.g., in Röhmel J (1998): Therapeutic equivalence investigations: Statistical considerations. Stat Med 17:1703-1714, and realised in, e.g., Röhmel J (2001): Statistical considerations of FDA and CPMP rules for the investigation of new anti-bacterial products. Stat Med 20:2561-2571, and Phillips KF (2003): A new test of non-inferiority for anti-infective trials. Stat Med 22:201-212.

Specific comments:

1. Showing superiority of test vs. placebo

Throughout the document, the term 'placebo' is used. Should this really always be placebo, or rather previous control? Suppose drug A has been proven to be better than placebo and then drug B has been proven to be better than drug A. If you now want to show that your new drug C is non-inferior to drug B, the document says you need to show additionally superiority to placebo. Should it be superiority to placebo or superiority to previous control A?

2. Use of the term ‘effect size’

The use of the term ‘effect size’ throughout the document has to be questioned. In the statistical literature, the ‘effect size’ is defined as the difference with respect to an outcome variable between two groups divided by the standard deviation of the outcome variable. In this document, the term ‘effect size’ seems to indicate the 'net' effect, i.e the difference between reference and placebo.

3. Relation to disease specific PtCs

This document should avoid contradictions to disease specific CPMP PtCs, e.g. the one on anti-infectives which specifies non-inferiority margins. It should be considered to make references to other CPMP documents.

4. Prescribing the use of confidence intervals (p.2 Introduction 2nd paragraph, 1st sentence)

There is a vast literature on statistical testing procedures tailored to problems in which the alternative hypothesis specifies that the population parameter of interest is less than its target value by some small tolerance margin Δ at most (characterizing equal effectiveness of the treatments under comparison). Typically, such a test provides considerably better power than the corresponding confidence interval inclusion rule. In view of this, the first sentence of paragraph 2 of the introduction should be less restrictive in prescribing the use of confidence intervals and should not preclude the use of fully efficient tests of non-inferiority hypotheses.

5. Conduct of three arm trials: T, R, P (p. 5 III.1.2.)

In the case of three arm trials some comments on multiplicity issues seem to be in place.

6. Surveying practitioners on the range of difference that they consider to be unimportant (p. 8 IV. 5th and 6th paragraph) or on the choice of significance level (p. 10 V.3)

In section IV at the end of the 5th paragraph (last 2 sentences) it is suggested that delta may be derived from a survey of practitioners on the range of difference that they consider to be unimportant. In the context of extreme areas where it is difficult to justify any non-inferiority margin, it is suggested to probably use a less stringent significance level and to justify its choice on a survey of practitioners. Both suggestions seem to be problematic. Many clinicians may not have a good idea of what the scoring systems used in clinical trials mean nor of what differences in clinical response rates are clinically important. It also raises the question of representativeness and validity of such surveys. The interpretation of the size of significance level seems to be even more difficult for practitioners.

7. Situations where differences between anticipated safety profiles exist (p. 8 IV. 6th paragraph

Small losses in efficacy might also be acceptable in exchange for other than safety benefits, such as improved regimen, more convenient route of administration, etc.

8. Specific comments on wording

- Page 4, 2nd paragraph, 2nd sentence:

  '…assurance that a test drug has a clinically relevant effect (relative to placebo) greater than zero…' instead of '… assurance that a test drug has a clinically relevant effect greater than zero…'

- Page 3, last paragraph of introduction:

  One should add to the last paragraph that efficacy parameters are used to illustrate the methods mentioned in this document. But non-inferiority margins can be defined for safety parameters as well.

- Page 10, V.3. 4th sentence:

  The term 'level of significance' could be misleading in this connection. If 'level of significance' is used here in the sense of level α, and if this is related to the suggestions from V.2, then it should rather read 'If an increased level of significance is used …'.

- Page 10, VI. 1st paragraph, last sentence:

  Change the term 'company' to a more general term considering clinical trials outside the pharmaceutical industry.

- Page 11, VII. last paragraph:

  Explain MAA.