EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

31 January 2019
EMA/CHMP/539146/2013
Committee for Medicinal Products for Human Use (CHMP)

Guideline on the investigation of subgroups in confirmatory clinical trials

Draft Agreed by Biostatistics Working Party	September 2013
Adoption by CHMP for release for consultation	23 January 2014
Start of public consultation	30 January 2014
End of consultation (deadline for comments)	31 July 2014
Agreed by Biostatistics Working Party	September 2018
Adopted by CHMP	31 January 2019
Date of coming into effect	1 August 2019

Keywords	***Subgroup analysis, confirmatory clinical trials, randomised controlled trials, internal consistency, heterogeneity, biostatistics, assessment of clinical trials, analysis plan, exploratory analysis, benefit/risk assessment.***

Guideline on the investigation of subgroups in confirmatory clinical trials

Table of contents

1. Introduction and problem statement

In line with DIRECTIVE 2004/27/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 31 March 2004 a marketing authorisation shall be refused if, after verification of the particulars and documents listed in Articles 8, 10, 10a, 10b and 10c, it is clear that:

(a) the risk-benefit balance is not considered to be favourable; or

(b) its therapeutic efficacy is insufficiently substantiated by the applicant; or

(c) its qualitative and quantitative composition is not as declared.

Consequent to (a) and (b), evidence of therapeutic efficacy and evidence to inform the risk-benefit decision is generated in the clinical development programme in particular in Phase III confirmatory clinical trials. Confirmatory clinical trials are performed in late-stage drug development to inform a risk-benefit decision and to justify a treatment recommendation. For confirmatory trials, usually robust evidence for therapeutic efficacy is required in a relatively broad patient population that is representative of patient population to be described in Section 4.1 of the SmPC (external validity). Evidence is considered to be more robust if treatment effects across the trials in the application, as well as in relevant subgroups within one trial (internal consistency), are consistent and substantiate the claim to be made for the experimental treatment. This justifies, following assessment of treatment effects on primary and secondary endpoints in the whole trial population, a regulatory assessment of relevant subgroups during assessment of Marketing Authorisation Application (MAA.

It is known that different patients will respond differently to the same intervention, and also that the same individual may respond differently to the same intervention on different occasions. This latter variability in response in the same patient usually remains unexplained but it is plausible, and widely accepted, that some of the variability in response between patients is caused by demographic, environmental, genomic or disease characteristics, co-morbidities, or by characteristics related to other therapeutic interventions (e.g. extent of pre-treatment or concomitant treatment). ICH E5 describes "genetic and physiologic (intrinsic) and the cultural and environmental (extrinsic) characteristics of a population" and the CHMP Points to consider (PtC) on multiplicity issues in clinical trials states "Some factors are known to cause heterogeneity of treatment effects such as gender, age, region, severity of disease, ethnic origin, renal impairment, or differences in absorption or metabolism", and indicates that "analyses of these important subgroups should be a regular part of the evaluation of a clinical study". Grouping together patients with similar characteristics in one or more of these factors is therefore an intuitive way to explore variability of response to treatment between different groups of patients within a clinical trial dataset and constitutes an integral part of the risk/benefit assessment.

It is widely understood that subgroup analyses need to be interpreted with caution because of the multiple data presentations that arise when investigating response to treatment within each level of the many possible intrinsic and extrinsic characteristics. Compounding the problem, when reviewing a display of subgroup analyses, the reviewer's eye may be drawn to those groups with extreme estimates of effect, whether smaller or larger (or in opposing direction) than the overall effect. An incautious review of subgroup analyses can result in unreliable inferences and, consequently, to poor decisions from the clinical trial sponsor or regulator. By definition, subgroups of a clinical trial population are reduced in size. The size of the respective subgroups often reflects the epidemiology of the disease and thus subgroups can be small and summary statistics are frequently associated with considerable uncertainty. In very rare diseases analysis of subgroups in clinical trial datasets might not be informative.   Notwithstanding these difficulties, whether the treatment effect is truly homogeneous in subgroups cannot be known and hence trial sponsors and regulatory decision makers are put in a difficult situation: whether to accept an assumption of homogeneity without further consideration and disregard extreme and/or pharmacologically plausible findings in subgroups, or whether to anticipate some heterogeneity and, with appropriate caution and investigation, attempt to use the results of subgroup analyses as one piece of evidence to inform decision making.

It is considered that the careful discussion of subgroups is an integral part of clinical trial planning, analysis and inference. However, the role of these subgroup analyses in decision-making is challenging and merits a dedicated guidance document.

2. Scope

This document is intended to provide assessors in European regulatory agencies with guidance on assessment of subgroup analyses in confirmatory clinical trials that are presented in a Marketing Authorisation Application. These considerations impact on the planning of the clinical trial and hence the document should also be useful to clinical trial sponsors and to assessors engaged in providing Scientific Advice. This guidance document describes principles and assessment strategies and does not dictate the use of any particular statistical methodology for estimating or testing the treatment effect and its consistency in subgroups of the trial population.

An important distinction is made between investigation of a subgroup as part of the confirmatory testing strategy and an exploratory investigation of subgroups during risk-benefit assessment. Whilst a number of the considerations outlined in this document will apply to the former, investigation of a subgroup as part of the confirmatory testing strategy is principally a problem related to multiple-testing because the trial seeks to test hypotheses relating to both the subgroup and the full trial population. Recommendations regarding pre-planned approaches for decision making in a confirmatory testing strategy based on subgroups are therefore not discussed in this guidance document: the guiding principles and examples for multiple-testing procedures that control the overall false positive rate are described in the Guideline on multiplicity issues in clinical trials.

Instead, the focus of this document is the exploratory investigation of subgroups as part of the assessment of the applicability of the treatment effect to the patient population. In principle, three situations can be distinguished in which this more exploratory investigation of subgroups might be pursued (see Sections 6.3-6.5). The first scenario is the most common, applying to all dossiers in which confirmatory clinical trials establish statistically persuasive and clinically relevant efficacy in the target population. The second two relate to more challenging scenarios where either the clinical relevance of the treatment effects, or the overall strength of evidence, is not compelling across the whole trial population:

- Assessment scenario 1: The clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated in the primary analysis population. It is of interest to verify that the conclusions of therapeutic efficacy (and safety) apply consistently across subgroups of the clinical trial population.
- Assessment scenario 2: The clinical data presented are statistically persuasive in the primary analysis population but with therapeutic efficacy or risk-benefit which is borderline or unconvincing and it is of interest to identify post-hoc a subgroup, where efficacy and risk-benefit is convincing.
- Assessment scenario 3: The clinical data presented fail to establish statistically persuasive evidence in the primary analysis population but there is interest in identifying a subgroup, where a relevant treatment effect and compelling evidence of a favourable risk-benefit profile can be assessed.

Whilst the decision-making problem differs, the principles outlined in the document apply equally to:

- subgroup investigations for efficacy or safety, or assessment of risk-benefit.

- confirmatory clinical trials without regard to choice of control group (placebo or active control) or primary hypothesis (superiority or non-inferiority / equivalence).

Section 4 presents some underlining principles. Sections 5 and 6 respectively give guidance on assessment strategies and discuss implications for trial planning regarding investigation of subgroups.

Considerations of efficacy and risk-benefit in subsets of the target population can be important in determining the Therapeutic Indication, in constructing posology, in giving warnings to the prescriber where information is lacking or where evidence is weaker, and these may influence requests for post-authorisation studies to address uncertainties in efficacy or specific safety concerns. There may also be interest in criteria for determining inclusion of information in subgroups in the European Public Assessment Report and in Section 5.1 of the Summary of Product Characteristics. This is predominately a consideration of whether information on subgroups would provide reliable information for the prescriber. The criteria outlined in Section 5 may help to decide whether the evidence generated may be considered reliable. Section 7 presents some high-level considerations on reporting of results from subgroups in assessment reports and product information.

3. Legal basis and relevant guidelines

The Guideline has to be read in conjunction with Annex I to Directive 2001/83/EC, as amended and all other relevant EU and ICH-guidelines. The following documents are of particular relevance:

Points to consider on multiplicity issues in clinical trials (CPMP/EWP/908/99)

Points to consider draft guideline on adjustment for baseline covariates (CHMP/EMA/295050/2013)

Points to consider on application with 1.meta-analyses, 2.one pivotal study (CPMP/EWP/2330/99)

ICH E9 Statistical Principles of Clinical Trials (CPMP/ICH/363/96)

ICH-E17 Multi-Regional Clinical Trials (EMA/CHMP/ICH/453276/2016)

Guideline on Summary of Product Characteristics, published by the European Commission, Revision 2, September 2009.

4. General considerations

4.1. Defining subgroups

In this guideline the term ‘subgroup’ will be used to refer to a subset of the clinical trial population defined by one or more intrinsic and extrinsic factors (see ICH-E5) of the patients under investigation, usually measured at baseline. The term ‘sub-population’ will be used to refer to a subset of the patient population described in the targeted therapeutic indication. Generalisability of findings in subgroups to subpopulations requires the same sort of consideration as the question whether the clinical trial population is representative of the population targeted in the therapeutic indication (external validity).

Examples of subgroups include demographic characteristics (including genetic or other biomarkers), disease characteristics (including severity or (pheno)type of disease), environmental aspects, and clinical considerations (e.g. use of concomitant medication), region or centre. Post-baseline covariates may be affected by treatment received and will not usually be appropriate to define subgroups for the investigation of a treatment effect. Patients excluded from a particular subgroup are described as the complementary subgroup.

Subgroups can be defined based on dichotomous (e.g. male / female), categorical (e.g. region), ordered categorical (e.g. disease score at baseline) or continuous (e.g. age) factors. The definition should reflect meaningful entities (subpopulations) of the patient population with relevance for decision making in clinical practice. In consequence, investigation of (ordered) categorical and continuous factors requires careful consideration when pooling across multiple levels of a single factor (e.g. centre or region), or cut-off points for categorisation of a continuous factor are needed.

Subgroups may be defined based on multiple factors (e.g. females aged >65) but for the considerations discussed in this document subgroups defined based on a single factor (e.g. gender) will suffice in most instances. The risks described in this document around analysis and interpretation of multiple subgroup analyses are exacerbated by also considering subgroups identified through multiple factors, though the need for this more complex type of investigation cannot be excluded.

Risk scores, usually based on multiple different clinical and demographic factors, can be useful for the definition of subgroups if the scores are well established in clinical practice and a separate assessment of the treatment effect in multiple subgroups formed by the individual factors might not be needed.

The research into biomarkers and genetic testing gives rise to issues for the definition of subgroups in two aspects. Firstly, an almost infinitive number of subgroups can be defined based on the combination and categorisation of multiple biomarkers. Secondly, knowledge about the prognostic value of the biomarker, and its clinical utility, is usually lacking. Particular diligence is needed therefore when considering defining subgroups in this manner.

4.2. The importance of subgroup analyses in decision making

Confirmatory trials should reflect the target population to be treated and thus the extent of heterogeneity in the clinical trial population is determined primarily by the extent of heterogeneity in the target population that the trial population represents. The extent of additional heterogeneity in the risk profile of a patient population included in a confirmatory clinical trial will depend on the inclusion / exclusion criteria of the study in respect of factors important for the prognosis of the disease course, the sites, countries or regions where the study is conducted and, in instances, by the experimental medicinal product under study. The more heterogeneous the population studied, the greater the importance of subgroup analyses to check that the estimated overall effect is broadly applicable to relevant subgroups.

In clinical practice treatment selection for a certain patient relies not only on benefits and risks demonstrated in an overall target population but considers knowledge of factors that modify treatment effects. Benefit and risks may be modified based on intrinsic and extrinsic factors (e.g. demographic factors, disease characteristics, required co-medication). In the clinical trial, investigations of the treatment effect in such subgroups, identified through relevant factors, supports not only the wording of an indication statement, but the possibility to provide information to prescribers on factors that modify treatment effects, or information that the effects of treatment are uncertain. The term relevant factors in first instance seems to be quite open and allowing for interpretation. However, in discussion of a specific disease or indication, there is often good agreement amongst stakeholders about what should be considered relevant (e.g. tumour stage in oncology, NYHA-stage in cardiovascular disease). Obviously, this knowledge evolves and information from other data sources (external to the trial) may emerge during the course of a trial.

An indication statement should describe a target population where it has been concluded that there is therapeutic efficacy and a positive risk-benefit. Sub-populations where evidence of efficacy is inadequate, or where risk-benefit is not concluded to be positive, should not be included. A well-defined subgroup with a (true) considerably smaller treatment effect will require a separate assessment of efficacy and risk-benefit in order to justify treatment. For example, for drugs associated with serious toxicity across the breadth of the population, the risk-benefit ratio may be negative in less severely ill patients if the absolute magnitude of benefit is smaller.

The anticipated heterogeneity within a target population, and the potential for inconsistency of therapeutic response to treatment justifies the exploration of subgroups after assessment of the whole trial population.

4.3. Key concepts in the assessment of subgroups

A number of concepts and the respective definitions required in the following discussion about the role of subgroups for decision making about efficacy and benefit of a new drug are introduced here:

a. **Heterogeneity** relates to the extent of differences within the target patient population, or a clinical trial population, in factors that are prognostic for outcome or predictive of treatment effects. The more heterogeneous the population, the more important the investigation of treatment effects in well-defined subgroups.

b. **Consistency** describes the extent to which estimated treatment effects in relevant subgroups assures that the overall treatment effect applies to the breadth of the trial population. In each specific indication, it is required to judge the extent of differences in effect size between subgroups that are of potential clinical importance, and hence that would indicate inconsistency. Inconsistency in estimated treatment effects is one indicator for further assessment in particular subgroups.

c. **Credibility** describes the extent to which subgroup findings can be concluded as being well substantiated and hence relied on for decision making. Credibility depends on the degree of well-founded, a priori definition, the **biological plausibility** for a particular finding and **replication** (see below).

d. **Biological plausibility** is a concept describing the extent to which a particular effect (in this case differential effects of treatment between subgroups) might be predicted, or might have been expected, based on clinical, pharmacological, and mechanistic considerations, and considerations of other relevant external data sources. Plausibility is primarily a clinical and pharmacological judgement and, unless already considered at the planning stage, is usually not a directly quantifiable or measurable concept.

e. **Replication** refers to whether an effect of a particular covariate, or differential effects between particular subgroups, is seen in multiple data sources: specifically, whether an inconsistency in one confirmatory clinical trial are also seen in independent clinical trial data (another phase III trial, phase II exploratory trial, or a trial from outside the development programme, but with similar experimental conditions).

4.4. Problems with conducting multiple subgroup analyses

The key problem of exploring subgroups is closely related to issues with multiple testing. Multiple factors are available on which subgroups can be identified and opportunities to select how the subgroup should be constructed (e.g. with different categorisations of a continuous factor) both introduce multiplicity and analysis of these subgroups may lead to contradictory conclusions simply due to the play of chance.

Considering multiple subgroups increases the probability of false positive findings, defined here as subgroups where the effect is concluded to differ from the effect seen in the primary analysis population when in fact it does not. False negative conclusions are possible as well and are equally important. These are defined as subgroups for which it is not detected that the effect is truly different from the overall treatment effect.

The possibility of false positive findings is often quoted as a reason to ignore or dismiss differential effects in a subgroup and its complement. Critically, this would mean not investigating the underlying hypothesis that effects across different subgroups are consistent with the overall outcome of the trial. It is not acceptable to assume consistent effects across important subgroups without further investigation or discussion.

When assessing results from a clinical trial there is the additional risk that the balance afforded by randomisation is not fully preserved when looking into subgroups, such that findings in one of multiple subgroups are more likely to be driven by baseline-imbalance in covariates between treatment groups than by an effect of treatment. It is expected that subgroup findings that are inconsistent with the overall outcome of the trial will first be explored with statistical methods that investigate whether proper adjustment for covariate imbalances can explain the differences between the overall- and the subgroup estimate for the treatment effect.

Erroneous conclusions that can arise from testing multiple subgroups within one clinical trial dataset is not only relevant in Phase III trials. Exploratory trials may result in an overall effect that is not impressive, but a signal of relevant efficacy may be apparent in a subgroup, and the sponsor might be tempted to pursue development of the drug in this subgroup. This type of selection will on average be associated with artificially extreme and potentially unreliable estimates of subgroup effects that would be, however, detected during the further drug development programme. Statistical methodology is available that allows to shrink the estimate and may be used to put overly optimistic outcome into perspective and this can reduce the risk for over-optimistic planning of subsequent trials.

4.5. Methods for assessment of the consistency of the treatment effect in subgroups and associated data presentations

The need to assess the consistency of the treatment effect in subgroups of the patient population is a direct consequence of the inclusion of a broad patient population, which is appropriate for a phase III clinical trial. It should be noted, however, that consistency of the treatment effect is not of value in itself. Differences in the estimated treatment effect in subgroups are of relevance where the treatment effect is credibly reduced to an extent that the global assessment of therapeutic efficacy and a positive risk-benefit ratio may no longer be valid for patients in the subgroup and a separate discussion of therapeutic efficacy and benefit/risk in subgroups is required.

Statistical tests for interaction (heterogeneity testing, e.g. the Breslow-Day test) have been used to decide about the need to further investigate subgroups of the trial population. It has, however, been demonstrated that currently available tests lack power (sensitivity) to detect inconsistency in treatment effects that are of potential clinical importance while, if assessed at increased levels of the type-1-error, lose specificity. This is also true for other measures of heterogeneity between different trials in a meta-analysis or the subgroups within one trial (e.g. I^2). The situation is further complicated by the fact that the size of the subgroup (and its complement) depends on the prevalence of the factor levels forming the subgroup and the test for heterogeneity loses power further if the subgroup and its complement are different in size. The role for tests of interaction is therefore limited, covering only statistical aspects, with a place to generate signals for further inspection. Interaction tests cannot be the only tool to identify or exclude inconsistent findings in subgroups that may be of relevance for clinical decision making. Because of their role in signal generation, if these tests are conducted nominal significance levels greater than 5% should be employed.

As a consequence, although still common practice, the sole reporting of an isolated p-value from a test for interaction is an inadequate basis for decision making. It is important to assess the estimated treatment effects in subgroups and to discuss the clinical relevance of observed differences.

Exploratory subgroup analyses are usually presented for a range of factors. Presentation of results should include estimates and confidence intervals. Whenever a subgroup analysis is displayed, the analysis of the complement subgroup should also be displayed. Where dichotomous or categorical variables are used to define subgroups it would be expected to see results presented in Forest plots.

A formal rule for the interpretation of subgroup findings presented in a Forest plot that is both sensitive to detect inconsistency in treatment effects and specific to avoid false-positive findings is not available. Visual inspection should consider the estimate and precision of the overall effect, the estimates and confidence intervals for the effect in each subgroup (reflecting the precision of the estimate as well as the amount of information available for a particular subgroup) and the overall number and size of (key and exploratory) subgroups presented in relation to the anticipated heterogeneity of the patient population. A subgroup result that is estimated with less precision (wide confidence intervals) indicates a sub-population that is rare or is under-represented within the trial. In case a key subgroup seems to be under-represented, regulatory review should consider whether, in totality (i.e. including considerations from other data sources and of biological plausibility), sufficient information for benefit-risk assessment is available.

A Forest plot that includes all relevant subgroups and shows consistency in the direction and magnitude of the treatment effect is generally accepted as adding validity to the overall conclusion that the outcome of the trial applies to the studied patient population. For continuous variables, plots should be presented to characterise how the estimated effect of treatment changes over the range of the factor. If multiple pivotal trials are presented, a Forest plot on a pooled dataset might be presented, in addition to assessing consistency between trials.

A signal that deserves further consideration might be a subgroup, in particular a 'key' subgroup, where it is judged that the point estimate differs markedly from the effect in the overall population and, in particular, where the point estimate lies below an effect size of minimal clinical relevance (or lies towards or below an agreed non-inferiority margin). Any signal for inconsistency in treatment effects taken from a Forest plot does not determine the credibility of the finding, which should be considered as a subsequent step in assessment (see Section 4.6).

If in one subgroup the treatment effect is larger than the average treatment effect, the complementary subgroup will by necessity show a smaller than average treatment effect. In instances where directional consistency is not obvious, and further inspection is warranted, the reviewer is cautioned that the different subgroups presented are not independent and the same underlying factor may be responsible for multiple observed inconsistencies. For example, renal insufficiency increases with age and a potential interaction between the treatment effect and kidney function will show up both in subgroups defined by level of renal function, in those defined by age, and in those also correlated with age e.g. use of a particular co-medication. Forest plots indicate the need for a close investigation of potential sources of the inconsistency and in how far they need to be considered in the overall risk-benefit assessment.

Statistical interactions are scale and model dependent. Interactions in linear regression models represent departures from additivity (differences in treatment effects on an absolute scale) while interactions in logistic/Cox regression models represent departures from a multiplicative model (differences in treatment effects on a relative scale). Even where the effects of a medicine in subgroups of the trial population are likely to be similar on the relative scale (e.g. 20% reduction regardless of baseline) the (larger) effect observed in patients with severe disease may offset the risks, while the (smaller) effect observed in patients with mild disease may not. It is recommended that the exploration of interactions and effects in subgroups proceeds first on the scale on which the endpoint is commonly analysed. It might be useful to present supplementary analyses on the complementary scale for those covariates or subgroups where inconsistency is observed.

If consistency turns out to be scale dependent, it is also appropriate to do the analyses in a scale where the treatment effect is consistent across levels of the factor. Irrespective of whether unexplained inconsistency remains, or where consistency is only seen in the relative scale, benefit/risk in subgroups needs to be assessed carefully as soon as the risk profile of the population is not homogeneous.

Estimates derived from exploratory subgroup analyses should be interpreted with caution. Not only might the play of chance impact the estimated effect, but it is tempting to focus on subgroups with extreme effects, which introduces a selection bias. Some methods have been proposed in the statistical literature to reduce the problem, in particular methods that shrink estimates based on certain underlying assumptions of consistency of the treatment effect. When discussing subgroups as part of clinical trial interpretation and regulatory decision making, these methods may be presented by sponsors, but the underlying assumptions must be carefully considered. In general, these methods may help to put unexpectedly positive estimates into perspective. However, they are not a basis to dismiss subgroup findings where the treatment effect is smaller than the average. In such situations, the finding of the subgroup analysis should instead be assessed for its credibility and, if needed, its potential implications on the assessment of risk-benefit.

It has been discussed that multiplicity associated with subgroup analyses and interaction tests should be addressed through changes to nominal significance levels for tests or presentation of confidence intervals. However, since these investigations mainly serve as an indicator for the need to initiate further exploration, adjustment would be counter-intuitive and is not recommended. The fact that multiple subgroups are examined to avoid that untoward effects of treatment are overlooked requires compensation by careful assessment of signals regarding their plausibility and the ability to find replication from within or outside the current application.

Development of other methods that are sensitive and specific for generating signals for subgroups that merit further scrutiny in the assessment of efficacy and risk-benefit, is supported.

4.6. 'Credibility' of a subgroup finding

As indicated in 4.3, biological plausibility and the ability to find replication are key elements to evaluate the credibility of a subgroup finding. Whilst some evidence will be available when planning the trial on which factors are likely to be prognostic for patient outcome or predictive of therapeutic response the credibility of findings in a subgroup of interest needs to be re-evaluated based on the data generated in the clinical trials supporting the MAA, and other external data or knowledge that has emerged during the course of the trial. The assessor must weigh all evidence that can be brought to bear on the question of whether findings in a particular subgroup can be considered credible.

In particular, where estimates of treatment effects in subgroups are made imprecisely, and inference is not definitive, strong biological plausibility, or absence thereof, or replication of evidence may well contribute at least as much weight to the regulatory assessment as the pattern of data that is observed across the range of subgroup analyses presented. In assessing the biological plausibility or replication of evidence it is useful to consider not only the subgroup of interest, but also explanation or support for observed findings in the complementary set.

Knowledge available when planning the trial will inform the choices that are made for which factors are used to stratify randomisation or analysis, and which further subgroup investigations are planned (Section 6). Such pre-specification, reflecting that there is plausibility for differential outcomes between subgroups, can lend credibility to positive or negative subgroup findings. Conversely, the absence of pre-specification cannot be taken as a direct argument that results in a particular subgroup lack credibility. In particular, for adverse findings in subgroups there should be no disincentive to properly consider and pre-specify all relevant subgroups at the planning stage. Instead, arguments for lack of credibility should focus on biological plausibility and (absence of) replication.

Having two or more relevant sources of evidence, and hence being able to assess replication, is of great assistance to interpretation. Where two or more trials from the same MAA can be used to investigate a subgroup effect, the weight of evidence from directly relevant clinical trial data increases credibility. A trial from outside the development programme, but with similar experimental conditions (similar intervention(s), variable(s) etc. with the relevant sub-population represented) might also be justified as providing relevant insight. Evidence for differential effects in well-defined clinical subgroups that are replicated across available clinical trials can be compelling even in the absence of a definitive mechanistic explanation. Conversely, an inconsistent finding in one trial is more readily disregarded if evidence from one or more other trials that are comparable in design does not replicate this inconsistency, in particular where there is no *a priori* reason to expect a differential effect. Because of the possibility of erroneous subgroup findings, a development programme with two trials in which a critical subgroup finding can be assessed is clearly advantageous. This is consistent with the guideline on applications based on a single pivotal trial which stresses the importance of internal consistency in a single pivotal trial.

In the end it is a major part of the regulatory assessment to weigh signals that have been generated during visual assessment and/or by means of statistical methods against the knowledge at the planning stage, from other trials in the same development program, or other relevant data sources. Algorithms for assessing credibility of findings in subgroups are presented below and in Annex 1. No algorithm can replicate the nuances and complexities of all possible decisions, but these should act as a guide to assessors in considering the strength of evidence from subgroups.

5. Issues to be addressed during assessment

5.1. A strategy for the assessment of clinical trials

Analysis of a clinical trial starts with a confirmatory test of the primary endpoint in the primary analysis population. The assessment of secondary endpoints is an important step to support the primary result of the trial or to put the primary result into perspective. Together with the assessment of safety in the overall population they form the basis for the risk-benefit assessment. The assessment of relevant subgroups is an important subsequent step to support the conclusion that the treatment effect as well as the overall conclusion applies to the potentially heterogeneous population of the trial. The analysis of subgroups would usually proceed with an adjustment for those factors that were also included to stratify randomisation in line with the primary analysis.

Factors that define subgroups of the target population should have been carefully considered at the planning stage (Section 6) and may be put in three categories. The assessor should expect to find in the trial report a discussion of subgroup investigations conducted and their categorisation, indicating any differences from the plan outlined in the study protocol or analysis plan.

1. For a particular factor there is strong reason to expect an inconsistent response to treatment across the different levels of the factor to an extent that would make an overall assessment uninterpretable. In this case separate trials should usually be planned.

2. For a particular factor there is reason to consider it prognostic for outcome, or at least some biological plausibility or external evidence such that an inconsistent response might be observed.

In addition to factors used to stratify randomization, it would be expected that key demographic factors, including genomic factors, and factors related to the mechanism of action / pharmacology would be included in this category. In addition, careful consideration should be given to other factors that might plausibly be predictive for different response to treatment such as stage, severity or phenotype of disease, use of concomitant medications at baseline and possibly region, country, or centre.

Unlike the factors that might be categorised under point 1, it is not required that a formal proof of efficacy is available individually in all important subgroups in order to conclude on effects across the breadth of the trial population.

3. For a particular factor there is good argumentation why consistency of response to treatment is plausible or for which there is absence of pharmacological rationale and absence of clinical evidence from which to determine the plausibility of consistent drug effects.

This categorisation is important to prioritize subgroup assessments: those used for stratification would be investigated first, followed by an assessment of those where there is a priori suspicion that they may influence the treatment effect. Once these have been fully evaluated, further inspection of relevant subgroups should be undertaken for factors where there is no a priori rationale that the treatment effect should be modulated. These exploratory analyses are mainly undertaken for completeness, to provide additional re-assurance that assumptions at the planning stage hold true in the actual dataset. If factors are mentioned as standard subgroup analyses in EMA guidelines or scientific literature they should be prioritised for evaluation, and at least included amongst the exploratory analyses.

Some guidance on the weight of evidence needed in order to confirm credibility in different scenarios is given in the subsequent sub-sections. As discussed in Section 4.6, replication, biological plausibility and pre-specification are at least as important as the trial data itself. Pre-specification of subgroups of interest will take different forms. Subgroups intended for confirmatory inference will be pre-specified as part of the formal statistical testing strategy. Outside the formal testing strategy, particular interest in, or relevance of, a factor has been indicated where it is included to stratify randomisation or has otherwise been mentioned as identifying a key subgroup (category 2, above). Merely listing a factor for future subgroup analysis is not regarded as pre-specification of particular interest in that subgroup. If an inconsistency finding is observed, triggering further investigation, lack of pre-specification is most relevant where full discussion has been given a priori, in advance of trial results being reported.

It has to be noted that subgroup investigations are one aspect of the risk-benefit assessment following the primary assessment of efficacy. Further investigations into subgroups are thus not only triggered from inconsistent findings from the assessment of efficacy in subgroups but may also arise as part of better understanding safety signals. So, in instances, further subgroup analyses may be needed, and discussion may be important even if treatment effects in subgroups are completely consistent. In particular, if a target population is to be restricted based on a toxicity, it will be necessary to verify the therapeutic efficacy in the subgroup.

5.2. Assessment scenario 1: The clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated in the primary analysis population.

Exploration of consistency of treatment effects should include subgroup analyses in line with the methods outlined in Section 4.5. If the assessment of key subgroups has been well planned, nothing has arisen during the course of the trial to change the scientific assessment of plausibility and no evidence of inconsistent findings is apparent, then investigation may be regarded as being complete.

Inconsistent or extreme data in other exploratory subgroups, where the absence of a plausible link to the effects of treatment response can be confirmed by the assessor, could generally be disregarded unless the finding is replicated across more than one trial, or particularly is extreme, in which case (absence of) plausibility should be re-considered. If the discussion and pre-specification of key subgroups is incomplete, the assessor will by necessity need to take a more ad-hoc approach and will be forced to rely more on the observed data and their own judgement of plausibility without the benefit of the structure given above that limits the number of subgroups that are prioritised for examination.

If some evidence of inconsistency is observed for the effect in a subgroup (compared to the whole trial population) it may be considered credible, and hence subject to further sponsor evaluation and regulatory consideration, if there is either:

a. biological plausibility for an inconsistent effect in the direction expected. Credibility is particularly strong if evidence is replicated across multiple data sources. In submissions with only one trial in which the subgroup can be properly assessed, the fact that there is no opportunity to find replication within the phase III clinical program cannot be taken as an argument to ignore the signal, or;

b. replication of the inconsistent finding across multiple data sources. Analogously, credibility is particularly strong if there is also biological plausibility.

This credibility is further supported if tests of interaction are nominally statistical significant, or borderline significant, and if there is some evidence of treatment-by-covariate interactions across different endpoints that cover distinct aspects of the efficacy of treatment and are not directly correlated.

5.3. Assessment scenario 2: The clinical data presented are statistically persuasive in the primary analysis population but with therapeutic efficacy or risk-benefit which is borderline or unconvincing

Formal proof of efficacy is of paramount importance for the development of new drugs. However, even where this is robustly established, the clinical relevance of the treatment effect might not be convincing across the breadth of, or in a subset of, the population studied. Furthermore, having established therapeutic efficacy, it is also necessary to confirm a favourable risk-benefit. Even where the evidence for efficacy is strong from a statistical perspective, it might be of interest to identify a subgroup that has not been pre-specified as part of the confirmatory testing strategy, where efficacy and risk-benefit would be convincing. In fact there are at least three scenarios where this might arise:

1. Benefit in the all-randomised population is statistically significant but clinically not persuasive across the breadth of the trial population.

2. Benefit in the all-randomised population is statistically and clinically persuasive, but risks and uncertainties are present in the all-randomised population to the extent that a positive risk-benefit cannot be concluded across the breadth of the trial population.

3. Benefit in the all-randomised population is statistically and clinically persuasive, but risks and uncertainties are present in a subset of the population to the extent that a positive risk-benefit cannot be concluded in that subset.

Here there exists not only the problems of multiplicity, but also of selection bias since the identification of a subgroup of interest (and its complement) would commence once the data from the trial are known and the eye of the assessor and the applicant will be drawn to those findings that are most extreme. Neither the estimated treatment effect, nor the associated strength of evidence is necessarily reliable. For a subgroup to be considered credible all of the criteria below would usually apply. This list applies in principle irrespective of whether it is the company or the regulator that is specifying additional investigations of interest:

- External evidence should exist that the subgroup of interest is a well-defined and clinically relevant entity. Usually it would be expected that the respective subgroup has been considered when planning the trial (e.g. stratified randomisation or that it has been mentioned amongst the key subgroups) with an argument provided as to why it is a clinically relevant entity.
- A pharmacological rationale, or a mechanistically plausible explanation, should exist, why the drug under investigation could have different efficacy (or risk-benefit) in a sub-population and its complement (after considering also the scale of assessment).
- The treatment effect observed in the subgroup would usually be larger than in the all-randomised population. The totality of statistical evidence, based on individual trials and pooled analyses, should meet the same standards of evidence as would usually be expected for the all-randomised population indicating that the size of the treatment effect in the subgroup is substantial as compared to the variability of the problem.
- Replication of subgroup findings from other relevant trials (internal to the MAA or external trials that are relevant). A particular challenge exists in applications based on a single pivotal study since replication is a key component of credibility. In this instance the biological plausibility and the clinical trial data from the subgroup would have to be exceptionally strong.
- Whenever a treatment recommendation is to be based on a subgroup, it is mandated that risk-benefit should be carefully inspected in that subgroup and the relevance of safety data from the all-randomised population to the subgroup is carefully considered. Rare serious adverse events might, by chance alone, not occur in the subgroup of interest in the clinical trial, but should be considered in the risk-benefit assessment unless justified otherwise.

A close inspection of the baseline profiles of the subgroup is required regarding the comparison between treatment groups. Adjustment for differences might be required to ensure that an estimated treatment effect is not introduced or enhanced by baseline imbalances between groups.

Unless all the aforementioned requirements can be convincingly argued it may not be possible to restrict the licence to the subgroup and, if substantial concerns remain with the size of the treatment effect or the overall benefit/risk in the whole trial, licensing of the drug may not be possible.

5.4. Assessment scenario 3: The clinical data presented fail to establish statistically persuasive evidence in the primary analysis population but there is interest in identifying a subgroup where a relevant treatment effect and compelling evidence of a favourable risk-benefit profile can be assessed.

This relates to the use of a subgroup to rescue a trial that has formally failed, such that the primary objective of the trial could not be demonstrated (usually classified as p>5%, two-sided). From a formal statistical point of view, no further confirmatory conclusions are possible in a clinical trial where the primary null hypothesis cannot be rejected.

One or more additional trials should usually be conducted. In rare instances there may a basis for pursuing regulatory approval without conducting additional studies. This may be the case for a clinical setting where trials are not feasible to repeat or situations where trials are of considerable size (like in cardiovascular diseases) and even subpopulations may bear considerable amounts of randomised evidence that can be assessed for decision making about efficacy and a positive risk-benefit. Careful assessment of the overall available evidence has to be performed and it must be indicated that this type of exercise would be regarded as inadequate to support a licensing decision in most instances.

If nevertheless a positive licensing decision is, exceptionally, considered in this circumstance then Section 5.3 represents the minimum criteria that should be fulfilled. In addition, in such a situation, a clear rationale must exist as to why a properly planned trial has failed despite the drug being regarded as efficacious and why additional prospective studies to establish formal proof of efficacy are unfeasible or unwarranted.

6. Implications for study planning

Because knowledge about the experimental drug is limited when planning confirmatory trials it will only rarely be possible to fully pre-specify the assessment of subgroups. However, good study planning will utilise knowledge at the planning stage, to the extent possible, to improve the strategy for trial analysis and evaluation. These considerations support the proper discussion of risk factors and subgroups at the planning stage because considerations of plausibility are usually more convincing when made in advance of the trial, so that they are not influenced by knowledge of trial data.

6.1. Considering heterogeneity within a target population

During the planning of a clinical trial the discussion of known prognostic (differentiating groups with different clinical outcomes) and predictive (differentiating groups with different response to treatment) factors is one of the most important steps. A decision has to be made on the target patient group for the clinical trial. In particular, whether the criteria for inclusion or exclusion should restrict the patient population to, say, one level of a certain factor (e.g. biomarker positive), or whether use of the drug is intended in the full population under the assumption that patients in all subpopulations defined by the levels of the factor will benefit to a relevant degree from treatment (e.g. without regard to biomarker status).

A trial recruiting a broad patient population will help to support a broad indication statement but this will also increase the importance of investigating consistency of response to treatment. Where feasible, it is optimal for decision making if one or more large trials are available that have recruited the breadth of the target population. This increases the likelihood to learn about effect modifiers so that appropriate conclusions about efficacy and risk-benefit can be made. Assessors should expect to find discussion in the trial protocol of the expected degree of heterogeneity of the patient population in terms both of factors likely to be prognostic for outcome and those that are plausibly predictive of different response to treatment. Restrictions of a trial population to a sub-population of the target population should be justified, detailing whether restrictions are made due to safety concerns, anticipated lack of efficacy, or other operational considerations. It might be important to generate evidence that the drug under investigation does not benefit the whole trial population (e.g. for restrictions based on biomarkers thought to be predictive).

It must be recognised of course that knowledge of the treatment will increase as the confirmatory trials are conducted and hence, not all potential sources of heterogeneity can be predicted in advance of the trial. Learnings during the course of the trial will need to be reflected at the analysis stage irrespective of whether reflected in the study protocol or not. A blind review, where conducted, may be an opportunity to review available evidence and revise planning without knowledge of the outcome of the current trial.

6.2. A strategy for selection and definition of subgroups for assessment

After careful consideration of the trial population to be recruited for a planned trial, and after excluding sub-populations for which a strong reason exists to predict that the treatment effect will not be consistent, planning can proceed on the basis that the treatment effect is consistent but must foresee investigations into whether this assumption holds true. A strategy that assumes consistency of a population in terms of its likely response to treatment, without discussion and without planning and conducting further investigation, is not sufficient.

It will usually be appropriate that the recruited population reflects the epidemiology of the disease in the target patient group (external validity of the trial). The need to stratify the randomisation should be considered, firstly (in combination with blocking the randomisation) to reduce the risk of imbalanced allocation of patients from different factor levels to the treatment groups and, secondly, to indicate some of the factors that should be subject to scrutiny as to whether patients with different risk profile will have the same benefit from the use of the experimental drug. Specifically, an investigation is needed whether known prognostic factors used for stratification are also predictive for efficacy and / or risk-benefit of an experimental drug. Stratified randomisation, however, only tolerates a limited number of prognostic factors to be included into the model (PtC on adjustment for baseline covariates), and at the planning stage a thorough discussion with investigators is of importance to identify the most important prognostic and plausibly predictive factors. If properly discussed and documented, this helps to decide about the importance of certain subgroup findings at the assessment stage.

It is recommended that two levels of investigation should routinely be considered. The first level would include investigation of 'key subgroups', including factors used in stratification of the randomisation and other factors covered by definition number 2 in Section 5.1. Second, analyses of truly 'exploratory subgroups' should be planned for the spectrum of demographic, disease and clinical characteristics, including those factors covered by definition number 3 in Section 5.1. A listing and categorisation of planned subgroup analyses should be presented in the trial protocol.

A clear understanding of the relative importance of different subgroup analyses at the planning stage helps to minimise the *a posteriori* discussion and is promoted in order to generate sufficient evidence for assessment of important subgroups and, importantly, as an attempt to reduce the risk for erroneous conclusions about efficacy, or lack thereof, in subsets of the population. It is well understood that once trial results and estimated effects in subgroup analyses are known, these influence the manner in which plausibility, or lack thereof, is discussed and justified. This justifies that dialogue between sponsor and regulator is important at the planning stage to determining what subgroups are of interest for more detailed exploration in the trial analysis. In case this discussion is missing, erroneous or incomplete, regulatory assessment will necessarily become more *post hoc*.

In general, initial investigations should be planned that respect the form of the factor (e.g. binary, categorical, and continuous) in how far a certain factor is prognostic or predictive. The functional form (e.g. linear relationship) of a continuous covariate should be well understood so that proper classification for the definition of subgroups is possible. If trial results indicate that the treatment effect may vary according to the levels of a particular factor, subsequent investigations might need to be based on a categorisation (for a continuous factor) or collapsing categories (for a ordered categorical variable with a higher number of levels), because these categories should relate to criteria that might ultimately be used in product labelling or clinical decision-making. This possibility should be carefully considered at the planning stage, pre-specifying categories that might ultimately serve this purpose. Analyses investigating the choice of cut-off on the robustness of conclusions should also be planned.

The sample size for a trial is usually planned in order to meet trial objectives based on analysis of the whole population. However, it is recommended also to consider whether sufficient evidence will be generated in key subgroups, to support the assessment of consistency of treatment effects. This needs to be considered on a case-by-case basis, depending on the risk, a priori, to observe an inconsistent treatment effect. This can be of particular interest when considering countries (or regions).

ICH E9 requests centre to be included as a stratifying variable for multi-centre clinical trials. This was based on the experience that centre may be not only a logistic entity, but a strong prognostic factor summarizing the potential impact of differences in hospital settings and patient populations included. With multi-regional trials it is recommended to include country or region as a factor into the randomisation model and the analysis (PtC on adjustment for baseline covariates), because including centre often becomes impractical as few patients are recruited per centre, across a large number of centres. Country (or region) can be similarly important prognostic factors covering important intrinsic and extrinsic factors, including different attitudes to diagnosis, co-medication and other aspects of the concomitant setting. Although it is recommended to address these aspects by directly addressing important intrinsic and extrinsic factors in the randomisation and analysis, country (or region) remains a plausible factor for learning about the robustness of the treatment effect and potential factors that modulate the treatment effect and were unknown at the planning stage. Further details can be found in the ICH-E17 guidance on multiregional clinical trials. The need to provide sufficient evidence in a single region to support the notion that the overall effect applies to that region depends on how much knowledge about similarities or differences in intrinsic and extrinsic factors is available and in how far evidence exists that the clinical setting is different in different regions of the world. Consistent findings in regional strata strengthen an application based on a multi-regional clinical trial and may justify an increase in sample size to investigate treatment effects by region to avoid trials being inconclusive overall due to substantial regional differences that were not foreseen at the planning stage.

7. Including results from subgroup analyses in assessment reports and product information

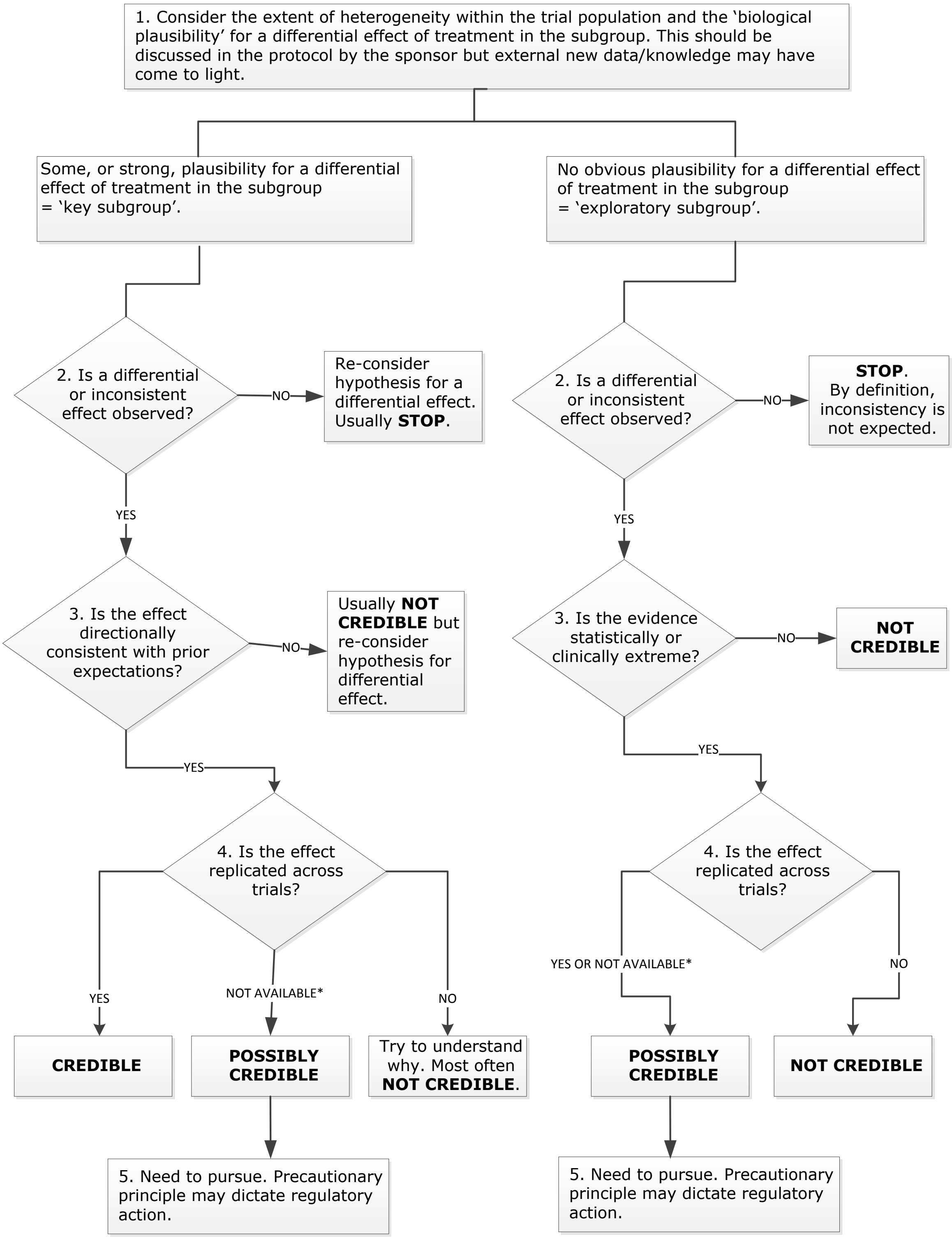
The assessment report should reflect the discussion provided by the applicant about heterogeneity in the target population and potential sources for inconsistency in treatment effects. The assessment should address whether that discussion is complete, balanced and contemporary. Trial results depicting consistency of effects should be presented in the assessment report. Any inconsistency should be discussed, in respect of credibility and, if found to be credible, in terms of clinical relevance, benefit-risk and communication through the Public Assessment Report or SmPC. If inconsistent results are found not be credible, this should be explicitly stated in the assessment report.

For the SmPC, it is not usually necessary to highlight consistency of effects in Section 5.1 of the SmPC, but this can be done where it represents important information for the prescriber. Where important uncertainty exists in a key subgroup despite overall positive findings, this can be expressed in a warning in Section 4.4 of the SmPC, with data presented in Section 5.1 if considered useful to the prescriber. Where a subgroup finding that is found to be credible indicates that therapeutic efficacy or positive risk-benefit is not established, or indeed that risk-benefit is negative, consequent labelling will be reflected in Section 4.1 or 4.3 of the SmPC, as appropriate.

A trial where a subgroup is used according to assessment scenarios 2 and 3, warrants extensive discussion in the AR, and justification why the results of the subgroup analysis are credible to the extent that they support a decision to approve the medicinal product. Where a target population is restricted to a subgroup of a trial population, findings in the overall trial population would usually be presented first in Section 5.1, followed by the findings in the subgroup. An explanation of why the subgroup findings can be considered credible can be provided if it can be communicated succinctly. If the subgroup finding is replicated in two trials, it is illustrative to display results of both trials to depict the degree of consistency of findings.

8. Annex

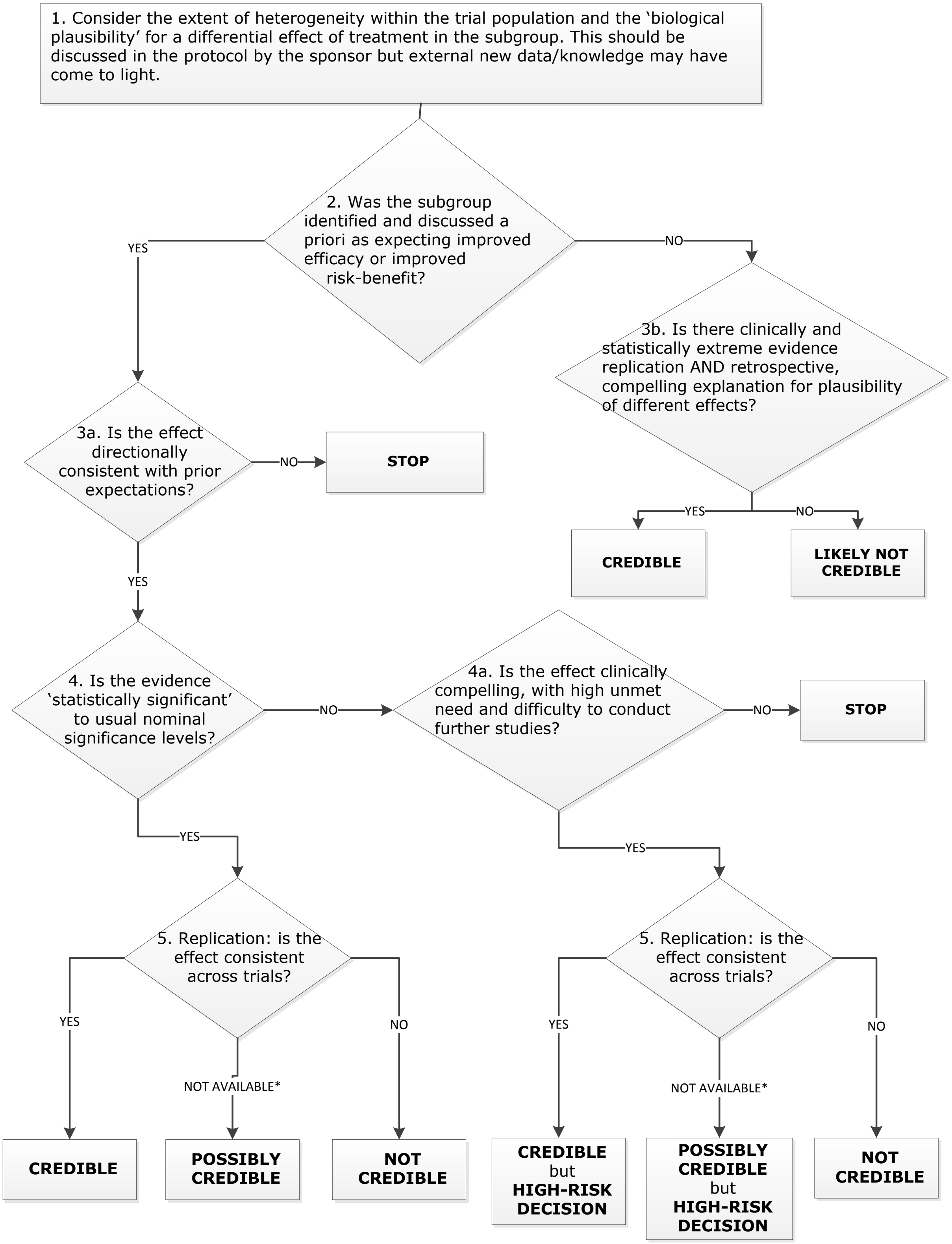
Annex 1 - Scenario 1 (Section 5.2) - establishing 'credibility' when considering 'consistency'



1. Consider the extent of heterogeneity within the trial population and the 'biological plausibility' for a differential effect of treatment in the subgroup. This should be discussed in the protocol by the sponsor but external new data/knowledge may have come to light.
Some, or strong, plausibility for a differential effect of treatment in the subgroup = 'key subgroup'.
No obvious plausibility for a differential effect of treatment in the subgroup = 'exploratory subgroup'.
2. Is a differential or inconsistent effect observed?
NO
Re-consider hypothesis for a differential effect. Usually **STOP**.
YES
3. Is the effect directionally consistent with prior expectations?
NO
Usually **NOT CREDIBLE** but re-consider hypothesis for differential effect.
YES
4. Is the effect replicated across trials?
YES
NOT AVAILABLE*
NO
CREDIBLE
POSSIBLY CREDIBLE
Try to understand why. Most often **NOT CREDIBLE**.
5. Need to pursue. Precautionary principle may dictate regulatory action.
2. Is a differential or inconsistent effect observed?
NO
STOP. By definition, inconsistency is not expected.
YES
3. Is the evidence statistically or clinically extreme?
NO
NOT CREDIBLE
YES
4. Is the effect replicated across trials?
YES OR NOT AVAILABLE*
NO
POSSIBLY CREDIBLE
NOT CREDIBLE
5. Need to pursue. Precautionary principle may dictate regulatory action.

*NOT AVAILABLE: Single large trial on the question of interest and insufficient external data.

Annex 2 - Scenario 2 (Section 5.3) - establishing 'credibility' to find a subgroup with clinically relevant efficacy or improved risk-benefit


1. Consider the extent of heterogeneity within the trial population and the 'biological plausibility' for a differential effect of treatment in the subgroup. This should be discussed in the protocol by the sponsor but external new data/knowledge may have come to light.
2. Was the subgroup identified and discussed a priori as expecting improved efficacy or improved risk-benefit?
YES
NO
3a. Is the effect directionally consistent with prior expectations?
NO
STOP
3b. Is there clinically and statistically extreme evidence replication AND retrospective, compelling explanation for plausibility of different effects?
YES
NO
CREDIBLE
LIKELY NOT CREDIBLE
YES
4. Is the evidence 'statistically significant' to usual nominal significance levels?
NO
4a. Is the effect clinically compelling, with high unmet need and difficulty to conduct further studies?
NO
STOP
YES
YES
5. Replication: is the effect consistent across trials?
YES
NOT AVAILABLE*
NO
CREDIBLE
POSSIBLY CREDIBLE
NOT CREDIBLE
5. Replication: is the effect consistent across trials?
YES
NOT AVAILABLE*
NO
CREDIBLE but HIGH-RISK DECISION
POSSIBLY CREDIBLE but HIGH-RISK DECISION
NOT CREDIBLE

*NOT AVAILABLE: Single large trial on the question of interest and insufficient external data.