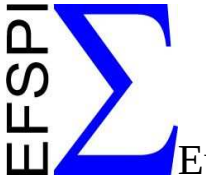

European Federation of Statisticians in the Pharmaceutical Industry (EFSPI)

COMMENTS ON DRAFT FDA "Guidance for Industry - Non-Inferiority Clinical Trials"

Rapporteur: Bernhard Huitfeldt (bernhard.huitfeldt@astrazeneca.com)

GENERAL COMMENTS

Single standard for proving efficacy (Test > Placebo) in a direct superiority (Test vs. Placebo) and an indirect superiority (Test vs. Control) study.

We are concerned about the recommendation in the guidance that suggests different standards to prove efficacy depending on whether a placebo controlled trial or an active controlled study is used. The use of an active controlled trial in combination with historical data, in order to indirectly demonstrate superiority to placebo, should not be used as a basis to require an arbitrarily higher standard for proving efficacy. In this case by introducing an additional fixed margin M2, which is addressing the relative efficacy of the test drug in relation to the control. This important principle has not been applied in the draft guidance. Arguments for a single standard of evidence for deciding whether a pharmaceutical treatment has demonstrated sufficient efficacy have also been presented in the PhRMA PISC Expert Team White Paper quite recently (BASS XV, Nov 3, 2008).

The true objective of an active controlled efficacy trial is to show that the drug is efficacious, i.e. would have beaten placebo if a placebo controlled trial could have been conducted. We acknowledge that there are weaknesses caused by the indirect comparisons, i.e. the assumptions of assay sensitivity and constancy. Thus, imposing some degree of conservativeness may well be motivated, e.g. through some kind of discounting, for example in the determination of M1. However, the primary purpose should still be to establish efficacy over placebo. The examination of relative efficacy for the new drug versus the control should not be an integrated part of analyzing the primary objective to prove the test drug > placebo.

We suggest therefore that it should suffice to require meeting the margin (M1) in order to demonstrate effectiveness of a test drug. This would be in line with the usual requirements in placebo-controlled trials of demonstrating that the effect is > 0 and it would eliminate the uncomfortable need for the subjective and probably in many cases not well-understood decision on a fraction of effect of an active control to be preserved (M2).

Fraction of effect to be preserved

The implementation of the requirement of a fraction of effect to be preserved in the form of a fixed margin M2 has two problems as it is described in the draft guidance.

First, the requirement seems to be based on the trial design (i.e. a non-inferiority design) rather than, more appropriately, on the existence of an effective therapy for the condition being studied. The document encourages the use of placebo-controlled trials to demonstrate a treatment effect when ethically feasible. Given this, it does not seem logical to require a certain fraction of effect to be preserved only when a non-inferiority design is chosen, but not when a placebo-controlled design is chosen. Our opinion is that the existence of effective therapy, not the trial design, should determine whether preservation of effect is required. See also above about a single standard!

The second problem regarding the implementation of effect to be preserved is that it bases the conclusions regarding a clinically meaningful effect on the lower end of the confidence bound for effect to be preserved, rather than on the point estimate. This is inconsistent with the customary approach for judging whether a treatment effect is clinically meaningful, and also may lead to serious logical inconsistencies in approval decisions. It may even prevent truly superior new drugs (or truly effective drugs with improved safety) to be approved. This has also been convincingly demonstrated in the above mentioned PhRMA PISC Expert Team report. A simple example follows. Assume that the margin M2 has been predefined and that the 95% CI for active/control (e.g. a hazard ratio) is confined between unity and M2, then non-inferiority would be concluded. Had the same study, with identical results, addressed superiority, the new drug would have been concluded to be inferior. I.e., identical results could imply conclusions of both inferiority and non-inferiority.

The EMA "Guideline on the choice of non-inferiority margin" (Jan 2006) states that a non-inferiority margin as a percentage of the active vs. placebo difference is deemed inappropriate not only for a study where relative efficacy is the primary purpose (sec 4, para 3 and 4), but also for studies where the purpose is to indirectly prove a new drug is superior to placebo (sec 2, 5th bullet). Thus, this guidance differs from the EMA guidance on the derivation of M2. Is there a plan for harmonization of review standards for international trials?

Statement of the primary hypothesis

One fundamental issue with this document has to do with the statement of the primary hypothesis to be tested, and the associated type 1 error rate. As described in this document, the margin M_1 is chosen in order to ensure that, by ruling out a difference between the treatment and control of M_1 or greater, one can conclude that the treatment has an effect greater than zero (i.e. is superior to a placebo). In other words, M_1 is simply a nuisance parameter of no direct interest, and the primary goal of the non-inferiority study is to demonstrate that the treatment has some effect. However, the document states the hypothesis to be tested, and the associated type 1 error rate, with respect to the difference between treatments of M_1 or greater. Since the goal is to demonstrate that the treatment has an effect, the most appropriate null hypothesis should be that the treatment has no effect, and the type 1 error rate of interest should be the probability of declaring an ineffective treatment to be effective. This probability can be controlled through appropriate choice of M_1, or through other statistical approaches such as the synthesis method. Stating the hypothesis in terms of M_1, which is simply a nuisance parameter, adds confusion throughout the document.

Synthesis method vs. fixed margin approach

The document is inconsistent regarding the fixed-margin approach vs. the synthesis method. In some places the document correctly refers to the synthesis method as an alternative to the fixed-margin approach that differs only in the way the variance terms from the historical data and the non-inferiority trial are pooled. That is, synthesis method is simply a more efficient alternative to the fixed-margin approach. Sometimes it refers to the inefficiency of the fixed-margin approach as a form of discounting that provides some additional assurance in the presence of concern over the constancy assumption. However, in other places it incorrectly refers to fundamental problems with the synthesis method that are not shared by the fixed-margin approach. In fact, there are no such problems; in some cases this would become much clearer if the problem with regard to the statement of the null hypothesis were fixed.

There are incorrect assertions that the fixed-margin approach is not affected by the constancy assumption. There are incorrect statements regarding advantages with respect to sample size calculation and the need for clinical judgment regarding efficacy preserved. Finally the document rejects the use of the synthesis method for M_1, but allows it for M_2. This is confusing. If the method is valid why shouldn't it be allowed for M_1? And if it's not then why is it allowed for M_2?

The "95-95" method may be justified as a conservative approach to show efficacy over placebo. However, this method suffers from several drawbacks. It is a fixed margin approach which does not treat the historical estimate of efficacy of the active control over placebo as a random variable. Further, since the method is based on the most unfavourable end of a 95% CI, it will lead to an overly conservative outcome. The synthesis approach, accounting for the variability in both the current and the historical study(ies), should be recommended.

Inconsistencies and repetitions

The guidance document provides useful and comprehensive guidance and describes a lot of situations, but as a consequence it is quite extensive. The general impression is that the document is overly wordy and unnecessarily complicated. The suggestion is to make it more condense by primarily removing repetitions. For example M2 is defined in several places (line 282, 292 and again in line 565 and 678).

It is suggested that also cross-referencing is used between sections to reduce repetitions. For example when discussing the implications of comparator response different from what anticipated in the sample size section (lines 1341-1342), the reader should also be directed to ponder the implications on the constancy assumption from section 4 (The Non-Inferiority Margin), 5 (Assay Sensitivity and Choosing M1), IV CHOOSING THE NON-INFERIORITY MARGIN AND ANALYZING THE RESULTS OF AN NI TRIAL, etc.

The document appropriately defines the concept of assay sensitivity, but does not apply that definition consistently.

See lines 79-80, 140, 334, 347

- "could have distinguished an effective from an ineffective drug"
- "control drug had at least the effect it was expected to have"
- "ability of the trial to have detected a difference between treatments of a specified size, M1 (the entire assumed treatment effect of the active control in the NI trial), if such a difference were present"
- "The active control would have had an effect of at least M1."

Whether or not the active control had its expected effect has nothing to do with the ability of the NI study to distinguish an effective treatment from an ineffective treatment.

There is inappropriate use of the word "effectiveness" throughout the document when NI designs are assessing efficacy. Effectiveness has been defined by many associations outside of the regulatory setting. Effectiveness is considered the extent to which an intervention does more good than harm when provided under usual circumstances of health care practice. Efficacy relates to clinical trials as the extent to which an intervention has a positive effect on the disease under ideal circumstances.

SPECIFIC COMMENTS ON TEXT

GUIDELINE SECTION TITLE: I. INTRODUCTION

Line Number	Comment and Rationale	Proposed change (if applicable)
	No comments!	

GUIDELINE SECTION TITLE: II. BACKGROUND		
Line Number	**Comment and Rationale**	**Proposed change (if applicable)**
33	The extensive list of references (page 59 and further) is very useful. Although it is mentioned that the references are not included in the text, it is considered as very helpful if the applicable reference(s) would be referred to in the text.	Suggestion to add the references in the text or at least at the end of the section/chapter.

GUIDELINE SECTION TITLE: III. GENERAL CONSIDERATIONS OF NON-INFERIORITY STUDIES: REGULATORY, STUDY DESIGN, SCIENTIFIC, AND STATISTICAL ISSUES		
Line Number	**Comment and Rationale**	**Proposed change (if applicable)**
114 and 144	Figure 1 versus Figure 2. It is confusing that the direction of effect is switched	Reverse direction on Figure 2.
133	In order for a non-inferiority trial to demonstrate effectiveness, it should have the same null hypothesis as that on line 107, i.e. superiority to placebo. That is how "effectiveness" is defined.	Recommendation - Further expand line 133 to refer to superiority to placebo
134-137	The hypotheses are stated as one-sided and the confidence interval is two-sided which is confusing and theoretically incorrect. In fact the one-sidedness of the hypotheses distinguishes non-inferiority from an equivalence setting. (See also line 832)	Suggestion to discuss the distinction between the hypotheses of non-inferiority and using a two-sided CI rather than a one-sided CI.
143	Examples here use M1 as the entire effect size of the control. The document is organized by separating M1 and M2, so including this example fits the logic of the document. However, the conclusion that "NI is demonstrated" for example 1 is inconsistent with the document's logic, because demonstration of NI requires that both M1 and M2 be met.	Recommendation – ensure M1 and M2 are both referred to for demonstrating NI

158-161	We understand such situation will present interpretive problems. However, the effect of control is still within pre-specified non-inferiority margin (M1 or M2). Should such a result still present interpretive problems if M1 is agreed?	Recommended change: A clarification of this point will be helpful. Are there any considerations in the choice of NI margin that are important for decreasing the chance of such a paradoxical outcome?
167	"Determining the NI margin is the single greatest challenge in the design, conduct, and interpretation of NI trials". We strongly disagree, the NI margin is not even needed, see general comments above, including the recommended use of the synthesis method.	The greatest challenges relate to estimating the efficacy advantage for the control over placebo, and in addressing the important assumptions of assay sensitivity and constancy, which are required due to the reliance on indirect comparisons.
173 and 332	Inconsistent use of headers.	It would be helpful to extent the table of contents, in particular for Chapter III. Also to use the same type of sub-headers. For example on page 6, under section 3 the header is 'a. Comparative effectiveness' and on page 10 under section 5 the header starts with a bullet.
201-212	The real world clinical "value" of any drug is determined primarily by consideration of both the efficacy and safety attributes of the test drug compared to the control drug. In more recent times, a third attribute of drug value (independent of safety and efficacy), the potential for improved compliance, has also become increasingly important in stakeholder assessment of the potential benefit to risk ratio of therapies for patients. The value of a drug B could be considered higher than that of a comparator drug A if a modest decrease in efficacy within a certain margin is shown but clear superiority is shown on safety and/or compliance attributes.	
201-212	There are situations where a non-inferiority trial can (and does) include placebo. These situations may include cases when control vs. placebo trials are non-existent and it is not unethical to treat patients for a short period with placebo. Inclusion of the placebo may allow a within-study assessment of assay sensitivity as a first step in non-inferiority.	We suggest adding a short discussion on NI in this situation when examining whether assay sensitivity may be derived wholly within the single current study (which may be unknown at the study start).
229-238	Agency's recommendation when constancy assumption is violated is vague.	Include more clear recommendation
260	Since this is the legitimate concern, why isn't it the null hypothesis to be	Consider if a note on controlling error rates should be added

	rejected? Should we control the probability of making this error?	
267	What is meant with 'the distribution of estimates'? The probability distribution of estimates other than the point estimate and the 95% upper bound?	Suggestion to explain what is meant with the term or to rephrase the sentence.
277	"Showing non-inferiority to M1 would provide assurance that the test drug had an effect greater than zero". OK, conditional on presence of assay sensitivity and constancy.	Add dependency on the assumptions of assay sensitivity and constancy.
282	"(M2) that reflects the largest loss of effect that would be clinically acceptable". How can the loss of, e.g., half the efficacy of the control group be "non-inferior", when we are designing superiority studies based on smaller differences between drugs.	Focus of the Guidance need to be changed, see general comments above.
286	This sentence "Note that the clinically ... secondary endpoint)" does not make sense, this would result in serious logical inconsistencies in demonstrating efficacy in many applications	Revise.
292	There are situations where any clinically acceptable margin is ethically difficult to justify, for example where the endpoint is 'death'. What would M2 be in this case; would it be '0'? (See also line 658)	Suggestion to provide in the document guidance in case of extreme situations where it is difficult to justify any non-inferiority margin.
301-302	Confusing statement. Should M1 be set after unblinding? How can M2 then be pre-specified?	Delete the statement
304	"M2 is a matter of clinical judgment". Hence, different companies will arrive at different M2's since different Advisors are being used, etc.	M2 is not meaningful, see general comments.
338	"of at least M1", why?, the point estimate should be M1, half the studies would be expected to have a smaller effect size, and half a larger, in the presence of assay sensitivity.	Revise.
338	This is not a correct alternative way to state the concept. A placebo control would have been subject to random error. The definition of assay sensitivity should be based on the "true" control effect size, not what might have been observed in a clinical trial.	Recommendation – note the true control effect size
342	Sentence starting "Even if the NI margin ...". Not necessarily true, this	Delete sentence, the degree of lack of constancy

	depends on how much less than M1 the effect of the active control was.	cannot be objectively assessed.
349	"(2) the similarity of the new NI trial to the historical trials (the constancy assumption)". This relates to assay sensitivity at least as much as constancy.	Revise.
353-363	Usually we do literature search for estimating the difference between active control and placebo for a medical condition in the intended population. There could be publication bias where studies with positive outcome are more likely to be published than those with unfavourable outcomes.	Include the agency recommendation in this kind of situation?
353 and 1069	Chapter with HESDE does not mention potential publication bias involved in historical evidence. Any literature search on historical effect sizes may be biased because the failed studies are not publicly displaced. Even the proposed 95% CI for the historical effect may be too optimistic. Concept of discounting could discuss this potential source of bias.	
378	As for the comment on line 349. The constancy assumption addresses whether the effect size of the control over placebo is still the same today as it was in the historical studies. This will not be guaranteed by using identical study designs. Changes in medical care, such as concomitant medications, etc., may also be of importance.	Revise.
405	In the case of selecting a new endpoint (not a composite endpoint) for an NI study, is it necessary to show that the new endpoint is a surrogate of the endpoint used in the historical trials, or can it be based on clinical judgment? How is the constancy assumption taken into consideration?	Recommendation – add further details of handling the use of different endpoints relative to historical endpoints in designing NI studies
442	"are not conservative in an NI study". ITT may still be conservative, e.g., if the new drug is truly more effective than the control, but not enough for superiority to be shown with reasonable sample sizes.	Add "necessarily" before "conservative".
447	The word 'perverse' is used in a strange way.	Use another word
472	It is stated that similarity could be concluded if C-T is close to showing superiority. This is counterintuitive, superiority and similarity can never	Revise.

	be concluded simultaneously.	
535-537	Two different and reasonable concerns are being mentioned here, but the position that historical effect size is unreliable does not lead to the argument about the need for more than one NI study. The second one is about protecting the Type I error from a single NI trial.	Recommended change: Remove the paragraph. The argument is made more clearly in the next paragraph on lines 539-543
574	"showing non-inferiority using M2 thus provides very strong evidence, analogous statistically to the 2 studies (at $p<0.05$) standard". Not at all obvious, still strongly dependent on constancy and assay sensitivity assumptions. What will be available is only one study which did not discriminate between two treatments.	Revise.
623-625	For the fixed margin method, it is indicated that the margin can be flexible (e.g., 90% CI, one-sided versus two-sided). It would be helpful to have guidance as to when to consider a less stringent margin.	Recommendation – consider adding further comments on appropriate use of margins
635	Since 1.28 is measured on a relative (multiplicative) scale, "half" of 1.28 would rather mean the square root of 1.28.	Change the calculation of half the relative risk of 1.28 to root(1.28) = 1.13

GUIDELINE SECTION TITLE: IV. CHOOSING THE NON-INFERIORITY MARGIN AND ANALYZING THE RESULTS OF AN NI TRIAL		
Line Number	**Comment and Rationale**	**Proposed change (if applicable)**
689	Should it be the size of the treatment effect compared to placebo instead of effect compared to no-treatment?	Recommended change: Suggest changing "no-treatment" to "Placebo".
725-727	With the texts around here, it sometimes appears that the (minimum) NI criterion is M1, and that M2 is an additional and less formal criterion.	Recommended change: Please re-word or clarify.
814-817	Considering the testing at an alpha of 0.001 for a single trial, if the expectation is to consider a 99.9% CI, please indicate.	Recommended change: Consider adding "(e.g., 99.9% instead of 95% CI) on line 815 to clarify.
928	While a qualitative heterogeneity is not desirable, "no evidence of statistical heterogeneity" seems unnecessary.	Recommended change: Consider adding appropriate test to clarify the intent.

962	The guidance document is mainly focussed on rates, like all four examples. Continuous endpoints (as used in QTc trials), counts and composite endpoints are not mentioned in this paragraph.	Suggestion to mention other types of endpoints in this paragraph as well.
1001-1015	Make it clear from the beginning that ½ on a relative scale implies the square root.	Revise.
1221	"And is only somewhat conservative". This is an understatement. In many applications the implications on sample size, relative to using the synthesis method, will be huge.	Revise.
1306-1308	Is determination of M2 as a fraction of M1 a well understood concept among clinicians? There is a concern that this choice may be somewhat arbitrary without realization of the full meaning of the concept.	Consider deleting the M2 requirement
1308	"The clinical judgment in determining M2 may take into account ... its impact on the practicality of sample sizes". This is in contradiction with scientific principles.	Delete.
1329	Mistake in % of effect ruled out. Ruling out a 48% loss is stronger than ruling out a 50% loss.	Correct "48%" to "52%"
1334-1352	Although the document recognizes that NI studies may or may not have event rates as endpoints, some sections are written exclusively using event rates as examples (see the section "Estimating the Sample Size for an NI Study").	A more general language (valid also for time to event or scores on continuous scales) would be welcome.
1337-1338	There is a contradiction between the recommendation to use the synthesis method for meeting M2 (lines 1128-1131, 1487-1488) and basing the design on meeting M2 using the more conservative Fixed Margin Method. Note that if M2=0.5*M1 this implies a quadrupling of the number of patients.	Sample size should be based on meeting M1. This relates to comments made elsewhere about omitting M2 from the requirements.
1383	Unclear why only re-estimation of sample size is addressed and not other options of adaptive designs like early stopping.	It would be helpful to describe other adaptive design features as well or to highlight the different aspects of adaptive design in NI studies compared to Superiority trials.
1383 -	The title suggests role of AD is to increase sample size. Are there	Recommended change: Clarify the intent.

1384	reasonable conditions under which the sample size may be reduced?	
1386-1389	This section briefly discusses adaptive techniques, mentioning exclusively sample size increases but nothing about other "sequential design" issues such as stopping for clear and adequate demonstration of efficacy or futility based on an interim analysis.	Recommended change: Please consider adding discussions clarifying the FDA's position.

GUIDELINE SECTION TITLE: V. COMMONLY ASKED QUESTIONS AND GENERAL GUIDANCE			
Line Number	**Comment and Rationale**	**Proposed change (if applicable)**	
	No comments!		

GUIDELINE SECTION TITLE: APPENDIX - EXAMPLES		
Line Number	**Comment and Rationale**	**Proposed change (if applicable)**
1674	Standard medical care is changing over the years.	Suggestion to add a column with the year of conduct of the study.
1683	"constancy of risk reduction"	Insert "relative" before "risk reduction" for clarity. Absolute risk reductions were not comparable.
1696 and 1697	"Race" is said both to be comparable, and not to be comparable.	Revise.
1706	"long duration"	Why should this matter, analyses are made on a per patient year basis. It is more common with decreasing event rates over time than the opposite.
1707	"including vascular deaths and non-fatal myocardial infarctions"	No, these events were not included in the meta-analysis. The 4% versus 13% are based on the

		correct primary endpoint. With those additional components, being part of the primary composite endpoint in EAFT, the event rates were 8% versus 17%.
1732	Delete “hazard”	The analyses were based on risk ratios not hazard ratios.
1830	In this example it is explicitly mentioned that the synthesis method is statistically more efficient than the fixed margin approach in most situations.	Suggestion to make this statement also in the main text, for example after line 1130 or in section 2 (line 1250).