EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

23 January 2014
EMA/CHMP/539146/2013
Committee for Medicinal Products for Human Use (CHMP)

Guideline on the investigation of subgroups in confirmatory clinical trials

DRAFT

Draft Agreed by Biostatistics Working Party	September 2013
Adoption by CHMP for release for consultation	23 January 2014
Start of public consultation	03 February 2014
End of consultation (deadline for comments)	31 July 2014

Comments should be provided using this template. The completed comments form should be sent to Biostatistics@ema.europa.eu

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

Guideline on the investigation of subgroups in confirmatory clinical trials

Table of contents

Executive summary

1. Investigation into the effects of treatment in well-defined subsets of the trial population is an integral part of clinical trial planning, analysis and inference that follows the inspection of the primary outcome of the trial. The guideline should assist in the planning and presentation of these investigations and in the understanding of factors to be discussed when considering the credibility of findings.

2. The more homogeneous the population studied, in terms of baseline risk and in terms of response to treatment, the lower the importance of exploratory subgroup analyses for regulatory assessment. The more heterogeneous the study population, the greater the importance of subgroup analyses to check that the estimated overall effect is broadly applicable and supports assessment of risk-benefit across the breadth of the proposed indication. Exploration of heterogeneity should include covariate-adjusted analyses and subgroup analyses.

3. Methodological complications related to multiple analyses mean that exploratory investigations into effects in subsets of the trial population must proceed with caution taking into consideration all available evidence, not only the point estimates from individual subgroup analyses. Despite the statistical complications, not investigating, or ignoring results of subgroup analyses could also lead to incorrect decisions.

4. 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 the course of disease and those that are plausibly predictive of differential response to treatment. A strategy that simply assumes homogeneity of a population in terms of its likely response to treatment, without discussion and without further investigation, is not sufficient. Analogously, it is not sufficient to dismiss all subgroup findings that indicate heterogeneity of response as being spurious. The benefits of this additional discussion are to maximise the *a priori* discussion of the importance of certain subgroups and thus to minimise the *a posteriori* discussion in an attempt to promote rational consideration of subgroups and to reduce the risk for erroneous conclusions. Done properly, this should minimise the need for data-driven investigations, relying instead on a well-reasoned pre-specified strategy.

5. Consistency of findings in relevant subgroups needs to be discussed in the analysis report: Forest plots graphing the treatment effect in a series of subgroups and statistical methods to assess heterogeneity of treatment effects estimated in subgroups play an important role for the provision of signals as to whether the overall treatment effect applies to the full trial population. Clinical and pharmacological knowledge are needed to evaluate the credibility and relevance of signals that are generated. A number of factors influence the credibility of a subgroup finding, including ‘biological plausibility’ and replication of evidence as well as the strength of evidence from the trial(s). Credible explanations for heterogeneity should be sought. Multiple analyses and data presentations may be required to properly inform an assessment of credibility.

6. A strategy for assessing the credibility of subgroup findings is presented for different situations that are commonly encountered. Key considerations for switching from the all randomised population to a subgroup for risk-benefit decision making are given. Subgroup analyses will not usually rescue failed trials.

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 and, 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. Assessment of these trials usually proceeds through investigation of the treatment effect on the primary and secondary outcome measures in the whole population, and through investigation into the safety profile of the experimental drug. For confirmatory trials, 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 a regulatory assessment of relevant subgroups with regard to relevant endpoints during assessment of Marketing Authorisation Application (MAA), as a second step subsequent to inspection of the primary and secondary trial outcomes on the whole trial population.

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 variability in response 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.” 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.

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. However, whether the true treatment effect is 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 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 controversial 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. These considerations for assessment 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 does not dictate any particular practical solutions in respect of statistical methodology for estimating or testing the treatment effect in subgroups of the trial population.

A differentiation is made between investigation of a subgroup as part of the confirmatory testing strategy and investigation of subgroups in a more exploratory manner. Whilst a number of the considerations outlined in this document will apply to the former, this 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 not discussed here. The guiding principles and examples for multiple-testing procedures that control the overall false positive rate are described in the respective guidance (PtC on multiplicity issues in clinical trials).

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 scenarios are focussed more on a *post hoc* restriction to the breadth of the target population:

- Scenario 1: The clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated globally. It is of interest to verify that the conclusions of therapeutic efficacy (and safety) apply consistently across subgroups of the clinical trial population.

- Scenario 2: The clinical data presented are overall statistically persuasive but with therapeutic efficacy or benefit/risk which is borderline or unconvincing and it is of interest to identify post-hoc a subgroup, where efficacy and risk-benefit is convincing.

- Scenario 3: The clinical data presented fail to establish statistically persuasive evidence 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.

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

The paper does not try to describe the appropriate regulatory decision in any particular circumstance. Whilst the decision-making problem differs, the principles outlined in the document apply equally to:

- subgroup investigations for efficacy or safety variables;

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

There may also be interest in criteria for determining inclusion of information in subgroups to Section 5.1 of the Summary of product characteristics. This is predominately a consideration of whether information on subgroups would be useful to the prescriber but, depending on the circumstance, criteria outlined in Section 6 may also be useful for a determination of whether the evidence generated may be considered reliable for presentation.

3. Legal basis and relevant guidelines

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

Points to consider on adjustment for baseline covariates (CPMP/EWP/2863/99)

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)

Concept paper on the need for a guideline on the use of subgroup analyses in randomised controlled trials (CHMP/EWP/117211/2010)

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

4. General considerations

4.1. Definition of a subgroup

The term 'subgroup' will be used to refer to a subset of a clinical trial population. The term 'sub-population' will be used to refer to a subset of the population described by the targeted therapeutic indication. Patients excluded from a particular subgroup are described as the complement subgroup.

In relation to a clinical trial, a subgroup can be defined as any subset of the recruited patient population that fall into the same category (level) with regard to one or more descriptive factors. These factors and the categorisation of patients will usually be identifiable prior to randomisation based on both intrinsic and extrinsic factors (see ICH E5), including demographic characteristics (including genetic or other biomarkers), disease characteristics including severity or (pheno)type of disease and clinical considerations (e.g. use of concomitant medications, region or centre). Post-baseline covariates may be affected by treatment received and will not usually be appropriate to define subgroups for investigation, in particular where the purpose of the investigation is to draw conclusions on the sub-populations in which it is appropriate to initiate treatment.

Factors can be dichotomous (e.g. male / female), categorical (e.g. region), ordered categorical (e.g. disease score at baseline) or continuous (e.g. age). Some categorisations of subgroups will be naturally defined (e.g. male / female). Others will need more careful consideration, in particular for factors based on continuous measures, or where pooling across multiple levels of a single factor is needed (e.g. centre or region). Cut-off points for continuous measures and groupings for categorical factors should generally be pre-specified and justified, considering the amount of information likely to be available for each level of the defining factor but, importantly, considering also the relevance as a threshold for decision making in clinical practice.

Most investigations will consider subgroups identified on the basis of a single factor. Subgroups defined on multiple factors (e.g. females aged >65) may be of interest on occasion but for simplicity, the descriptions in this document will make reference to a subgroup defined on a single factor (e.g. gender categorised as male and female), and this will suffice for most investigations. The risks described in this document around analysis and interpretation of subgroup analyses are exacerbated by also considering subgroups based on multiple factors, though the need for this more complex type of investigation cannot be excluded. Another type of investigation is to categorise patients according to a 'risk score' based on their profile considering multiple prognostic or predictive characteristics. If the risk score is informative, this may represent a worthwhile investigation into understanding response to treatment. The risk score itself may serve as a factor by which subgroups of patients may be defined in addition to a categorical factor against which response to treatment may be modelled.

For factors where categorisation depends on a biological measure there is a risk of misclassification, in particular due to measurement or diagnostic error. Information will be needed to quantify the influence of this risk on the classification of patients into subgroups and on the inferences that can reliably be made therefrom.

4.2. Problems with conducting multiple subgroup analyses

The heterogeneity of a patient population included in a confirmatory clinical trial will vary depending on the specific therapeutic indication, the inclusion / exclusion criteria of the study, factors important for the prognosis of the disease course, the experimental medicinal product under study and the countries / regions selected for conducting the trials. The more homogeneous the population studied, the lower the importance of subgroup analyses is likely to be in regulatory assessment (though as indicated in PtC on Multiplicity Issues in Clinical Trials a narrow population may have implications for generalisability of trial outcome and a consequent restriction to the indicated population). The more heterogeneous the study population, the greater the importance of subgroup analyses to check that the estimated overall effect is broadly applicable.

The problem of exploring subgroups is closely related to the problem of multiple testing. Initial inference should be based on analysis of a primary endpoint in a primary analysis population, usually the Full Analysis Set, supported by analysis of secondary endpoints in the primary analysis population. When multiple subgroups are considered, problems relating to multiple testing arise, specifically the increased probability of false-positive findings (subgroups where effect is concluded to differ from the primary analysis population when in fact it does not) which, if interpreted incautiously, will lead to erroneous conclusions. This supports the position outlined in ICH E9 that “any conclusion of treatment efficacy (or lack thereof) or safety based solely on exploratory subgroup analyses are unlikely to be accepted.”

To extend this, it is basic knowledge in statistics that repeated testing the same data for different variables or different subgroups can lead to false-positive conclusions unless proper consideration is given to multiplicity adjustment at the planning stage of the trial. Specifically even if a medicine is associated with no benefit, if a large number of subgroups are examined it will inevitably appear disproportionately beneficial in one or more subgroups. Conversely, if a medicine is associated with benefit, it will, by chance alone, appear not to work or even harm in some category or categories of patient. This is often quoted as a reason to ignore or dismiss investigation of effects in different subgroups but, critically, this ignores an examination of the underlying hypothesis that effects across different subgroups will be homogenous. This will not always hold. There is a tension therefore between the widely appreciated statistical phenomenon related to multiplicity and the issues outlined above relating to the potential heterogeneity of a target population and potential heterogeneity of response to treatment. Despite the statistical limitations, not investigating, or ignoring results of subgroups could also lead to incorrect decisions.

This phenomenon is not only relevant to Phase III trials of course. 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.

4.3 Basic considerations for investigation of heterogeneity, analysis of subgroups and associated data presentations

Analysis of subgroups would proceed only after confirmatory testing on the primary analysis population is complete. ICH E9 (Section 5.7) indicates that analyses of subgroups should proceed first through the addition of interaction terms to the statistical model in question.

A key question here is how to parameterise the factors for use in treatment*covariate interaction tests. In general, the form of the factor (e.g. binary, categorical, continuous) should be respected in the initial subgroup investigations. In particular, initial investigations of continuous factors should be performed without dichotomisation or categorisation of the factor if possible since this would result in loss of information. However, caution needs to be taken when specifying the functional form (e.g. linear relationship) of a continuous covariate since mis-specification may lead to misleading conclusions with respect to interactions, and in instances where the relationship is unclear it may still be wise to start investigations into heterogeneity of subgroup findings by categorising a continuous factor. A justification should be provided for the functional form selected. If a signal for heterogeneity effect is observed, subsequent investigations might also involve categorising or collapsing factors that are measured on the continuous scale or that have a higher number of levels so that the investigations presented relate to criteria that might ultimately used in product labelling or clinical decision-making. If categorising a continuous covariate, sensitivity analyses using different cut-offs should routinely be performed. Some thought may be given in the clinical trial protocol on how this might proceed.

The test for interaction will be associated with a p-value. Although still common practice, the sole reporting of a p-value from a test for interaction cannot be considered adequate. It is recommended to add estimates and corresponding confidence intervals, and graphical representations may prove particularly useful in more complicated settings. These additional statistics and data presentations can give a guide as to what the data is capable of showing with regard to differences in effects among subgroups and what can reasonably be excluded by the available data in terms of the size of the interaction.

Tests of interaction on important variables can be complemented by additional exploratory subgroup analyses within relevant subsets of the trial population, or within strata defined by the covariates. It is common to present exploratory subgroup analyses for a range of factors. Presentation of results should include estimates and confidence intervals in the context of baseline values. Whenever a subgroup analysis is displayed, the analysis of the complement subset should also be displayed. For continuous variables, plots should be presented to characterise how the estimated effect of treatment changes over the range of the factor. Where dichotomous or categorical variables are used to define subgroups, it would be expected to see results presented in Forest plots. When interpreting Forest plots it is tempting to find reassurance in directional consistence of estimated effects. The reviewer is cautioned that the subgroup presentations are not independent and do not provide mutually exclusive confirmation of findings. Also, if in one subgroup the treatment effect is larger than the average treatment effect, the complementary subgroup will by necessity worse than the average treatment effect.

A key question here is the scale on which to assess the influence of covariates on the estimated treatment effect. 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). Commonly it is more realistic to expect homogeneity of treatment effect on the relative scale (e.g. patients with mild disease at baseline do not have the capacity to experience beneficial effects as large as might be possible in patients with severe disease at baseline). Contrary to this, absolute effects tend to be more intuitive for understanding the magnitude of effect and are more commonly used in risk-benefit decision-making. Even where the effects of a medicine 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, with supplementary analyses presented on the complementary scale for those covariates or subgroups that become important for the risk-benefit decision. The assessor needs to be aware of the scale being used and to question whether additional analyses would be informative.

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 heterogeneity. These methods may be presented by sponsors but the underlying assumptions must be carefully considered and discussed.

It might be questioned whether the 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 serve as an indicator for further exploration, adjustment would be counter-intuitive and is not recommended. The fact that multiple subgroups are examined, and the number of subgroups examined, is of course a key matter for consideration during assessment and regulatory decision-making.

In summary, the price to be paid for the inclusion of a broad patient population into the phase III clinical trials is the need to check that the overall treatment effect applies to relevant subgroups of the patient population. It may well be that the treatment effect is not the same in all subgroups or may depend on a continuous covariate. This is called heterogeneity or treatment-by-covariate interaction. In case the treatment effect in relevant subgroups of the patient population is different, a separate benefit/risk assessment may be required. While it is important to understand, how certain patient characteristics impact on the overall treatment effect and to model the treatment-by-covariate interaction, or to assess heterogeneity, it is in the end the benefit/risk assessment for some subgroups that is needed to describe the efficacy of a drug appropriately.

4.4 Key considerations that underpin assessment of subgroups

Whilst the observed clinical trial data are important, the utility is influenced by many factors, not least the size of the trial and the relative prevalence of the subset of interest in the trial population. Analysis of a subset of the population that is not well represented, at least in relation to the variability and effect size of the outcome measure of interest, will not provide informative data for assessment of heterogeneity. A number of key additional considerations are outlined below; their relevance to planning is described in Section 5 and their relevance to assessment in Section 6.

a. The **heterogeneity** of the clinical trial population; the more heterogenous the population, the more important the investigation of **consistency (homogeneity)** of effects in well-defined subgroups. **Consistency** of effect is most relevant where the clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated globally and it is of interest to verify that the conclusions of therapeutic efficacy and safety apply across subgroups of the clinical trial population.

b. **Biological plausibility**; a concept describing the extent to which a particular effect (in this case a differential effect of treatment in a particular subgroup of patients) might be predicted, or might have been expected, based on clinical, pharmacological, and mechanistic considerations, and considerations of other relevant external data sources (often referred to collectively as 'Biological Plausibility'). Plausibility is primarily a clinical and pharmacological judgement and is usually not a directly quantifiable or measurable concept. Ideally, those factors where biological plausibility exists will be pre-specified for use as stratification factors or as being of particular interest for exploratory investigations in the clinical trial protocol.

c. **Replication** of evidence; the possibility to examine an effect of a particular covariate, or effect within a particular subgroup, from multiple sources of relevant clinical trial data.

5. Issues to be addressed at the planning stage

5.1. Considering heterogeneity within a target population

During the planning of a clinical trial the discussion of known prognostic (differentiating groups with different clinical progression) 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 from treatment (e.g. without regard to biomarker status). Similarly, the inclusion and exclusion criteria will define the breadth of the population recruited with regards to other clinical, demographic and disease characteristics. A broad patient population will tend to support a broad indication statement but will also increase the importance of investigating heterogeneity of response to treatment.

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 response and those that are plausibly predictive of different response to treatment. 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. Consistent with the text quoted below from the CHMP PtC on multiplicity issues in clinical trials *"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. Analyses of these important subgroups should be a regular part of the evaluation of a clinical study (when relevant), but should usually be considered exploratory, unless there is a priori suspicion that one or more of these factors may influence the size of effect"*, factors that define a target population may be put in three categories:

1. For a particular factor there is strong reason to expect a heterogeneous response to treatment across the different levels of the factor. In this case separate trials should usually be planned.

2. For a particular factor there is at least some biological plausibility or external evidence such that a heterogeneous response might be hypothesised. In this case it is relevant to discuss and plan for an assessment of consistency of effects.

   In addition to factors used to stratify randomisation, it would be expected that key demographic factors, including genomic 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 and possibly region, country, or centre, see section 5.3.

Unlike the factors that might be categorised under point 1, it is not usually 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. It would, however, be prudent to design the trial accordingly such that a sufficient number of patients are recruited to the subgroup to ensure an estimate of effect that can be made with reasonable precision so that the applicant is able to substantiate therapeutic efficacy and a favourable risk-benefit in important subgroups.

3. For a particular factor there is good argumentation why homogeneity of response to treatment is plausible.

A strategy that assumes homogeneity of a population in terms of its likely response to treatment, without discussion and without 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 to reduce the risk of imbalanced allocation of patients from different factor levels to the treatment groups, and, secondly, to indicate at the planning stage that whether patients with different risk profile will have the same benefit from the use of the experimental drug is a question to be examined. Stratified randomisation, however, only tolerates a very limited number of prognostic factors to be included into the model (see also 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 predictive factors. This discussion should impact on the assessment strategy and evaluation of subgroup findings.

5.2. Prioritising the exploratory analyses

Investigation of homogeneity of response should always be planned, but the associated multiplicity needs to be considered. It is recommended that two levels of investigation should routinely be considered, excluding any subgroups planned as part of the confirmatory testing strategy. 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, truly exploratory analyses should be planned for the spectrum of demographic, disease and clinical characteristics, including those factors covered by definition number 3 in Section 5.1.

The benefits of this additional discussion and clarity are to maximise the *a priori* discussion of the importance of subgroups and thus to minimise the *a posteriori* discussion in an attempt to reduce the risk for erroneous conclusions about efficacy in subsets of the population. Done properly, this should minimise the need for data-driven ‘for-cause’ investigations, relying instead on a well reasoned pre-specified strategy. It must be recognised however that this leads to a potential disincentive for the sponsor to properly plan the investigation of subgroups, arguing instead that no relationships between baseline factors and response to treatment are plausible and therefore that no key subgroup analyses are needed and any findings of concern in any subgroup analysis must be ascribed to chance.

It is therefore clear that the assessor will have a key role in determining what subgroups are of key interest for more detailed exploration. Again this would, ideally, be discussed at the planning stage of the trial. By necessity, if the sponsor has not provided well-reasoned arguments and a comprehensive strategy for analysis, regulatory assessment will become more *post hoc*. In addition, factors for which there is absence of evidence of scientific knowledge to make a classification will necessarily need to be considered *post hoc*. Whilst 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, it is re-iterated that a fully comprehensive discussion on biological plausibility will not always be possible prior to the Phase III trial. Hypotheses for heterogeneity of response might emerge as scientific knowledge about the drug or drug class accumulates.

In general studies are planned for a certain primary endpoint in the full population. In case heterogeneity of the patient population is foreseen at the planning stage increases of the total sample size of the trial may be justified in order to allow the assessment of the consistency of the treatment effect in relevant subgroups. Alternatively a decision could be made to refine the full population to an extent that heterogeneity of the treatment effect in different subgroups is less likely (see also the respective discussion in Section 5.3).

In summary, pre-planning helps to reduce the risk that abundant analyses are requested or performed, but assessors have to recognize that accumulating information may necessitate further investigations into subgroups of a trial. Indication for harm in subgroups should be understood in the same way as signal generation during assessment: findings should not be dismissed as pure chance findings at the outset, but carefully assessed for their plausibility and relevance, before they are either classified as requiring further observation or dismissed as a chance finding.

5.3. Country or region used for pre-stratification

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 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. In recent years the experience has grown that country (or region) can be similar 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 the respective variables, country (or region) as entity for checking the context-sensitivity (or robustness) of the treatment effect is of importance to regional drug licensing bodies and as a plausible source for learning about the robustness of the treatment effect.

As with other factors, whether or not trials should be planned only to meet their primary objective or whether consideration should be given to how much of a trend for a positive treatment effect should be available for the results in countries (or regions) should depend on how much knowledge about similarities or differences in intrinsic and extrinsic factors is available and in how far evidence exists that the concomitant setting is different in different regions of the world. Consistent findings in regional strata strengthen such an application 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.

5.4. Documenting the exploratory analyses

The Clinical Trial Protocol and Statistical Analysis Plan are used to document key aspects of clinical trial design, conduct, analysis and reporting. Statistical approaches relating to conduct and analysis are pre-specified in these documents prior to the trial commencing and updated through formal amendments during the course of the trial and, if appropriate, in response to a blind review (see ICH E9).

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. As described above, the trial documents should also discuss, identify and prioritise some key factors and subgroups for exploratory analysis from the background of indication specific knowledge (e.g. gender in cardiovascular disease). In addition, stratification factors may have been identified for the randomisation and indicate that these are important (prognostic or predictive) covariates for statistical modelling.

It is important to note that these different types of reference in trial documents do not have the same weight in terms of pre-specification. This is important when considering whether emphasis may be switched from the FAS to a subgroup. Concluding that a subgroup has been pre-specified should be reserved for the use of a subgroup for its intended purpose. For example, a subgroup identified as exploratory has, by definition, not been pre-specified for positive confirmatory inference, neither have subgroups classified by stratification factors, though it has at least been recognised *a priori* that these are of some importance and balance of randomisation is addressed.

6. Issues to be addressed during assessment

6.1. Assessing 'consistency' ('homogeneity') and 'inconsistency' ('heterogeneity')

As outlined above, there is justification to carefully assess important subsets of the patient population within a Phase III clinical trial and to search for descriptive consistency of treatment effects estimated in subgroups (Scenario 1, Section 6.3, below). It is repeated that both the subgroup of interest and its complement should be routinely presented. When checking consistency of the treatment effect in subgroups beyond those that have also been used to stratify randomisation, baseline balance of important risk factors is important and should be checked, as well. If there is indication that this is violated, an adjusted analysis should be provided before drawing conclusions.

Historically, it has been argued that the absence of statistically significant treatment-by-covariate interactions implies consistency of the treatment effect in the studied population. This is not accepted. It is a general principle that absence of statistical significance should not be taken to imply equality or consistency. It has also been argued, say in a superiority trial, that observing all points estimates to be going in the same direction, an absence of qualitative interaction, is adequate to establish consistency. To require only absence of statistical significance in an interaction test, or only directional consistency, would not be sufficiently sensitive filters to detect differences of potential interest. Instead investigations into the homogeneity of the treatment effect in relevant subsets of the study population may be likened to the assessment of safety of new drugs: in both situations statistical tests can be of help to "flag" potential problems, but descriptive assessments and clinical considerations need to be combined to evaluate potential signals.

There is no widely accepted definition for consistency. Presented below are some working definitions to use when reviewing a series of subgroup analyses. Inconsistencies in one or more subgroups might give rise to concern about the applicability of the overall treatment effect, if the subgroup analysis result is found credible. The assessment of consistency is different from that of credibility, see Section 6.2. It is recommended (see Annex 1) that assessment of credibility is based primarily on biological plausibility and external evidence, but it is also appreciated that an investigation of results within subgroups is an important part of data review. It is important to recognise that the mere identification of inconsistency without full consideration of other important factors outlined in this paper should not generally be used as a basis for regulatory action, for example with regard to restricting the licence.

Some statistical measures have been identified for the purpose of assessing heterogeneity (e.g. I^2 test or chi-squared test, the heterogeneity test statistic Q from a generalised Breslow & Day test). These are not commonly presented in the analysis of confirmatory clinical trials and experience in their utility is limited. Criteria to draw inferences from these tests such that they are sensitive and specific for detecting heterogeneity are not well defined.

Visual inspection of a Forest plot that describes the results for multiple subgroup analyses can help, specifically where interrogation of subgroup analyses is to flag subsets of the trial population for further inspection and consideration (see Section 6.2 and Annex 1). However, here too, a formal rule for interpretation that is both sensitive to detect heterogeneity of potential interest and specific 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 and the overall number of subgroups (the more groups the more likely to observe one or more groups with extreme findings, by chance). Further research into statistical methods to trigger inspections into subgroups of a confirmatory clinical trial is needed.

A reassuring pattern of results is where all point estimates from subgroup analyses are rather similar to the overall effect with all confidence intervals overlapping with the confidence interval for the overall effect. This will rarely occur and it is worth repeating that estimates will differ by chance alone, or by imbalances in subgroup characteristics. Two further scenarios are described for purpose of illustration based, for convenience, on a superiority trial, with effects in the positive direction on the scale of measurement being desirable. First consider a trial within which the overall effect is estimated precisely (in relation to the effect size) such that both the point estimate and the lower confidence bound are well away from the point of no difference. For subgroups where the effect can also be estimated with reasonable precision (such that the width of the relevant confidence interval is up to approximately 2x or 3x as wide as for the overall effect) a flag for inconsistency would be an estimated effect that is outside the span of the CI for the overall effect such that the confidence intervals for the subgroup and the overall effect are largely non-overlapping. Of course, this flag for inconsistency does not speak to other aspects of interpretation; in particular the estimated effect in the subgroup may still indicate clinical relevance (and indeed be statistically significant). For other subgroups estimated with lower precision, and in particular for subgroups of low size (and consequently with wide confidence intervals) estimated effects well removed from the estimate and confidence interval for the overall effect may give some cause for concern but confidence intervals that largely overlap the confidence interval for the overall effect give little information and it must be recognised that there will be subsets of the trial population where the trial simply provides too little information for inference. For these groups an assessment of consistency may not be possible and the majority of assessment will be based on considerations relating to biological plausibility for a differential effect and other sources of evidence.

Secondly consider a trial for which the effect is less precisely estimated (in relation to effect size) such that the confidence interval for the overall effect approaches the point of no difference. Usually this will not be a single pivotal trial since this would not constitute sufficiently extreme evidence of efficacy and so replication, or otherwise, is an important consideration (see Section 6.2). In terms of a flag for potential inconsistency in subgroup analyses the above rules would also apply, noting that in this case the estimated effects in subgroups would now be negative, but a flag for further consideration may also apply to subgroups where effects are reduced in comparison to the overall effect in the region of an effect considered to be of limited clinical significance and where confidence intervals are only partially overlapping.

6.2. Defining ‘credibility’

As indicated in Sections 5.1 and 5.2, plausibility will be considered in the absence of trial data at the planning stage of the trial. Based on the clinical trial data generated and other data or knowledge emerging during the course of the trial, the credibility of findings of interest in subgroups must then be re-considered.

The assessor must consider all evidence that can be brought to bear on the problem including the key considerations outlined in Section 4.4 above in addition to the clinical trial data. Strong biological plausibility, or absence thereof, or replication of evidence may well contribute greater weight to the overall assessment as the pattern of data observed across the range of subgroup analyses presented. In particular, having two or more relevant sources of evidence is of great assistance to interpretation. Where two or more trials can be interrogated on effects in a particular subgroup the weight of evidence from directly relevant clinical trial data rather than from external evidence of lesser relevance or arguments of biological plausibility increases. Evidence for differential effects in subgroups that are replicated across available clinical trials can be compelling irrespective of the fact that it may be larger or smaller than the (average) effect that is overall observed in this trial. This holds true even in the absence of a plausible mechanistic explanation. Conversely, an inconsistent finding in one trial is more readily disregarded if evidence from one or more other trials 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 the subgroup 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 the assessment of internal consistency in a single pivotal trial.

Of course, when multiple trials are available that bear on the same question, a pooled analysis is possible. The possibility to look at two or more sources of evidence provides stronger evidence on the question of consistency, or otherwise, of effect in a subgroup than the mere presentation of a more precise estimate obtained through pooling of the respective subgroups from two trials. However, sources of evidence should always be presented separately, as well (see PtC on application with 1.meta-analyses, 2.one pivotal study).

The sponsor may use absence of pre-specification as an argument for lack of credibility, in particular for adverse findings. Because there may exist a disincentive to specify some key subgroup analyses, the absence of pre-specification, in particular where accompanied by absence of a comprehensive discussion, does not in itself constitute reason to ignore results in a particular subgroup.

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 with the knowledge from other trials in the development program or in the same class, pharmacology and/or mechanistic considerations. 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 available.

6.3. Scenario 1: The clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated globally. It is of interest to verify that the conclusions of therapeutic efficacy and safety apply consistently across subgroups of the clinical trial population.

Exploration of heterogeneity should include covariate-adjusted analyses and subgroup analyses. If well-reasoned in the trial protocol, assessment of subgroups may be based primarily on the pre-specified strategy described in Sections 5.1 and 5.2 above and be followed, for completeness, by a review of other exploratory analyses.

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 extreme, in which case plausibility should be re-considered. If the discussion and pre-specification of key subgroups is incomplete then 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 and the inconsistency is in the direction expected. Credibility is particularly strong if evidence is replicated across multiple data sources, though in submissions with only one trial in which the subgroup can be properly assessed, the precautionary principle dictates that replicated evidence cannot be required to confirm credibility of an untoward effect of the experimental treatment.

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 statistically significant, or borderline significant, and if there is some evidence of treatment-by-covariate interactions across different endpoints (notwithstanding correlation between endpoints; the stronger the correlation, the less credibility is enhanced).

Subgroup findings that do not meet the above criteria will not usually be considered credible. If there is evidence of heterogeneity / inconsistency and the findings are regarded as credible because of the biological plausibility, directional consistency and/or replication, the magnitude of the estimated effects, and the uncertainty, must be set in the context of a risk-benefit consideration.

6.4. Scenario 2: The clinical data presented are overall statistically persuasive but with therapeutic efficacy or benefit/risk which is borderline or unconvincing and it is 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.

Formal proof of efficacy is of paramount importance for the development of new drugs. However, drug development does not rely on one clinical trial only and situations may exist where there is interest in drawing positive conclusions about efficacy of the drug under investigation at least in a subset of the population that has been investigated in the clinical trial programme.

This scenario would usually arise because:

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 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. Therefore, and because the aim of this exercise is to draw a positive conclusion for marketing authorisation from a clinical development programme that has not provided persuasive evidence from a statistical and clinical point of view, the level of evidence needed to establish credibility is arguably higher. 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.
- A pharmacological rationale, or a mechanistically plausible explanation, should exist, why a certain drug or treatment could have different efficacy (or benefit/risk) in a sub-population and its complement (considering also the scale of assessment).
- The estimated effect of treatment in the subgroup would usually be more pronounced in absolute terms (i.e. indicating a greater benefit) 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.

Usually it would be expected that pre-stratification (i.e. stratified randomisation) clearly has identified the respective subpopulation, or that it has been mentioned amongst the key subgroups. If the factor of interest has not been used to stratify the randomisation, a close inspection of the baseline profiles of the subgroups identified between treatment groups, and eventually adjustment for differences, is needed. Whenever a treatment recommendation is to be based on a subgroup, it is mandated that benefit/risk should be carefully inspected in that subgroup and the extrapolation of safety data from the all-randomised population to the subgroup is carefully considered.

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, licensure of the drug may not be possible.

6.5. Scenario 3: The clinical data presented fail to establish statistically persuasive evidence but there is interest in identifying a subgroup, where a relevant treatment effect is evident and there is compelling evidence of a favourable risk-benefit.

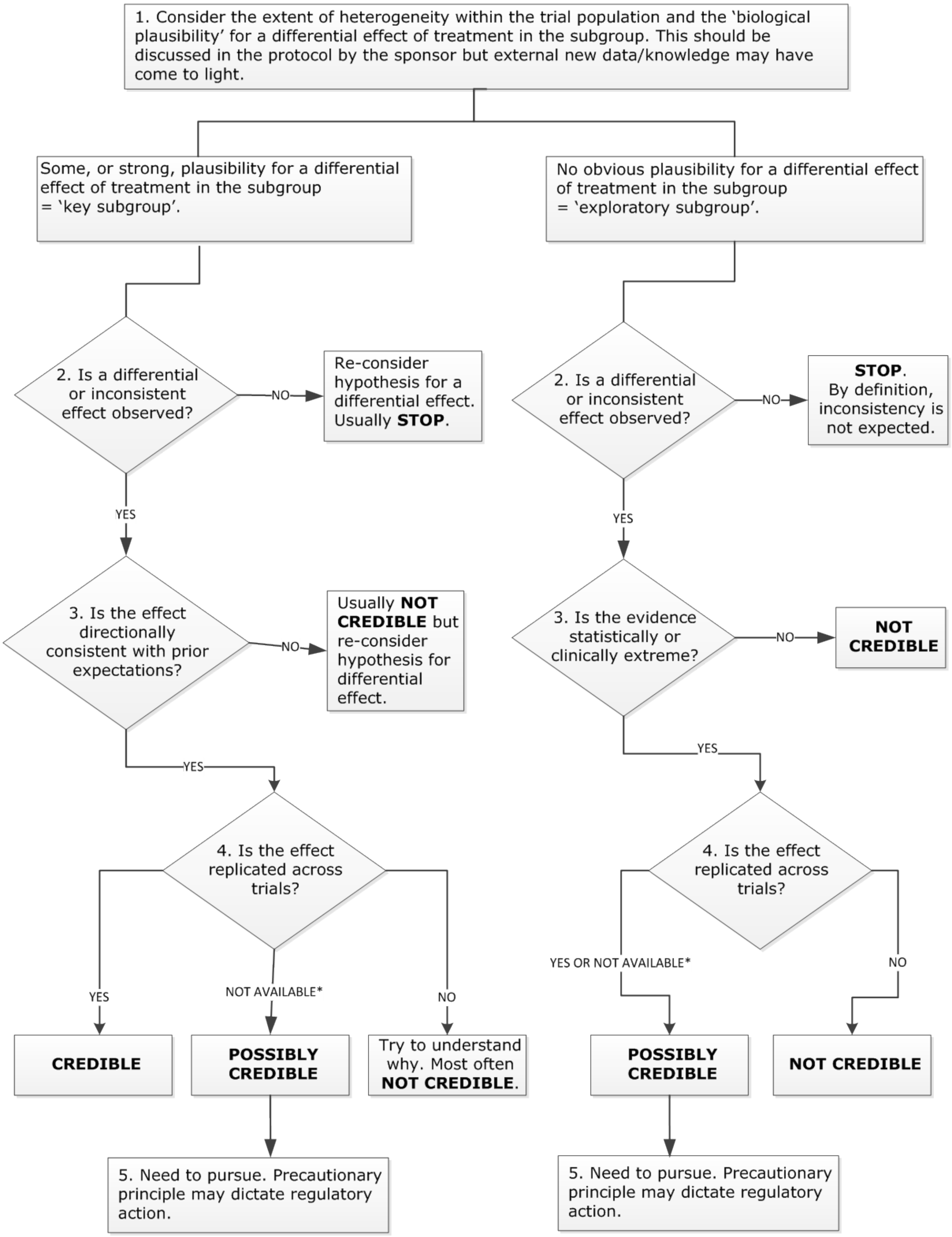
This relates to the use of a subgroup to rescue a trial that has formally failed, such that the primary analysis fails (usually classified as p>5%, two-sided). It is a well-known fact, from a formal statistical point of view, that no further confirmatory conclusions are possible in a clinical trial where the primary null hypothesis cannot be rejected. No formal proof of efficacy is possible under such circumstances and the potential for bias is such that data cannot be considered reliable.

In this case there may be interest to try to rescue the trial in order to gain regulatory approval without conducting expensive and time-consuming additional studies, in particular for the clinical setting of high unmet medical need or situations where trials are usually of considerable size (like in cardiovascular diseases) careful assessment of the overall available evidence has to be performed and substantial limitations need to be identified before replication is requested. However, it must be indicated that this type of exercise would be regarded as inadequate to support a licensing decision in most instances. One or more additional trials should usually be conducted.

If nevertheless a positive licensing decision is, exceptionally, considered in this circumstance then Section 6.4 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.

Annex

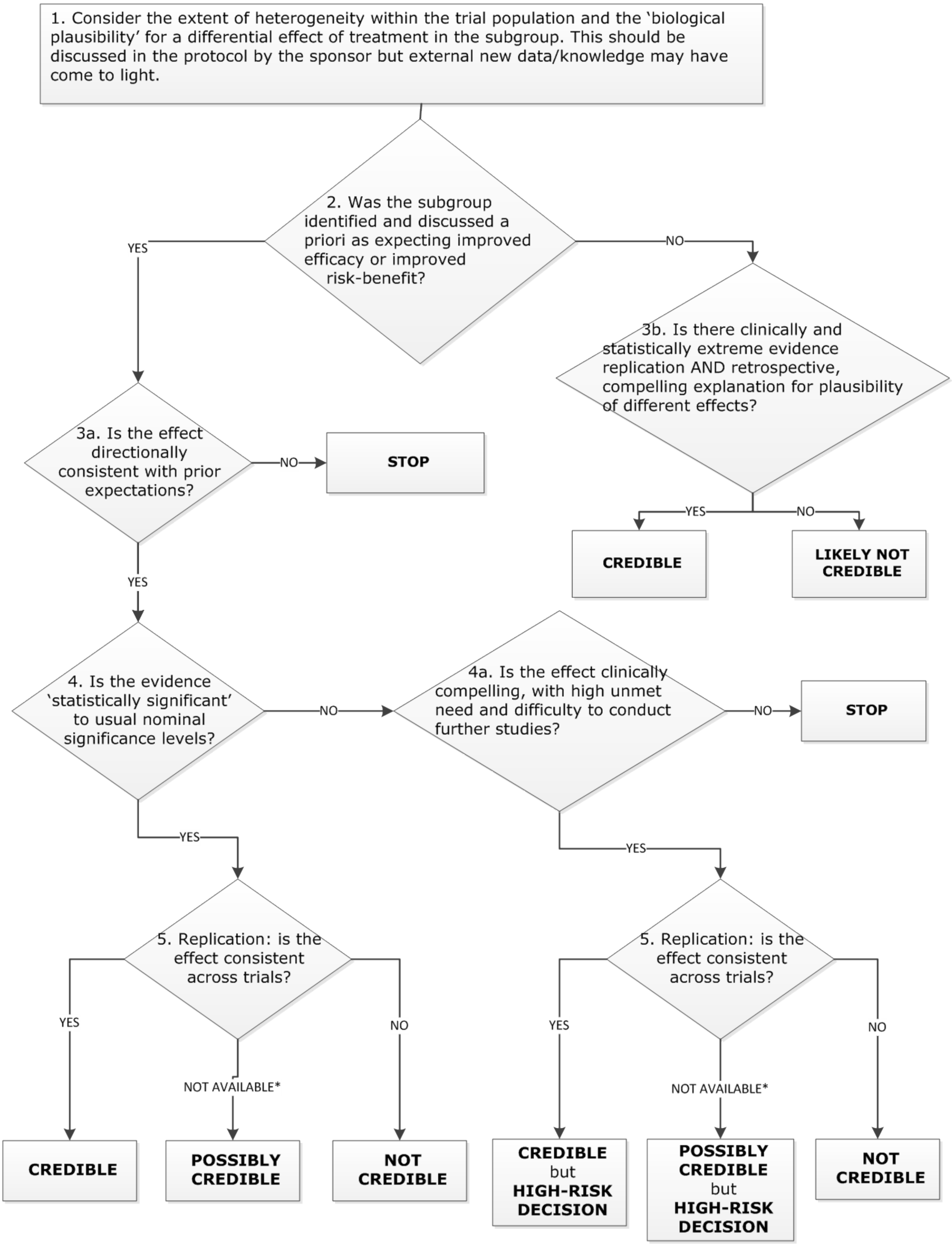
Annex 1 - Scenario 1 (Section 6.3) - establishing 'credibility' when considering 'consistency'



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
CREDIBLE
NOT AVAILABLE*
POSSIBLY CREDIBLE
NO
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*
POSSIBLY CREDIBLE
NO
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 6.4) - 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.