European Medicines Agency

London, 27 July 2006
Doc. Ref. EMEA/CHMP/EWP/281490/2006

OVERVIEW OF COMMENTS RECEIVED ON DRAFT GUIDELINE "CLINICAL TRIALS IN SMALL POPULATIONS"

Table 1: Organisations that commented on the draft Guideline as released for consultation

	Organisation
1.	Protherics PLC
2.	EFPIA
3.	Emerging Biopharmaceutical Enterprises
4.	EORTC Data Centre, Brussels
5.	International Plasma Fractionation Association
6.	Medicines Evaluation Board, Netherlands
7.	Merck Sharp & Dohme
8.	National Cancer Centre, Singapore
9.	PSI (Statisticians in the Pharmaceutical Industry)
10.	Prof Eva Skovlund, NoMA
11.	Utrecht University, The Netherlands

General Overview

Comments were received from 11 organisations/individuals. Comments were quite mixed. Some were quite neutral in tone but with suggestions for improved wording, highlighting sections that were unclear, etc. Other comments were very positive and supportive of the initiative – and generally supportive of the specific suggestions. Others were much more critical – either of the general 'relaxing of regulatory standards' or of specific methods/approaches that have been suggested (or of both).

In the summary of comments given below does not contain every small detailed point but tries to include a balance of the major comments received. Hence this includes the positive comments (which may require no change to the document) as well as the more critical comments. Some opinions were sent in a similar form by more than one commentator; these are only listed once.

The final version of the document is much shorter than the original because where there has been substantial difference of opinion, issues and ideas have been deleted from the document, rather than present one non-consensual view. In particular, the Appendix (consisting of examples and references has been deleted).

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

Table 2: Discussion of comments

Section	Comment	Response
GENERAL COMMENTS		
	the term "*Small populations*" is not well defined in the proposed guideline	No definition has been given – but it is stressed that deviation from other, existing guidance will always need careful justification and sponsors are encouraged to seek scientific advice or protocol assistance, as appropriate
	It should be clarified whether this guideline also applies to medical conditions where there are very few patients in Europe but many elsewhere such as tropical diseases. It would help understanding whether this guideline may apply in those conditions where large scale conventional and multi-centre clinical trials are not practical or applicable.	No change, but see above comment.
	Two general areas that are not addressed are: 1. instances where the indication with a small population is potentially a second indication for a product (eg aplastic anaemia for a haematopoetic growth factor) and prior knowledge can be leveraged for this follow-on indication 2. the application and use of clinical trial simulation (where feasible) to substantiate the clinical claims.	Both of these points might form part of a sponsor's justification (see above)
	It is recommended to state more clearly that this document deals with an exceptional case were two conditions have to be fulfilled: - The condition evaluated should meet the orphan status - It is impossible to perform a conventional randomised controlled trial	Comment included
	The document focuses almost exclusively on efficacy; safety should be addressed in a specific paragraph although it is recognized that small samples will provide limited safety information. A discussion of the level of evidence, and timing of such evidence, for safety evaluation of a new drug under investigation in a small population would be helpful.	Comment added that additional data (post approval) may be required (see, also, point below)

Section	Comment	Response
	…it should be stated that follow-up data for efficacy and safety as part of a post marketing commitments could be required. Especially in an extreme small number of subjects, when a surrogate parameter is used, the maintenance of effect, efficacy on the more relevant clinical outcome and safety are largely unknown. In those situations long term follow data of good quality is required.	This comment added
	The document does not provide any real guidelines, but rather lists different methods and techniques that could be applied in small studies. As such, this document may be dangerous because it appears to encourage what in some cases may be inappropriate or questionable statistical methodology. There are some sections of the document that are felt to be unacceptable. Other sections discuss methodology that could reasonably be employed in the case of a small study but which require further clarification.	Agreed, there is little 'guideline'. Sponsors can choose their own approach and ideas/examples are given. Any approach will need to be justified
	Methodological and statistical aspects are in general very well covered by the ICH E9, and I do not see any advantage of producing a guideline which to some extent seems to oppose the design and analysis strategies recommended in E9. Furthermore, deviations from ICH E9 have always been acceptable in specific situations when appropriately justified. Thus, it is not obvious that specific guidance on small populations is actually needed or even advisable. At best the guideline could be regarded as superfluous, but as it now reads, I fear some of the statements in the document are actually dangerous to put in writing and will produce problems for drug regulation rather than help solving them. To me it is very surprising to see a guideline stating for instance that anecdotal case reports could be seen as providing valuable information with regard to granting a marketing authorisation.	Much of the statistical/methodological section has been deleted.
	[We] regret this guidance and strongly suggest reconsidering its need. [We] unanimously agreed that there is no need for this guidance. The guidance is considered superfluous and, above all, dangerous. Superfluous, as all topics discussed are covered in ICH9 and deviations from ICH9 always will be acceptable if justified. Dangerous as the methodological and statistical principles formulated in ICH9 adopted in 1998 and which improved the quality of applications since, are relaxed in case a small population can be defined. It appears that many justifications for insufficient research can be found in the document.	This version has is less emphasis on relaxing principles set out in ICH E9 and more on the need for sponsors to justify any relaxing or deviating to that (and/or other) guideline(s)
	In general, the document should be more cautious and state more clearly the risks and benefits associated with each of the methods that are described.	This version is more cautious than the previous version

Section	Comment	Response
	In orphan drug applications the data submitted often are of poor quality. It might be more emphasized that at least data of good quality are required.	Appropriate comment added
	The following area has not been fully addressed: - logistical problems in conducting trials in small populations, e.g. the fact that for small populations the patients may be spread over several centres, so that each clinical centre may have only 1 or a few patients.	Comment on such practical problems has been added
	We believe that scientific advice and/or protocol assessment (sic) should be recommended for studies in small populations. [We] would, therefore, suggest that the Guideline clearly and strongly recommend discussions with the EMEA about the development plan and the pivotal studies' protocols.	This recommendation is added
	It is noted that some alternatives discussed further in the document e.g. use of surrogate variables, Bayesian analyses, sequential designs do not increase the efficiency of the design and analyses like stratification, block randomisation or matching. Instead elements are introduced that may decrease the level of validity of the study results (e.g. data driven design modifications, uncertainty of the relation between surrogate variable and clinical outcome, introducing a historical comparison). The discussion could have been focussed more on the trade off between a larger uncertainty of the validity of the results as compared to a RCT weighted against the risk of disapproving a truly effective agent and approving a truly ineffective agent in the context of the rare disease and safety data available.	Comment on trade-off between uncertainty and validity has been added
1 INTRODUCTION		
Page 3	Support given to statistical methods that may increase power or precision, including the use of stratification, covariates and statistical modelling	No change
Page 3, para 5	Not all decisions are uncertain and in some cases, evidence will be "beyond doubt". Therefore, [we] recommend deleting paragraph 5 of the introduction on page 3.	Text deleted
Page 3, para 9	In both situations, (1) and (2), the statistical tests will be underpowered. The distinction lies in the severity of the lack of power. In the second situation, the power is so low that a statistical test will be of very limited value. It is suggested to present the distinction based on the severity of the lack of power.	Text revised to better reflect this point

Section	Comment	Response
Page 3, para 9	A third situation should be addressed: “A therapeutic effect in an uncontrolled clinical trial may be so dramatic that there is no reasonable doubt that a randomised controlled trial would give the same result. This is more frequently the case in small populations than in others.	No change. The benefits of concurrent controls is placed very highly
Page 3, para 9	The document addresses methods to increase the efficiency of design and analysis but does not give guidance to address situations where such methods are not applicable.	The only advice we can give is to seek scientific advice/protocol assistance. This has been stressed
2 LEVELS OF EVIDENCE		
	pleased to see recognition within the guideline… that in very rare diseases, the combination of single case studies may be the only way to accumulate evidence	No change
whole section	Could retrospective analysis of data from patient files be considered in the case of a rare disease, for example, where a prospective clinical trial may be impractical? If so, please indicate points to consider in such circumstances.	The emphasis remains on concurrent 9and randomised) controls, wherever possible
Page 4, 1st main para	The wording "effect size highly statistically significant" should be replaced by "effect size highly clinically relevant and statistically significant"	Text changed
Page 4, last para	Last sentence of the last paragraph: This paragraph is misleading because it is mixing together the meta-analysis of phase 3 trials and the pooling together of results from non randomized studies. It states that meta-analyses of individual case reports or of observational studies should be considered. It is misleading to call this a meta-analysis and may give the readers a false sense of security in the results. Overview might be a better term to use.	Text changed (see below)

Section	Comment	Response
Page 4, last para	It is advised to change this paragraph as follows: "In very rare diseases, the combination of single case studies may be the only way to accumulate evidence. In such situations, treatment regimens and data collection ~~may~~ **should** still be carried out in a ~~controlled~~ **standardised** manner. **A systematic review** ~~and this~~ will add weight to the evidence. ~~Furthermore, if careful consideration is given to the statistical analysis, (including methods such as formal 'cumulative meta-analyses' of randomised controlled trials) then this will carry more strength than ad hoc pooling of several case reports~~ Meta-analyses of individual case reports or of observational studies should be considered."	Text changed as suggested
Page 5, para 1	The first sentence states that "Generally, a larger sample size and/or smaller variance will result in narrower confidence intervals and more extreme level of statistical significance". This is not a sufficient condition. The level of statistical significance depends on the treatment effect too. In this paragraph, the guideline only focuses on the concern on statistical significance level and not in terms of clinically relevant effect and power to obtain this effect. The guideline should discuss difficulties to a powered trial with a relevant clinical effect when only a small sample size is applicable. Then a less extreme significance level should be considered.	Text clarified and emphasis placed on the need for estimated of effect sizes
Page 5, para 1	The proposed flexibility in interpreting observed treatment effects is welcomed. It is encouraging to note that the CHMP emphasizes the use of confidence intervals in this context and explicitly states that 'No [P-value] such value is adequate to confirm that a treatment effect truly does exist'. A firm and unambiguous statement on the use of confidence intervals instead of P-values would be beneficial in providing clear guidance on this issue. Otherwise it would be very helpful to have some specific guidance with regards to which level of statistical significance might be acceptable in which circumstances. Should this be a topic for discussion with the SAWG/COMP via Scientific Advice/ Protocol Assistance it might be useful to indicate it in the guideline.	No specific advice on acceptable *P*-values can be given. It will always be a case-by-case decision – and the size of the treatment effect is likely to be at least as important as the *P*-value
Page 5, para 1	The discussion of the p-values is very welcome. [We] feel that in addition to this, more guidance on dealing with outliers would be necessary, as the impact of an outlier in a small study is quite significant.	No specific changes – but note comments on use of non-parametric methods in Section 6.2

Section	Comment	Response
3 PHARMACOLOGICAL CONSIDERATIONS		
	For a data package that is heavily reliant on preclinical pharmacology data, general guidance on how to show that this is relevant for man would also be welcomed	No specific guidance included. It will always be on a case-by-case basis (see, for example, next comment)
	Animal pharmacology studies may not always be predictable for the design of clinical trials. This is especially the case for oncology products.	Comment added that PK/PD data may not always be very reliable (or predictive) of clinical outcomes
Page 5, para 4	… the statement that "regulatory requirements for licensing substitution products may sometimes be less rigid than for other compounds" should be qualified [i.e. to say this applies particularly to recombinant products]	Text changed
page 5, para 6	Use of the term 'Black box designs' is assumed to refer to situations where limited or no scientific or pharmacological hypothesis or model is available. Use of such a term however is potentially confusing and ambiguous; we would therefore recommend clearer language is used in place of this term.	This term has been deleted
4 CHOICE OF ENDPOINTS		
	recognition that the most appropriate end-point for a study cannot always be pre-specified at the design stage	Agreed; point added
Page 6, para 2	This regards the sentence "If *quality of life* is measured, it should always be assessed using scales validated for the particular indication ….". For rare diseases, it's not easy to find a specific scale of the given disease and it's still more difficult to find a validated specific scale. So, it's an ideal to use "scales validated for the particular indication" but not realistic.	This section re-worded and clarified
Page 6, para 2	We would also consider QOL data as an appropriate stand alone endpoint in certain cases, as QOL can be considered strong evidence for clinical benefit.	This was not agreed to. QoL is likely only to be supportive
Page 6, para 2	It is proposed to change the last sentence "It may be one means of helping to place the product in context with other available treatments." into "It may be one means of assessing the impact of the observed effect for activities of daily life and social functioning"	Text has been modified to reflect this comment

Section	Comment	Response
Page 6, para 3	The results of all endpoints are typically presented in the study report. Their hierarchy helps deciding and concluding. We suggest adding a paragraph on multiplicity in the Data Analysis section. This paragraph could also mention methods for combining evidence from multiple endpoints (O'Brien and Pocock's methods [1,2]) to help assessing collectively the treatment effect.	Since the topic of multiplicity is covered in a separate Points to Consider document (CPMP/EWP/909/99), it is not discussed again here
Page 6, para 3	… the guideline would be strengthened if the dangers of multiple testing of endpoints and the need for appropriate adjustment for p-values (or confidence intervals) were made more explicit.	Text on evaluating multiple endpoints has been deleted. See, also, comment above
Page 6, para 4,5	In general terms, the guidance on the validation of surrogate endpoints and biomarkers is vague. The document would greatly benefit from further clarification/guidance on this topic, particularly where it concerns surrogate endpoints or biomarkers that may be difficult to validate. … The addition of further examples of different surrogate endpoints and how they can be used/justified and validated would be extremely valuable (comment also refers to fifth bullet in the 'Summary and Conclusions section). Furthermore, a distinction should be made between endpoints that are established as a surrogate for clinical benefit and those that are likely to predict a clinical benefit.	It was decided that the guidance should not contain further comment on methods for validating surrogate endpoints. This is such a large (and difficult) subject that it should not divert attention from the main issues to be covered The examples (Appendix) have been deleted
Page 6, para 5	Generally, once the potential effectiveness of a drug product has been demonstrated, especially in the case of an unmet medical need, it is difficult to perform further clinical trials in a timely fashion, particularly controlled studies, due to patient recruitment, and/or continue patient enrolment in the current trial(s). This issue is increased in the case of a small population. Please clarify how trials in small populations could be supplemented to provide evidence of clinical benefit, as well as to further assess safety and risk/benefit? What level and form of supplemental evidence could suffice in such circumstances?	This topic is likely to be covered in further documents relating to conditional approval
5 CHOICE OF CONTROL GROUPS		
	The word "placebo" or "placebo control" is not mentioned in the whole paragraph [section]. [We] would suggest rewriting the section in a way that starts with the "ideal" situation, i.e., a placebo-controlled study design and then discuss alternatives to placebo-controls when they are not feasible or not possible.	Placebo is now mentioned at the beginning of the paragraph

Section	Comment	Response
	There should be some mention of using standard of care as a randomised control arm	This was an important omission. Best standard of care is now included
Page 7, para 1	"note the recognition that randomized controlled studies cannot be performed for some agents, including those where patients require emergency treatment for a life-threatening condition and cannot ethically be entered onto a study in which they could be randomized to receive an ineffective control, instead of the test agent"	This comment was not agreed with. Even for acutely life threatening situations, randomisation is often still possible and should be strived for
page 7, para 3	The document recommends using strong predictor factors in the randomization procedure to achieve balance. The terminology of stratified randomization may be misleading since it refers to a randomization process that seeks to achieve balance in all combinations of the strata. In small populations, this approach is almost always impractical. In addition, covariate-adaptive randomization seeks to achieve marginal balance for the prognostic factors. The following wording is suggested to insist on marginal balance: 'If there are any strong prognostic factors for the outcome, then a ~~stratified~~ balanced randomisation procedure, combined with a suitably stratified/modelled analysis can greatly increase the efficiency of the trial. Similarly, such ~~stratification~~ balancing – across as many factors as possible – will usually increase credibility of the results by ensuring balance on these factors across the treatment groups. ~~Stratification~~ Balance for many factors in small studies becomes almost impossible unless dynamic/covariate-adaptive randomization schemes are used.'	Text changed to refer to balance instead of stratification
page 7, para 4	We welcome the recognition that in certain circumstances, it will be appropriate to use historical controls. Guidance and examples on when this may be appropriate is requested. This paragraph uses the terminology "exceptional circumstances" however it should be clarified by the CHMP whether it refers to the "exceptional circumstances" provided by the EU Directive 2001/83/EC amended by Directive 2004/27/EC under which marketing authorizations may be granted subject to a requirement for the applicant to meet certain conditions.	This phrase ("exceptional circumstances") was not intended in its regulatory meaning. Text has been changed to clarify this
6 METHODOLOGICAL AND STATISTICAL CONSIDERATIONS		
	An additional sub-section on medical methods should be included to describe options to improve precision and accuracy of clinical endpoints (e.g., by repeat measurements, by independent evaluation consultants, …)	No specific section has been added but comment is included about the importance of high quality, precise data

Section	Comment	Response
	While the guideline provides an extensive summary of possible statistical approaches to enrolment and analysis of clinical trials where small populations are involved, it is limited in its guidance on current thinking with respect to the applicability or acceptability of such approaches in providing evidence of safety and efficacy suitable for product registration	As the document states at the outset, all forms of evidence can contribute to making an overall benefit–risk assessment. It will be up to the sponsor to decide which forms of evidence are attainable and which forms of evidence will be sufficiently convincing. Scientific advice/protocol assistance should be sought to help with this
	It is … puzzling to see that some methodological approaches, not acceptable in large trials, may be considered acceptable for trials in small populations (why?), and that clinically relevant endpoints may be substituted by less relevant but more (statistically) efficient endpoints. To me it is less than clear how to assess benefit/risk if benefit cannot be estimated.	In general, any section of the document suggesting acceptability of methods that would not be acceptable in large trials has been deleted
	A paragraph on multiplicity handling, as a strong control of the family-wise type I error is not appropriate in this setting	The document stresses the value of estimates and confidence intervals over that of controlling Type I error
	In some circumstances (especially where there is little experience with the drug product and its pharmacology which may be poorly understood), it may be difficult to predict the size of the treatment effect. In the case of underestimation of the sample size, would it be acceptable to either combine the results from this study to another study, or other form of evidence, e.g., meta-analysis, in order to increase sampling power? Alternatively, would it be acceptable to continue/re-start recruitment in a given study, following the analysis of primary endpoint(s)? It is also recommended to also discuss with more details the use of interim data analysis and how this may be helpful in evaluating/validating at an early stage the design of a study and further continuation of a clinical development program in small populations.	Other guidance on meta-analysis (CPMP/EWP/2330/99) and multiplicity (CPMP/EWP/909/99) already exist and so this document does not overlap with those. Guidance in those documents will still be applicable to small populations.
page 8, para 2	The idea of using an efficient endpoint even if it is not the most clinically relevant can be misleading. A difference which is statistically significant is not necessarily medically significant. If the endpoint is not really clinically relevant, then why use it? What is its relevance as far as clinical decisions are concerned? One should rephrase the sentence to clarify its meaning. In this section one should discuss the possible role of surrogate endpoints, tumor markers that may give an indication of biological activity, consistency of results based on different a priori defined endpoints, etc. In addition, one could suggest that enough evidence from pre-clinical studies, laboratory studies or animal models are needed to justify the use of biological markers if any are being used.	The emphasis of this paragraph is been changed to confirm the importance of clinically relevant endpoints

Section	Comment	Response
page 8, para 3	The discussion on stratification factors is unclear. On page 7 it states that one should stratify for as many factors as possible and this can only be done using dynamic methods. However depending on the algorithm used, there is still a risk of "over stratification" in small studies even using dynamic methods, for example if center is one of the factors. Then on page 8, dynamic methods are criticized because they are not strictly random and conventional methods cannot be used for the analysis. CPMP/EWP/2863/99, "Points to Consider on Adjustment for Baseline Covariates", heavily criticizes the use of dynamic allocation methods. However the drawbacks of block randomization (stratification using blocks) are not adequately discussed. Thus the reader is left wondering whether or not dynamic allocations methods can and should actually be used in small studies. We also believe that in small trials, and independent of the method used, that stratification for a very large number of factors could actually be worse than no stratification at all. The question of whether or not to stratify for center in small randomized studies should also be addressed.	This section has mostly been deleted. All reference to dynamic methods of treatment allocation have been deleted to avoid any contradiction with other guidance on use of baseline covariates (CPMP/EWP/2863/99)
Page 8, para 4	A problem is that there is seldom reliable information on which variables are of prognostic value. Perhaps it should be highlighted that information on the prognostic value of factors may be limited and unreliable in small populations and that selection of covariates should be based on quantitative rather than anecdotal evidence.	No mention is made on methods to select covariates. This is adequately covered in ICH E9
Page 8, para 5	"Covariate-adaptive methods" page 8: With respect to "…. conventional statistical methods cannot be used for data-analysis." the guideline should give an example of "non-conventional" statistical methods?	All reference to dynamic allocation/covariate-adaptive methods has been deleted
Page 8, para 5	The statement that conventional statistical methods cannot be used for data-analysis is too strong and is a topic of debate. Model based approaches may be appropriate. The following statement 'Further, if centre is used as a stratum and there are many centres but few patients per stratum this may lead to simple alternate allocation, thus jeopardising allocation concealment' may not be correct. When center is the only factor to balance upon, and Taves's version of minimization is used to allocate the patients, the sequence of treatment assignments within any given center will simply be a permuted blocks schedule with blocking factor of 2 (each odd allocation will be selected at random, followed by the allocation to the other group) - a widely accepted allocation procedure. If there are other factors in addition to center, blocks of sizes >2 will be also present.	As above, all reference is deleted

Section	Comment	Response
Page 8, para 7	We suggest include the sentence: “It should be realised that this approach departs from simple randomisation and that in certain circumstances, such as strong time trends in the response data, conventional statistical tests may have unsatisfactory properties. Applicants should therefore perform additional sensitivity analyses to support the conclusions from the primary analysis or discuss why these are not necessary.	As above, all reference is deleted
Page 8, para 8	If play-the-winner rules are not used in adequately powered studies, why are they now to be considered acceptable in small studies?	This text has remained. We believe that play-the-winner designs are sometimes used in large studies
Page 8, para 8	"Response-adaptive methods": Another limitation and drawback of the "play-the-winner" design is that the allocation ratio must be based on a short-term outcome with a quick recruitment of patients. For (very) rare diseases, the recruitment is however slow and it's often important to assess the long-term efficacy/safety of a study treatment compared to a control. Therefore [we] do not recommend using this design for small populations.	Their use will depend on the clinical indication. For long term endpoints such methods would not be suitable, but for short term endpoints they might be. Hence, text left largely unchanged
page 9, para 2	Continual Reassessment Methods (CRM) are advocated in dose finding (phase I) studies. At the end of the paragraph it states that they are to be encouraged during all stages of development. How can they be used in phase 2 and phase 3 studies? So far it has only been applied in phase I studies and no methodology (as far as we know) has been developed for using this method in phase II or in phase III studies.	Dose finding may continue beyond phase I; hence text left unchanged
Page 9, para 4	n-of-1 trials: it seems dangerous to say that each trial can be tailored to each patient. This may introduce quite a lot of heterogeneity. In the last sentence, how can they then be combined together in a meta-analysis?	This comment remains but it is stressed that pre-planned (and pre-planned series) of *n*-of-1 trials will be more convincing
Page 9, para 5	Yes, assumptions add to the data, but what if the assumptions are not correct? Sensitivity analyses may produce inconsistent results and too many sensitivity analyses will lead to problems of multiplicity. Like above, we suggest to the authors to more precisely detail their recommendations and to provide recommendations that all sensitivity analyses should be pre-planned. They should emphasize the multiplicity problem and describe the conclusions to take whenever the results of the sensitivity analyses are conflicting.	Little change to this text. It is agreed that incorrect assumptions may lead to false conclusions and this is a problem with small datasets
page 10, para 4	Phase II studies are not used to identify important prognostic factors. Phase II studies are performed in a different patient population and use different endpoints as compared to phase III studies. In addition, they are not powered to identify important prognostic factors. With small patient numbers, how can you do the proper modeling? We believe it is incorrect to state that prognostic factors are known after phase II trials.	This piece of the text has been deleted

Section	Comment	Response
Page 9, para 3	It should also be added that sequential designs usually require that only one primary end-point is selected.	Newer methods do allow for efficacy and safety endpoints to be considered simultaneously
page 9, para 3	"Sequential designs, as with response-adaptive designs, require treatment outcomes to be available quickly (relative to the patient recruitment rate). This will almost never be the case if we are looking for long term survival data…" This is seen as a major issue as we strongly believe that this statement is incorrect. Indeed Group sequential designs are successfully used for survival studies and in fact nearly all survival studies adopt a sequential design. Proposed amendment: Please amend to indicate that "**sequential designs have a particularly useful role for long-term survival data**" "Stopping boundaries for benefit and harm need not be symmetrical" This statement implies stopping boundaries are for "benefit or harm". An important reason to use sequential design is to allow the company to stop the trial early if there is a chance of showing a difference between treatments; however it should also be explicitly acknowledged that it is also possible to stop for futility where the data show that the investigational product may not be enough efficacious**.**	If long term survival is the clinically relevant endpoint then is not likely that short term survival would be accepted as a surrogate. Hence, no change has been made to the text
page 10, para 2	Non-parametric Methods: [We] believe that the use of non-parametric methods generally require slightly greater sample sizes.	There is mixed opinion about this and so no comment has been made
Page 10, para 2	References should be provided concerning the appropriateness of using Bootstrap methods in small samples. Although they may make no assumptions about data distributions, the original data still need to be representative of the overall patient population.	Reference to bootstrap methods has been deleted
page 10, para 2	Non-parametric methods, page 10: Bootstrap is a resampling method and usually applied for skewed distributions (ex. cost parameters in pharmaco-economics trials). Bootstrap method needs to determine a number of resampling (100, 1000, 10000?). It's therefore surprising that this method is described in this paper!!	
page 10, para 3	Alfa and Beta errors: It is important to specify what a good prior justification could be in order to choose a different level of significance for the Confidence Intervals.	As no consensus can be reached on appropriate alpha and beta errors, and the text was unclear, this section has been deleted
Page 10, para 3	Alpha and beta errors, page 10: This section is not clear. What is allowed and what not?	

Section	Comment	Response
Page 10, para 3	At the top of page 5 it is stated that 0.05 is arbitrary but on page 10 it is stated that the use of any other significance level needs a good prior justification. We appreciate than in really rare populations, a higher alpha or beta value might be acceptable. But the text should also emphasize the associated increased risks of a false positive or false negative conclusion. The inflation of the error thresholds might be more or less acceptable depending on the types of treatment tested and on the actual risks to future patients that would result from incorrect conclusions.	
Page 10, para 4	"It is mandatory for proper statistical inference that factors used to stratify the randomisation in a study should be used to stratify the analysis." While this general principle is well understood, it may be impractical if not impossible in small populations, when several stratification factors are considered. In practice, one will probably only include the strong predictors in the model. The illustrative example [Falk et al, 2002] had 768 strata, making it impossible to include all stratification factors in the analysis. This problem is also encountered in large populations and the center factor may not be included in the analysis model although randomization by centre was used. The effect of adding a relevant covariate in a linear model usually increases the precision. In other types of models (logistic regression, Cox regression), the effect of adding a covariate is different (it also affects the treatment estimate). Proposed language: "It is ~~mandatory~~ **recommended** for proper statistical inference that factors used to stratify the randomisation in a study should be used to stratify the analysis." "…including stratification variables in the analysis that, in fact, have very little prognostic value rarely has any detrimental effect on the analysis." It should be considered that this is not always the case especially for small centers, as it is in some circumstances possible to effectively eliminate the center's results from the analysis. In addition, this could have a major impact on non-parametric analysis. Proposed amendment: "…including stratification variables in the analysis that, in fact, have very little prognostic value ~~rarely has any~~ **often has no** detrimental effect on the analysis."	The stronger statements in this paragraph have been deleted but comments on the benefits of appropriately stratifying the analysis have been retained
page 10, para 5	how can you test the assumptions with small sample sizes ?	It is agreed that this is often not possible. Comment is included that sensitivity analyses (sensitively to different assumptions) should be presented

Section	Comment	Response
Page 11, para 1	"As with sensitivity analyses mentioned above, a variety of reasonable prior distributions should be used to combine with data from small studies to ensure that conclusions are, at least, reasonably data-dependent and not almost entirely belief dependent." For greater clarity, could the CHMP clarify what is meant by this statement and illustrate with a practical example?	We believed this message was sufficiently clear – no change made
Page 10, para 6	Greater prominence should be given to Bayesian methods as a cohenrent framework for combining disparate sources of evidence in a transparent manner.	No change to the text. Sponsors are free to give their preferred methods greater prominence if they believe they help decision-making
Page 11, para 1	in small studies, will there be enough data from the study to have a sufficient impact on the prior distribution?	No change to the text. Sponsors will have to demonstrate that prior distributions are not unreasonably dominant over the data
Page 11, section 6.3	It is recommended skip this section 6.3. Motivation: It is in the assessment where uncertainty of the validity of observed effect should be weighted against the risk of disapproving a truly effective agent and approving a truly ineffective agent in the context of the underlying disease and safety data available. Moreover, we should avoid comments (or even legal steps) from MAH's that according to the guideline the assessment should have been more flexible. The criteria formulated by Bradford-Hill are general clues for assessing a causal relationship not rules. As such are used in the context of etiological studies and not clinical trials. Assessing a treatment effect relationship implies that the results of a study are valid. It is the degree of validity of study results that differs between observational studies, small RCTs or large RCTs. Once the validity of the study results is accepted the treatment effect relationship can be assessed. Moreover the Bradford-Hill criteria should not be used as a tick off system. In non-RCT there is always the danger of confounding. If an effect size can mainly be explained by a confounding factor (basically baseline difference in prognosis) the specificity of association, strength of association and dose response relationship is explained by a confounding factor as well. Hence the presence of the latter is not enough to decide on a treatment effect relationship.	This section has been deleted

Section	Comment	Response
Page 11	It is recommended to add a new section in line with the general comments made/ The following text is proposed: **6.4 Reporting** In the study report, preferable study protocol, a justification of the deviations from the principles laid down in ICH10 is expected as well as a justification of the alternative study design chosen above other alternative designs (e.g. choice surrogate end point, lack of randomisation, lack of control) and a justification of statistical analysis. It is noted that an orphan drug status as the only justification not sufficient. Further a sensitivity analyses in order show that different assumptions about the data do not alter the conclusions is expected. In addition, the need for a long term follow for efficacy and safety should be discussed.	This section has been added
7 SUMMARY AND CONCLUSIONS		
page 11, bullet 8	This bullet offers an interesting approach. Suggest making it more optional by replacing "When planned statistical (analysis) methods fail to show treatment effects, ~~alternative approaches should be sought out~~ **it may be possible that alternative approaches do show convincing evidence of an effect. Such analyses however should be carefully justified** (and preferably anticipated in the study protocol We also suggest to add that results of different statistical approaches should be in **quantitative** agreement	This point has been deleted
Page 11, bullet 8	An even more disturbing example of unacceptable statements is the following: "When planned statistical methods fail to show treatment effects, alternative approaches should be sought out." This is exactly what we do not want! Using a large number of approaches will inevitably increase the probability of a false positive finding due to multiplicity issues. This one example of what is frequently referred to as data dredging or data torturing.	As above, this point has been deleted

Section	Comment	Response
Page 11, bullet 8	the use of alternative approaches when planned analyses fail to show a difference encourages data dredging. This leads to problems of multiplicity and type I error rates. The authors should clarify their suggestion. We appreciate that if a number of analyses on supposedly correlated endpoints lead to similar results, the evidence for a treatment effect might be enhanced. The authors should also provide guidelines for the situation where the results would be contradictory (consider the study as negative) and emphasize the problem of the risk of increased false positive results.	As above, this point has been deleted
APPENDIX		
	[We] appreciated the examples provided in this section.	The appendix has been deleted: much of the text of the main body of the document has been deleted, so that the references were no longer applicable many of the references did not apply to problems of rare diseases (and it has not been easy to find examples that do)
p. 13/16	The reference for the paper of Falk et al. is not provided. [BMJ 2002;325:465]	
Page 14, para 2.	Although this practical example is useful and illustrates a principle, it may have been more relevant to tackle a condition with a small populations rather than using a condition such as osteoarthritis which is probably not covered by this guideline.	