

DRAFT (Final) Concept Paper
On choosing appropriate estimands and defining sensitivity analyses in confirmatory clinical trials

EFSPI Comments

Page	Priority (H/M/L)	Comment
General		<p>The concept to develop a definitive guidance on what constitutes an appropriate primary estimand and the choosing of sensitivity analyses is very much appreciated. A goal combined with this concept should be to reduce a large amount of meaningless sensitivity analyses but concentrate rather on the most appropriate estimands. Extreme assumptions should be avoided (e.g. worst case analyses) as the choice for the primary estimand. In addition considerations for multiplicity are also important.</p> <p>However, others noted whether the proposed guidance to be developed could be covered in E9 (not as an addendum) and other existing specific guidelines, for example missing data, subgroup analyses etc? The existing statistical principles in E9 could be emphasised on the choice of estimands to ensure appropriate consideration is given to the assumptions being made. In addition, what may be appropriate for a confirmatory registration trial may not be so for an early-phase or market support study so any new guidance should take this into consideration. The latest FDA draft guidance on defining secondary variable as those targeted for potential labelling offers clear thinking on these aspects, and EMA could consider doing something similar.</p>
1	M	<p>First definition of estimand is not very clear (“... <i>summary quantities are termed estimands</i>”).</p> <p>The <i>constructs for illustration</i> on page 2 and 3 are more helpful in clarifying the concept but are only examples and not a definition. A stringent definition would be appreciated especially for readers who are less familiar with the concept of this topic.</p>
1	M	<p>There should be more precise use of terms – such as estimating parameters from models. The randomised patients in the trial are the sample from a population around which the hypothesis is formed and that can be parameterised, but not the sample itself.</p>
1	H	<p>Current text</p> <p>Statement of the Perceived Problem</p> <p>- <i>Poorly defined estimands lead to problems in relation to trial planning, conduct and analysis and the</i></p>

		<p><i>potential for inconsistencies in inference and decision making.</i></p> <p>Suggest changing to “Incorrect choice of or poorly defined estimands can lead to...” As some estimand can be well defined, but it is the wrong choice to answer the study objective. There are two dimensions of the estimands. First is to choose the right one, and the second is to clearly define it.</p>
1, 6	H	<p>“Considerations relating to trial conduct differ according to therapeutic area, and it may be expected that the estimand of primary relevance will also differ according to the experimental situation.”</p> <p>We think this is a very important aspect. The scientific question of interest will vary from one therapeutic area / indication to another. In order to specify a precise and clinically interpretable trial objective and an associated primary analysis, the input of our clinical colleagues will generally be needed. We would therefore welcome the nomination of clinical colleagues for the Expert Working Group.</p> <p>The discussions with clinical colleagues may provide critical input when assessing scientific questions of interest in challenging settings where patients may</p> <ul style="list-style-type: none"> • need rescue medication (e.g. in diabetes); • switch treatments (e.g. in oncology studies after disease progression); • discontinue the study treatment; • discontinue from the study for reasons that are associated with efficacy or safety; • die for diseases-related versus disease-unrelated reasons.
1, 2	H	<p>“... no definitive guidance is available on what constitutes an appropriate primary estimand for a confirmatory clinical trial.”</p> <p>We welcome the efforts of the ICH to create such important guidance. In the light of extensive differences between therapeutic areas, creating an overarching, generic guidance appears to be a challenging endeavour.</p> <p>Ideally, disease specific guidelines would follow to provide concrete guidance on the choice of the ‘primary estimand’.</p> <p>This is particularly important as a harmonisation of the precise trial objectives would alleviate some of the challenges that arise in multi-regional trials. For the ICH E9 addendum, we would suggest to include as many examples as possible so that the reader can appreciate the range of different settings.</p>
2	M	<p>This concept paper considers sensitivity analyses of different estimands. But an additional source of sensitivity analyses</p>

		and confusion for decision makers is the application of different statistical methods to make statistical inferences. For example both parametric and non-parametric statistical tests with corresponding different preference of applicants and health authorities. Could this issue also be addressed?
2	M	<p>The ‘misinterpretation by different regional authorities’ seems to relate to authorities wanting an ensured ‘directional consistency’ in their subgroups, knowing that the study is not designed to show a meaningful difference in that country/region’s data, especially if that country/region contributes <20% of the total in the program. Guidelines aimed at aligning regulatory agencies principles in this area would be useful.</p> <p>“Furthermore, similar submissions may lead to different inferences being drawn in different regions”, is this referring to differences in regions in assessing benefit risk rather than a difference in the inference of efficacy? If so, the issue may not be on clarity on the estimands of efficacy but how decision making is undertaken on the balance of benefit-risk in the different regions. In large global trials regions can be a proxy to enrichment.</p> <p>“Consequently, decisions of regional authorities may appear to weight...” is the issue on ‘defining an appropriate set of sensitivity analyses’ or is it instead defining for each regional authority how they would each consider the sensitivity analyses conducted in their decision making of balancing benefit-risk?</p>
2	H	<p>Current text A series of relevant estimands may be identified and harmonised guidance given on circumstances where it is appropriate to choose each on as an estimand of primary interest.</p> <p>Suggest The estimands should be considered based on the full analysis set (ITT). Additional estimands based on per-protocol set should be also considered. When the ITT and per-protocol analyses come to essentially the same conclusions, confidence in the study results is increased.</p> <ul style="list-style-type: none"> • Estimands are based on full analysis set will be aligned with the original study design. Therefore, its' results will be statistically meaningful. • Both FDA and CPMP guidelines refer to these approaches being preferred.
2-3	H	<p>Estimands of type 1) and similarly type 6) of the <i>constructs for illustration</i> call for follow-up of patients after end of the randomized treatment observation period and generally lead to dilution of treatment effects and so to larger and more difficult trials in addition to rather complicated interpretation. This seems appropriate in some contexts, e.g. if the outcome is survival, but will complicate trials in other situations.</p> <p>Whenever appropriate, estimands of type 3), 4) or 5) are preferable in the sense of feasibility and meaningful interpretation. Type 2) is rather close to the classic Per-Protocol definition, fully excluding a number of patients from</p>

		the analysis, and would usually not be appropriate for the primary analysis, at least for a superiority study.
3	H	<p>The concept paper focuses on so-called ‘estimands’. We note that the term ‘estimand’ is not very commonly used in the clinical trial setting. From a pure language perspective it means “what is to be estimated”.</p> <p>A very clear and precise definition of the word ‘estimand’ in the setting of confirmatory clinical trials is therefore crucial. The definition currently provided in the concept paper lacks some clarity and may cause confusion.</p> <p>The reasons are two-fold.</p> <p>Firstly, the concept paper states: “In a confirmatory clinical trial data are collected to measure outcomes that quantify the impact of experimental interventions in comparison to a known control group, typically over a defined period of time. Inference focuses on summaries of these measures (such as the mean) for the target population of interest. These summary quantities are termed estimands.”</p> <p>Based on this statement, an estimand could be interpreted as the <i>estimator</i> which is used to infer the value of an unknown parameter in a statistical model. However, the National Academy of Sciences (NAS) report states “These summary quantities are termed parameters or estimands.”</p> <p>We understand that an estimand can be regarded as a parameter in a statistical model to be determined in a specific population. Do you concur with this interpretation?</p> <p>Secondly, from the definition and the examples in the concept paper, it is not clear to what extent an estimand depends on analysis model assumptions like</p> <ul style="list-style-type: none"> • inclusion/exclusion of certain baseline covariates in a (generalized) linear model; • inclusion/exclusion of interactions in a linear regression. <p>Would one of the above model modifications lead to different estimands?</p> <p>We would like to illustrate some of our insecurities with regard to the ‘estimand’ definition via the following examples:</p> <ol style="list-style-type: none"> 1. Consider a placebo-controlled study where one is interested in assessing the following scientific question: “What is the difference in mean outcome improvement for all randomized patients after 12 weeks of treatment?” The estimand is the ‘mean difference at week 12 based on all randomized patients’. Further, assume that all data to assess this question are available. Two linear regression models are considered: one is

		<p>including an interaction between gender and treatment, the other one not.</p> <p>Both analyses yield different estimators for the same estimand. In the regression model which includes the interaction, the estimator will be a linear combination of gender-specific treatment effect estimates. This estimator will therefore depend on the proportion weights (male versus female) that are used to estimate the population-averaged treatment effect. Will equal weights or sample weights be used? And do all these analyses (with/without interaction, equal/unequal weights) address the same estimand? When are we talking about varying model assumptions and when about changing the estimand?</p> <p>2. Consider a placebo-controlled study where one is interested in the following scientific question: “What is the difference in mean outcome improvement attributable to the initially randomized treatment for all randomized participants after 12 weeks of treatment?” Now suppose some patients in the active arm discontinue study treatment, thus the primary endpoint measurements at week 12 are confounded by whatever the patients take after study treatment discontinuation. Would an analysis including all randomized patients address the question of interest?</p> <p>3. The issue in example 2 would become more complex when patients discontinue the study after they stop treatment, which would lead to missing data. In this case one would need assumptions on the response behaviour after stop of treatment, e.g.</p> <ul style="list-style-type: none"> • the drug effect is lost straight after discontinuation of the study drug and the response follows that of ‘similar’ placebo arm patients; • the drug effect is lost only slowly after discontinuation of the study drug and the response follows that of ‘similar’ placebo arm patients; • the drug effect is not lost after the discontinuation of the study drug and the response follows that of ‘similar’ patients in the active arm. <p>Do these three assumptions imply the same or different estimands? All assumptions try to assess the same scientific question of interest. Dependent on the plausibility (e.g. half-life of the drug, single dose versus multiple dose design) one could use one of these assumptions within the primary analysis and consider the other assumptions in a sensitivity analysis. Based on the current definition and examples we are not clear as to whether we are confronted with one, two or even three estimands.</p>
3	M	The document states ‘The extent to which this type of analysis is needed once an estimand of primary importance is

		<p>agreed by all parties should be discussed.'</p> <p>What is the scope of these parties? Are there likely to be situations in which different stakeholders may be interested in different estimands? If the parties are limited to the regional regulatory authorities (as the final paragraph seems to suggest) this may potentially be less of a concern?</p>
3	H	<p>"A common understanding of what is meant by sensitivity analyses should be derived and this should inform a statement to clarify the objectives for sensitivity analyses."</p> <p>We very much agree that a clearer understanding of what constitutes a sensitivity analysis is much needed.</p> <p>Questions that you may want to consider in this context are:</p> <ul style="list-style-type: none"> • Should a sensitivity analysis address (only) the primary estimand of interest? • Should a sensitivity analysis focus on the same population as is used for the primary analysis? • Which assumptions should be assessed through a sensitivity analysis? <ul style="list-style-type: none"> ▪ Covariate adjustment; ▪ Interactions; ▪ Error distribution; ▪ Missing data mechanism;
3	H	<p>"To confirm robustness, should results be consistent in terms of the strength of evidence presented from a statistical perspective, or in terms of estimated effects from a clinical perspective? It may be questioned whether it suffices for decision makers to be presented with explanations for discrepancies between different analyses." should not be an "or" but should really be "and" as both statistical and clinical relevance are important. The analyst has first to confirm the statistical robustness of the results, and then the clinical relevance of the balance of benefit-risk is then key for decision making. The clinical relevance will take into account other factors outside of demonstrating the statistical robustness of the estimands. In addition, as far as possible sample size considerations should try to make sure that clinical and statistical perspectives coincide.</p> <p>Agreement is needed on how decisions are made based on primary and sensitivity analyses results.</p> <p>In this context, different criteria could be considered. For example, a violation of the robustness could be claimed if the sensitivity analyses lead to</p> <ol style="list-style-type: none"> 1. qualitatively different results (e.g. effect reversal); 2. non-significant treatment-differences;

		<p>3. effect sizes which are not clinically meaningful; 4. ... Ideally, sponsors and health agencies will agree on a criterion for decision making prior to initiation of the study. We suggest you refer to the subgroup guidance.</p> <p>Generally, we believe that criterion 1 above is useful. Criteria 2 and 3, in contrast, may often be too strict. Reason being that sensitivity analyses make less plausible (and often deliberately conservative) assumptions, compared to the primary analysis. The resulting effect sizes and p-values should therefore be interpreted within that context.</p> <p>The estimated effects from a clinical perspective are particularly important for label considerations. This raises several questions:</p> <ol style="list-style-type: none"> 1. What is the 'effect' of primary interest for the label? 2. Should this quantity match the estimand? 3. Should the results from the primary analysis be reported in the label? 4. What is the role of the sensitivity analyses results?
3	M	<p>The document states 'An improved framework should restrict to the number of sensitivity analyses that are provided to those that add value to decision making by targeting a particular assumption that is critical for inference. '</p> <p>A framework should provide a structure for sensitivity analyses it is not clear to me that it should restrict sensitivity analyses. If there is a clear rationale for alternate sensitivity analyses the framework should not restrict these analyses.</p> <p>In addition to the identified issues to be addressed, would this framework provide guidance on the determination of levels of the importance of different sensitivity analyses? In addition, should one issue a guidance regarding inference when different conclusions could be drawn from sensitivity analyses compared to primary analysis?</p> <p>'It is anticipated that specific definitions will need to be generated through discussions" this highlights that there is a not a one size fit all approach and that what is being advocated is more discussion with each regional authority in agreeing how sensitivity analyses will be used to support the decision making for assessing the balance of benefit-risk.</p> <p>In terms of sensitivity analyses, and their purpose, it would be helpful to consider regional analyses (often for regional approvals). The document talks of understanding the role of sensitivity analysis, how it is interpreted jointly with primary result, and this is a key area where this may be the case. Similarly, it is often the case that many regions</p>

		perform similar reviews of data for their area, so understanding this particular sensitivity/subgroup analysis in the context of all such analyses conducted, and considering how such analyses will be considered in the context of the primary analysis is important.
3	M	<p>Current Text <i>Mallinckrodt et al</i>, give a further illustration: 6. For all randomized participants at the planned endpoint of the trial attributable to the initially randomized treatment</p> <p>Suggest To parallel the above five points, in the confirmatory clinical trial setting we would still be interest in (Difference in)</p>
3	M	If sensitivity analysis is conducted based on subgroups, heterogeneity should be examined or discussed.
3	H	“.. and to present analyses purported to be sensitivity analyses without indication of which choice or assumption is being investigated”. One would hope the statistical analysis plan would describe the role of each sensitivity analysis relative to the assumptions of the planned primary analysis. When presenting the results of the sensitivity analyses, whilst the aim is to show the robustness of the conclusions drawn from both the primary and supportive sensitivity analyses, if the range of assumptions are not being summarised along with the interpretation of all the results, then this is an important detail lacking and should be addressed. Is the deficit in both defining sensitivity analyses and in interpreting sensitivity analyses?
3-4	M	<p>The document states.’ It is increasingly common that methods for primary analysis are proposed which focus on a particular type of ‘estimand’ and that rely on assumptions that cannot be verified and are unlikely to be true.’ The “unlikely to be true” is overstated and should be removed.</p> <p>Is it the intent of this document to provide guidance of the choice of primary analysis? The primary analysis for the protocol is usually the one quoted in the label and in publication. So consistency of choice of primary analysis is important in the broader context. Is there an option to provide more enriched information in label by including the sensitivity analyses?</p> <p>The role that each sensitivity analysis has in supporting the primary analysis should be clearly described in the statistical analysis plan. This should include an assessment of the ability of the assumptions supporting each analysis to be verified. Where assumptions are not verifiable, should these sensitivity analyses have less weight compared to other sensitivity analyses where assumptions can be verified. This level of detail could be discussed in the statistical analysis plan which then enables a more structured discussion in the interpretation of sensitivity analyses.</p>
4	H	“Nevertheless, it is proposed that the discussion of estimands and of a framework for sensitivity analyses should cover these topics comprehensively and not only in relation to the problem of missing data.”

		<p>The concept paper has focused on efficacy endpoints and referred mainly to challenges that arise due to missing data or intake of rescue medication / treatment switching. We welcome that a broader discussion is planned in the future.</p> <p>Some areas of particular interest and remarks are:</p> <ul style="list-style-type: none"> • Discrete longitudinal data analysis and survival analyses: The scientific question of interest changes when adjusting for covariates and/or random effects. For an example, see Chapter 16 in: G. Molenberghs and G. Verbeke. Models for discrete longitudinal data. Springer, 2005. • For a structured discussion on estimands, do we need to distinguish between superiority versus non-inferiority trials / placebo-controlled versus active controlled trials / symptomatic versus outcome studies / etc.? <p>Should the topic of 'appropriate estimands and defining sensitivity analyses in confirmatory clinical trials' also be discussed in terms of safety endpoints? What are the implications for benefit-risk assessments?</p>
4	H	<p>One thought, when sample sizes are small, and reliance on the central limit theorem is not appropriate, it would be interesting to know by how much non-parametric methods "should" differ should the assumption of normality be true (for example there is ~ 95% relative asymptotic efficiency for the sign-rank test vs t-test). One could run the non-parametric analysis knowing what to expect should the data be normally distributed, and if the results are much different than that, one may look more closely at the non-parametric result. The sensitivity analyses usually seemed to be aimed at deviations from the treatment effect, variance, etc. while still assuming the normality assumption is correct. Again, this applies more to trials with smaller sample sizes.</p>
5	H	<p>Could ICH E9 be updated to discuss in more detail the role of sensitivity analyses in terms of definition and interpretation of them in establishing efficacy rather than create an addendum? Increasing the specificity of sensitivity analyses would result in a higher quality and more comprehensive evidence base on which the regulatory authorities can use for their decision making for the balance of benefit-risk.</p>