



1 23 January 2014  
2 EMA/CHMP/539146/2013  
3 Committee for Medicinal Products for Human Use (CHMP)

4 **Guideline on the investigation of subgroups in**  
5 **confirmatory clinical trials**  
6 **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

7  
8

Comments should be provided using this [template](#). The completed comments form should be sent to [Biostatistics@ema.europa.eu](mailto:Biostatistics@ema.europa.eu)

9

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

10  
11



12 **Guideline on the investigation of subgroups in**  
13 **confirmatory clinical trials**

14 **Table of contents**

15 **Executive summary ..... 3**

16 **1. Introduction and Problem statement ..... 4**

17 **2. Scope ..... 5**

18 **3. Legal basis and relevant guidelines ..... 6**

19 **4. General considerations ..... 6**

20 4.1. Definition of a subgroup .....6

21 4.2. Problems with conducting multiple subgroup analyses .....7

22 4.3. Basic considerations for investigation of heterogeneity, analysis of subgroups and  
23 associated data presentations.....8

24 4.4. Key considerations that underpin assessment of subgroups .....9

25 **5. Issues to be addressed at the planning stage ..... 10**

26 5.1. Considering heterogeneity within a target population ..... 10

27 5.2. Prioritising the exploratory analyses ..... 11

28 5.3. Country or region used for pre-stratification..... 12

29 5.4. Documenting the exploratory analyses ..... 12

30 **6. Issues to be addressed during assessment ..... 13**

31 6.1. Assessing ‘consistency’ (‘homogeneity’) and ‘inconsistency’ (‘heterogeneity’) ..... 13

32 6.2. Defining ‘credibility’ ..... 15

33 6.3. Scenario 1: The clinical data presented are overall statistically persuasive with  
34 therapeutic efficacy demonstrated globally. It is of interest to verify that the  
35 conclusions of therapeutic efficacy and safety apply consistently across subgroups of  
36 the clinical trial population. .... 15

37 6.4. Scenario 2: The clinical data presented are overall statistically persuasive but with  
38 therapeutic efficacy or benefit/risk which is borderline or unconvincing and it is of  
39 interest to identify a subgroup that has not been pre-specified as part of the  
40 confirmatory testing strategy, where efficacy and risk-benefit would be convincing. 16

41 6.5. Scenario 3: The clinical data presented fail to establish statistically persuasive  
42 evidence but there is interest in identifying a subgroup, where a relevant treatment  
43 effect is evident and there is compelling evidence of a favourable risk-benefit. .... 18

44 **Annex ..... 19**

45

46

47

## 48 Executive summary

- 49 1. Investigation into the effects of treatment in well-defined subsets of the trial population is an  
50 integral part of clinical trial planning, analysis and inference that follows the inspection of the  
51 primary outcome of the trial. The guideline should assist in the planning and presentation of these  
52 investigations and in the understanding of factors to be discussed when considering the credibility  
53 of findings.
- 54 2. The more homogeneous the population studied, in terms of baseline risk and in terms of response  
55 to treatment, the lower the importance of exploratory subgroup analyses for regulatory  
56 assessment. The more heterogeneous the study population, the greater the importance of  
57 subgroup analyses to check that the estimated overall effect is broadly applicable and supports  
58 assessment of risk-benefit across the breadth of the proposed indication. Exploration of  
59 heterogeneity should include covariate-adjusted analyses and subgroup analyses.
- 60 3. Methodological complications related to multiple analyses mean that exploratory investigations into  
61 effects in subsets of the trial population must proceed with caution taking into consideration all  
62 available evidence, not only the point estimates from individual subgroup analyses. Despite the  
63 statistical complications, not investigating, or ignoring results of subgroup analyses could also lead  
64 to incorrect decisions.
- 65 4. Assessors should expect to find discussion in the trial protocol of the expected degree of  
66 heterogeneity of the patient population in terms both of factors likely to be prognostic for the  
67 course of disease and those that are plausibly predictive of differential response to treatment. A  
68 strategy that simply assumes homogeneity of a population in terms of its likely response to  
69 treatment, without discussion and without further investigation, is not sufficient. Analogously, it is  
70 not sufficient to dismiss all subgroup findings that indicate heterogeneity of response as being  
71 spurious. The benefits of this additional discussion are to maximise the *a priori* discussion of the  
72 importance of certain subgroups and thus to minimise the *a posteriori* discussion in an attempt to  
73 promote rational consideration of subgroups and to reduce the risk for erroneous conclusions.  
74 Done properly, this should minimise the need for data-driven investigations, relying instead on a  
75 well-reasoned pre-specified strategy.
- 76 5. Consistency of findings in relevant subgroups needs to be discussed in the analysis report: Forest  
77 plots graphing the treatment effect in a series of subgroups and statistical methods to assess  
78 heterogeneity of treatment effects estimated in subgroups play an important role for the provision  
79 of signals as to whether the overall treatment effect applies to the full trial population. Clinical and  
80 pharmacological knowledge are needed to evaluate the credibility and relevance of signals that are  
81 generated. A number of factors influence the credibility of a subgroup finding, including 'biological  
82 plausibility' and replication of evidence as well as the strength of evidence from the trial(s).  
83 Credible explanations for heterogeneity should be sought. Multiple analyses and data  
84 presentations may be required to properly inform an assessment of credibility.
- 85 6. A strategy for assessing the credibility of subgroup findings is presented for different situations  
86 that are commonly encountered. Key considerations for switching from the all randomised  
87 population to a subgroup for risk-benefit decision making are given. Subgroup analyses will not  
88 usually rescue failed trials.

## 89 **1. Introduction and Problem statement**

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

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

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

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

96 Consequent to (a) and (b), evidence of therapeutic efficacy and evidence to inform the risk-benefit  
97 decision is generated in the clinical development programme and, in particular, in Phase III  
98 confirmatory clinical trials. Confirmatory clinical trials are performed in late-stage drug development to  
99 inform a risk-benefit decision and to justify a treatment recommendation. Assessment of these trials  
100 usually proceeds through investigation of the treatment effect on the primary and secondary outcome  
101 measures in the whole population, and through investigation into the safety profile of the experimental  
102 drug. For confirmatory trials, robust evidence for therapeutic efficacy is required in a relatively broad  
103 patient population that is representative of patient population to be described in Section 4.1 of the  
104 SmPC (external validity). Evidence is considered to be more robust if treatment effects across the  
105 trials in the application, as well as in relevant subgroups within one trial (internal consistency), are  
106 consistent and substantiate the claim to be made for the experimental treatment. This justifies a  
107 regulatory assessment of relevant subgroups with regard to relevant endpoints during assessment of  
108 Marketing Authorisation Application (MAA), as a second step subsequent to inspection of the primary  
109 and secondary trial outcomes on the whole trial population.

110 It is known that different patients will respond differently to the same intervention, and also that the  
111 same individual may respond differently to the same intervention on different occasions. This  
112 variability in response usually remains unexplained but it is plausible, and widely accepted, that some  
113 of the variability in response between patients is caused by demographic, environmental, genomic or  
114 disease characteristics, co-morbidities, or by characteristics related to other therapeutic interventions  
115 (e.g. extent of pre-treatment or concomitant treatment). ICH E5 describes “genetic and physiologic  
116 (intrinsic) and the cultural and environmental (extrinsic) characteristics of a population” and the CHMP  
117 Points to consider (PtC) on multiplicity issues in clinical trials states “Some factors are known to cause  
118 heterogeneity of treatment effects such as gender, age, region, severity of disease, ethnic origin, renal  
119 impairment, or differences in absorption or metabolism.” Grouping together patients with similar  
120 characteristics in one or more of these factors is therefore an intuitive way to explore variability of  
121 response to treatment between different groups of patients within a clinical trial dataset.

122 It is widely understood that subgroup analyses need to be interpreted with caution because of the  
123 multiple data presentations that arise when investigating response to treatment within each level of  
124 the many possible intrinsic and extrinsic characteristics. Compounding the problem, when reviewing a  
125 display of subgroup analyses, the reviewer’s eye may be drawn to those groups with extreme  
126 estimates of effect, whether smaller or larger (or in opposing direction) than the overall effect. An  
127 incautious review of subgroup analyses can result in unreliable inferences and, consequently, to poor  
128 decisions from the clinical trial sponsor or regulator. However, whether the true treatment effect is  
129 homogeneous in subgroups cannot be known and hence trial sponsors and regulatory decision makers  
130 are put in a difficult situation: whether to accept an assumption of homogeneity and disregard extreme  
131 and/or pharmacologically plausible findings in subgroups, or whether to anticipate some heterogeneity  
132 and, with appropriate caution and investigation, attempt to use the results of subgroup analyses as  
133 one piece of evidence to inform decision making.

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

## 137 **2. Scope**

138 This document is intended to provide assessors in European regulatory agencies with guidance on  
139 assessment of subgroup analyses in confirmatory clinical trials. These considerations for assessment  
140 impact on the planning of the clinical trial and hence the document should also be useful to clinical trial  
141 sponsors and to assessors engaged in providing Scientific Advice. This guidance document describes  
142 principles and does not dictate any particular practical solutions in respect of statistical methodology  
143 for estimating or testing the treatment effect in subgroups of the trial population.

144 A differentiation is made between investigation of a subgroup as part of the confirmatory testing  
145 strategy and investigation of subgroups in a more exploratory manner. Whilst a number of the  
146 considerations outlined in this document will apply to the former, this is principally a problem related to  
147 multiple-testing because the trial seeks to test hypotheses relating to both the subgroup and the full  
148 trial population. Recommendations regarding pre-planned approaches for decision making in a  
149 confirmatory testing strategy based on subgroups are not discussed here. The guiding principles and  
150 examples for multiple-testing procedures that control the overall false positive rate are described in the  
151 respective guidance (PtC on multiplicity issues in clinical trials).

152 In principle, three situations can be distinguished in which this more exploratory investigation of  
153 subgroups might be pursued (see Sections 6.3-6.5). The first scenario is the most common, applying  
154 to all dossiers in which confirmatory clinical trials establish statistically persuasive and clinically  
155 relevant efficacy in the target population. The second two scenarios are focussed more on a *post hoc*  
156 restriction to the breadth of the target population:

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

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

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

- 170 • subgroup investigations for efficacy or safety variables;
- 171 • confirmatory clinical trials without regard to choice of control group (placebo or active control) or  
172 primary hypothesis (superiority or non-inferiority / equivalence).

173 There may also be interest in criteria for determining inclusion of information in subgroups to Section  
174 5.1 of the Summary of product characteristics. This is predominately a consideration of whether  
175 information on subgroups would be useful to the prescriber but, depending on the circumstance,

176 criteria outlined in Section 6 may also be useful for a determination of whether the evidence generated  
177 may be considered reliable for presentation.

### 178 **3. Legal basis and relevant guidelines**

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

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

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

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

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

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

### 187 **4. General considerations**

#### 188 **4.1. Definition of a subgroup**

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

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

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

209 Most investigations will consider subgroups identified on the basis of a single factor. Subgroups  
210 defined on multiple factors (e.g. females aged >65) may be of interest on occasion but for simplicity,  
211 the descriptions in this document will make reference to a subgroup defined on a single factor (e.g.  
212 gender categorised as male and female), and this will suffice for most investigations. The risks  
213 described in this document around analysis and interpretation of subgroup analyses are exacerbated  
214 by also considering subgroups based on multiple factors, though the need for this more complex type  
215 of investigation cannot be excluded. Another type of investigation is to categorise patients according  
216 to a 'risk score' based on their profile considering multiple prognostic or predictive characteristics. If

217 the risk score is informative, this may represent a worthwhile investigation into understanding  
218 response to treatment. The risk score itself may serve as a factor by which subgroups of patients may  
219 be defined in addition to a categorical factor against which response to treatment may be modelled.

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

## 224 **4.2. Problems with conducting multiple subgroup analyses**

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

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

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

256 This phenomenon is not only relevant to Phase III trials of course. Exploratory trials may result in an  
257 overall effect that is not impressive, but a signal of relevant efficacy may be apparent in a subgroup,  
258 and the sponsor might be tempted to pursue development of the drug in this subgroup. This type of  
259 selection will on average be associated with artificially extreme and potentially unreliable estimates of  
260 subgroup effects that would be, however, detected during the further drug development programme.

261 **4.3 Basic considerations for investigation of heterogeneity, analysis of**  
262 **subgroups and associated data presentations**

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

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

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

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

300 A key question here is the scale on which to assess the influence of covariates on the estimated  
301 treatment effect. Statistical interactions are scale and model dependent. Interactions in linear  
302 regression models represent departures from additivity (differences in treatment effects on an absolute  
303 scale) while interactions in logistic/Cox regression models represent departures from a multiplicative  
304 model (differences in treatment effects on a relative scale). Commonly it is more realistic to expect  
305 homogeneity of treatment effect on the relative scale (e.g. patients with mild disease at baseline do  
306 not have the capacity to experience beneficial effects as large as might be possible in patients with



307 severe disease at baseline). Contrary to this, absolute effects tend to be more intuitive for  
308 understanding the magnitude of effect and are more commonly used in risk-benefit decision-making.  
309 Even where the effects of a medicine are likely to be similar on the relative scale (e.g. 20% reduction  
310 regardless of baseline) the (larger) effect observed in patients with severe disease may offset the  
311 risks, while the (smaller) effect observed in patients with mild disease may not. It is recommended  
312 that the exploration of interactions and effects in subgroups proceeds first on the scale on which the  
313 endpoint is commonly analysed, with supplementary analyses presented on the complementary scale  
314 for those covariates or subgroups that become important for the risk-benefit decision. The assessor  
315 needs to be aware of the scale being used and to question whether additional analyses would be  
316 informative.

317 Estimates derived from exploratory subgroup analyses should be interpreted with caution. Not only  
318 might the play of chance impact the estimated effect, but it is tempting to focus on subgroups with  
319 extreme effects, which introduces a selection bias. Some methods have been proposed in the  
320 statistical literature to reduce the problem, in particular methods that shrink estimates based on  
321 certain underlying assumptions of heterogeneity. These methods may be presented by sponsors but  
322 the underlying assumptions must be carefully considered and discussed.

323 It might be questioned whether the multiplicity associated with subgroup analyses and interaction tests  
324 should be addressed through changes to nominal significance levels for tests or presentation of  
325 confidence intervals. However, since these investigations serve as an indicator for further exploration,  
326 adjustment would be counter-intuitive and is not recommended. The fact that multiple subgroups are  
327 examined, and the number of subgroups examined, is of course a key matter for consideration during  
328 assessment and regulatory decision-making.

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

#### 338 **4.4 Key considerations that underpin assessment of subgroups**

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

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

- 351 b. **Biological plausibility**; a concept describing the extent to which a particular effect (in this case a  
352 differential effect of treatment in a particular subgroup of patients) might be predicted, or might  
353 have been expected, based on clinical, pharmacological, and mechanistic considerations, and  
354 considerations of other relevant external data sources (often referred to collectively as 'Biological  
355 Plausibility'). Plausibility is primarily a clinical and pharmacological judgement and is usually not a  
356 directly quantifiable or measurable concept. Ideally, those factors where biological plausibility  
357 exists will be pre-specified for use as stratification factors or as being of particular interest for  
358 exploratory investigations in the clinical trial protocol.
- 359 c. **Replication** of evidence; the possibility to examine an effect of a particular covariate, or effect  
360 within a particular subgroup, from multiple sources of relevant clinical trial data.

## 361 **5. Issues to be addressed at the planning stage**

### 362 **5.1. Considering heterogeneity within a target population**

363 During the planning of a clinical trial the discussion of known prognostic (differentiating groups with  
364 different clinical progression) and predictive (differentiating groups with different response to  
365 treatment) factors is one of the most important steps. A decision has to be made on the target patient  
366 group for the clinical trial. In particular, whether the criteria for inclusion or exclusion should restrict  
367 the patient population to, say, one level of a certain factor (e.g. biomarker positive), or whether use of  
368 the drug is intended in the full population under the assumption that patients in all subpopulations  
369 defined by the levels of the factor will benefit from treatment (e.g. without regard to biomarker  
370 status). Similarly, the inclusion and exclusion criteria will define the breadth of the population  
371 recruited with regards to other clinical, demographic and disease characteristics. A broad patient  
372 population will tend to support a broad indication statement but will also increase the importance of  
373 investigating heterogeneity of response to treatment.

374 Assessors should expect to find discussion in the trial protocol of the expected degree of heterogeneity  
375 of the patient population in terms both of factors likely to be prognostic for response and those that  
376 are plausibly predictive of different response to treatment. It must be recognised of course that  
377 knowledge of the treatment will increase as the confirmatory trials are conducted and hence, not all  
378 potential sources of heterogeneity can be predicted in advance of the trial. Consistent with the text  
379 quoted below from the CHMP PtC on multiplicity issues in clinical trials *"Some factors are known to  
380 cause heterogeneity of treatment effects such as gender, age, region, severity of disease, ethnic  
381 origin, renal impairment, or differences in absorption or metabolism. Analyses of these important  
382 subgroups should be a regular part of the evaluation of a clinical study (when relevant), but should  
383 usually be considered exploratory, unless there is a priori suspicion that one or more of these factors  
384 may influence the size of effect"*, factors that define a target population may be put in three  
385 categories:

- 386 1. For a particular factor there is strong reason to expect a heterogeneous response to treatment  
387 across the different levels of the factor. In this case separate trials should usually be planned.
- 388 2. For a particular factor there is at least some biological plausibility or external evidence such that a  
389 heterogeneous response might be hypothesised. In this case it is relevant to discuss and plan for  
390 an assessment of consistency of effects.

391 In addition to factors used to stratify randomisation, it would be expected that key demographic  
392 factors, including genomic factors, related to the mechanism of action / pharmacology would be  
393 included in this category. In addition, careful consideration should be given to other factors that  
394 might plausibly be predictive for different response to treatment such as stage, severity or

395 phenotype of disease, use of concomitant medications and possibly region, country, or centre, see  
396 section 5.3.

397 Unlike the factors that might be categorised under point 1, it is not usually required that a formal  
398 proof of efficacy is available individually in all important subgroups in order to conclude on effects  
399 across the breadth of the trial population. It would, however, be prudent to design the trial  
400 accordingly such that a sufficient number of patients are recruited to the subgroup to ensure an  
401 estimate of effect that can be made with reasonable precision so that the applicant is able to  
402 substantiate therapeutic efficacy and a favourable risk-benefit in important subgroups.

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

405 A strategy that assumes homogeneity of a population in terms of its likely response to treatment,  
406 without discussion and without further investigation, is not sufficient.

407 It will usually be appropriate that the recruited population reflects the epidemiology of the disease in  
408 the target patient group (external validity of the trial). The need to stratify the randomisation should  
409 be considered, firstly to reduce the risk of imbalanced allocation of patients from different factor levels  
410 to the treatment groups, and, secondly, to indicate at the planning stage that whether patients with  
411 different risk profile will have the same benefit from the use of the experimental drug is a question to  
412 be examined. Stratified randomisation, however, only tolerates a very limited number of prognostic  
413 factors to be included into the model (see also PtC on adjustment for baseline covariates), and at the  
414 planning stage a thorough discussion with investigators is of importance to identify the most important  
415 prognostic and predictive factors. This discussion should impact on the assessment strategy and  
416 evaluation of subgroup findings.

## 417 **5.2. Prioritising the exploratory analyses**

418 Investigation of homogeneity of response should always be planned, but the associated multiplicity  
419 needs to be considered. It is recommended that two levels of investigation should routinely be  
420 considered, excluding any subgroups planned as part of the confirmatory testing strategy. The first  
421 level would include investigation of 'key' subgroups, including factors used in stratification of the  
422 randomisation and other factors covered by definition number 2 in Section 5.1. Second, truly  
423 exploratory analyses should be planned for the spectrum of demographic, disease and clinical  
424 characteristics, including those factors covered by definition number 3 in Section 5.1.

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

433 It is therefore clear that the assessor will have a key role in determining what subgroups are of key  
434 interest for more detailed exploration. Again this would, ideally, be discussed at the planning stage of  
435 the trial. By necessity, if the sponsor has not provided well-reasoned arguments and a comprehensive  
436 strategy for analysis, regulatory assessment will become more *post hoc*. In addition, factors for which  
437 there is absence of evidence of scientific knowledge to make a classification will necessarily need to be  
438 considered *post hoc*. Whilst considerations of plausibility are usually more convincing when made in  
439 advance of the trial, so that they are not influenced by knowledge of trial data, it is re-iterated that a

440 fully comprehensive discussion on biological plausibility will not always be possible prior to the Phase  
441 III trial. Hypotheses for heterogeneity of response might emerge as scientific knowledge about the  
442 drug or drug class accumulates.

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

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

### 455 **5.3. Country or region used for pre-stratification**

456 ICH E9 requests centre to be included as a stratifying variable for multi-centre clinical trials. This was  
457 based on the experience that centre may be not only a logistic entity, but a strong prognostic factor  
458 summarizing potential impact of differences in hospital settings and patient populations included. With  
459 multi-regional trials it is recommended to include country or region as a factor into the randomisation  
460 model and the analysis (PtC on adjustment for baseline covariates), because including centre often  
461 becomes impractical as few patients are recruited per centre, across a large number of centres. In  
462 recent years the experience has grown that country (or region) can be similar important prognostic  
463 factors covering important intrinsic and extrinsic factors, including different attitudes to diagnosis, co-  
464 medication and other aspects of the concomitant setting. Although it is recommended to address these  
465 aspects by directly addressing the respective variables, country (or region) as entity for checking the  
466 context-sensitivity (or robustness) of the treatment effect is of importance to regional drug licensing  
467 bodies and as a plausible source for learning about the robustness of the treatment effect.

468 As with other factors, whether or not trials should be planned only to meet their primary objective or  
469 whether consideration should be given to how much of a trend for a positive treatment effect should be  
470 available for the results in countries (or regions) should depend on how much knowledge about  
471 similarities or differences in intrinsic and extrinsic factors is available and in how far evidence exists  
472 that the concomitant setting is different in different regions of the world. Consistent findings in regional  
473 strata strengthen such an application and may justify an increase in sample size to investigate  
474 treatment effects by region to avoid trials being inconclusive overall due to substantial regional  
475 differences that were not foreseen at the planning stage.

### 476 **5.4. Documenting the exploratory analyses**

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

482 Pre-specification of subgroups of interest will take different forms. Subgroups intended for  
483 confirmatory inference will be pre-specified as part of the formal statistical testing strategy. As  
484 described above, the trial documents should also discuss, identify and prioritise some key factors and  
485 subgroups for exploratory analysis from the background of indication specific knowledge (e.g. gender  
486 in cardiovascular disease). In addition, stratification factors may have been identified for the  
487 randomisation and indicate that these are important (prognostic or predictive) covariates for statistical  
488 modelling.

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

## 496 **6. Issues to be addressed during assessment**

### 497 **6.1. Assessing 'consistency' ('homogeneity') and 'inconsistency'** 498 **('heterogeneity')**

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

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

517 There is no widely accepted definition for consistency. Presented below are some working definitions  
518 to use when reviewing a series of subgroup analyses. Inconsistencies in one or more subgroups might  
519 give rise to concern about the applicability of the overall treatment effect, if the subgroup analysis  
520 result is found credible. The assessment of consistency is different from that of credibility, see Section  
521 6.2. It is recommended (see Annex 1) that assessment of credibility is based primarily on biological  
522 plausibility and external evidence, but it is also appreciated that an investigation of results within  
523 subgroups is an important part of data review. It is important to recognise that the mere identification  
524 of inconsistency without full consideration of other important factors outlined in this paper should not  
525 generally be used as a basis for regulatory action, for example with regard to restricting the licence.

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

531 Visual inspection of a Forest plot that describes the results for multiple subgroup analyses can help,  
532 specifically where interrogation of subgroup analyses is to flag subsets of the trial population for  
533 further inspection and consideration (see Section 6.2 and Annex 1). However, here too, a formal rule  
534 for interpretation that is both sensitive to detect heterogeneity of potential interest and specific is not  
535 available. Visual inspection should consider the estimate and precision of the overall effect, the  
536 estimates and confidence intervals for the effect in each subgroup and the overall number of  
537 subgroups (the more groups the more likely to observe one or more groups with extreme findings, by  
538 chance). Further research into statistical methods to trigger inspections into subgroups of a  
539 confirmatory clinical trial is needed.

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

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

571 **6.2. Defining ‘credibility’**

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

576 The assessor must consider all evidence that can be brought to bear on the problem including the key  
577 considerations outlined in Section 4.4 above in addition to the clinical trial data. Strong biological  
578 plausibility, or absence thereof, or replication of evidence may well contribute greater weight to the  
579 overall assessment as the pattern of data observed across the range of subgroup analyses presented.  
580 In particular, having two or more relevant sources of evidence is of great assistance to interpretation.  
581 Where two or more trials can be interrogated on effects in a particular subgroup the weight of evidence  
582 from directly relevant clinical trial data rather than from external evidence of lesser relevance or  
583 arguments of biological plausibility increases. Evidence for differential effects in subgroups that are  
584 replicated across available clinical trials can be compelling irrespective of the fact that it may be larger  
585 or smaller than the (average) effect that is overall observed in this trial. This holds true even in the  
586 absence of a plausible mechanistic explanation. Conversely, an inconsistent finding in one trial is more  
587 readily disregarded if evidence from one or more other trials does not replicate this inconsistency, in  
588 particular where there is no *a priori* reason to expect a differential effect. Because of the possibility of  
589 erroneous subgroup findings, a development programme with two trials in which the subgroup can be  
590 assessed is clearly advantageous. This is consistent with the guideline on applications based on a  
591 single pivotal trial which stresses the importance of the assessment of internal consistency in a single  
592 pivotal trial.

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

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

603 In the end it is a major part of the regulatory assessment to weigh signals that have been generated  
604 during visual assessment and/or by means of statistical methods with the knowledge from other trials  
605 in the development program or in the same class, pharmacology and/or mechanistic considerations.  
606 Algorithms for assessing credibility of findings in subgroups are presented below and in Annex 1. No  
607 algorithm can replicate the nuances and complexities of all possible decisions but these should act as a  
608 guide to assessors in considering the strength of evidence available.

609 **6.3. Scenario 1: The clinical data presented are overall statistically**  
610 **persuasive with therapeutic efficacy demonstrated globally. It is of**  
611 **interest to verify that the conclusions of therapeutic efficacy and safety**  
612 **apply consistently across subgroups of the clinical trial population.**

613 Exploration of heterogeneity should include covariate-adjusted analyses and subgroup analyses. If  
614 well-reasoned in the trial protocol, assessment of subgroups may be based primarily on the pre-

615 specified strategy described in Sections 5.1 and 5.2 above and be followed, for completeness, by a  
616 review of other exploratory analyses.

617 If the assessment of key subgroups has been well planned, nothing has arisen during the course of the  
618 trial to change the scientific assessment of plausibility and no evidence of inconsistent findings is  
619 apparent, then investigation may be regarded as being complete. Inconsistent or extreme data in  
620 other exploratory subgroups, where the absence of a plausible link to the effects of treatment response  
621 can be confirmed by the assessor, could generally be disregarded unless the finding is replicated  
622 across more than one trial, or particularly extreme, in which case plausibility should be re-considered.  
623 If the discussion and pre-specification of key subgroups is incomplete then the assessor will by  
624 necessity need to take a more ad-hoc approach and will be forced to rely more on the observed data  
625 and their own judgement of plausibility without the benefit of the structure given above that limits the  
626 number of subgroups that are prioritised for examination.

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

- 630 a. biological plausibility and the inconsistency is in the direction expected. Credibility is particularly  
631 strong if evidence is replicated across multiple data sources, though in submissions with only one  
632 trial in which the subgroup can be properly assessed, the precautionary principle dictates that  
633 replicated evidence cannot be required to confirm credibility of an untoward effect of the  
634 experimental treatment.
- 635 b. replication of the inconsistent finding across multiple data sources. Analogously, credibility is  
636 particularly strong if there is also biological plausibility.

637 This credibility is further supported if tests of interaction are statistically significant, or borderline  
638 significant, and if there is some evidence of treatment-by-covariate interactions across different  
639 endpoints (notwithstanding correlation between endpoints; the stronger the correlation, the less  
640 credibility is enhanced).

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

645 ***6.4. Scenario 2: The clinical data presented are overall statistically***  
646 ***persuasive but with therapeutic efficacy or benefit/risk which is borderline***  
647 ***or unconvincing and it is of interest to identify a subgroup that has not***  
648 ***been pre-specified as part of the confirmatory testing strategy, where***  
649 ***efficacy and risk-benefit would be convincing.***

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

654 This scenario would usually arise because:

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



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

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

663 Here there exists not only the problems of multiplicity, but also of selection bias since the identification  
664 of a subgroup of interest would commence once the data from the trial are known and the eye of the  
665 assessor and the applicant will be drawn to those findings that are most extreme. Therefore, and  
666 because the aim of this exercise is to draw a positive conclusion for marketing authorisation from a  
667 clinical development programme that has not provided persuasive evidence from a statistical and  
668 clinical point of view, the level of evidence needed to establish credibility is arguably higher. For a  
669 subgroup to be considered credible all of the criteria below would usually apply. This list applies in  
670 principle irrespective of whether it is the company or the regulator that is specifying additional  
671 investigations of interest:

- 672 • External evidence should exist that the subgroup of interest is a well-defined and clinically relevant  
673 entity.
- 674 • A pharmacological rationale, or a mechanistically plausible explanation, should exist, why a certain  
675 drug or treatment could have different efficacy (or benefit/risk) in a sub-population and its  
676 complement (considering also the scale of assessment).
- 677 • The estimated effect of treatment in the subgroup would usually be more pronounced in absolute  
678 terms (i.e. indicating a greater benefit) than in the all-randomised population. The totality of  
679 statistical evidence, based on individual trials and pooled analyses, should meet the same  
680 standards of evidence as would usually be expected for the all-randomised population indicating  
681 that the size of the treatment effect in the subgroup is substantial as compared to the variability of  
682 the problem.
- 683 • Replication of subgroup findings from other relevant trials (internal to the MAA or external trials  
684 that are relevant). A particular challenge exists in applications based on a single pivotal study  
685 since replication is a key component of credibility. In this instance the biological plausibility and  
686 the clinical trial data from the subgroup would have to be exceptionally strong.

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

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

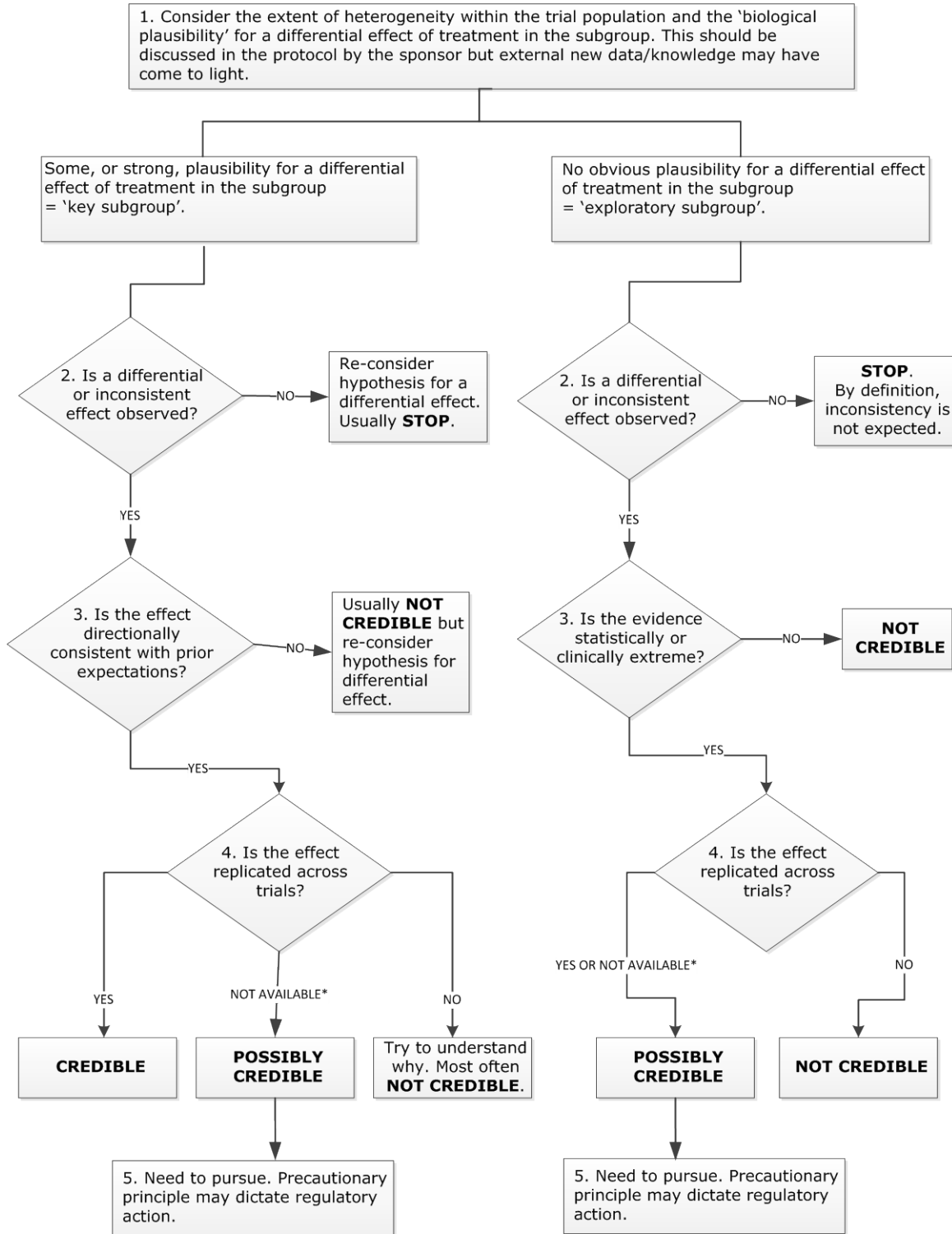
697 **6.5. Scenario 3: The clinical data presented fail to establish statistically**  
698 **persuasive evidence but there is interest in identifying a subgroup, where a**  
699 **relevant treatment effect is evident and there is compelling evidence of a**  
700 **favourable risk-benefit.**

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

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

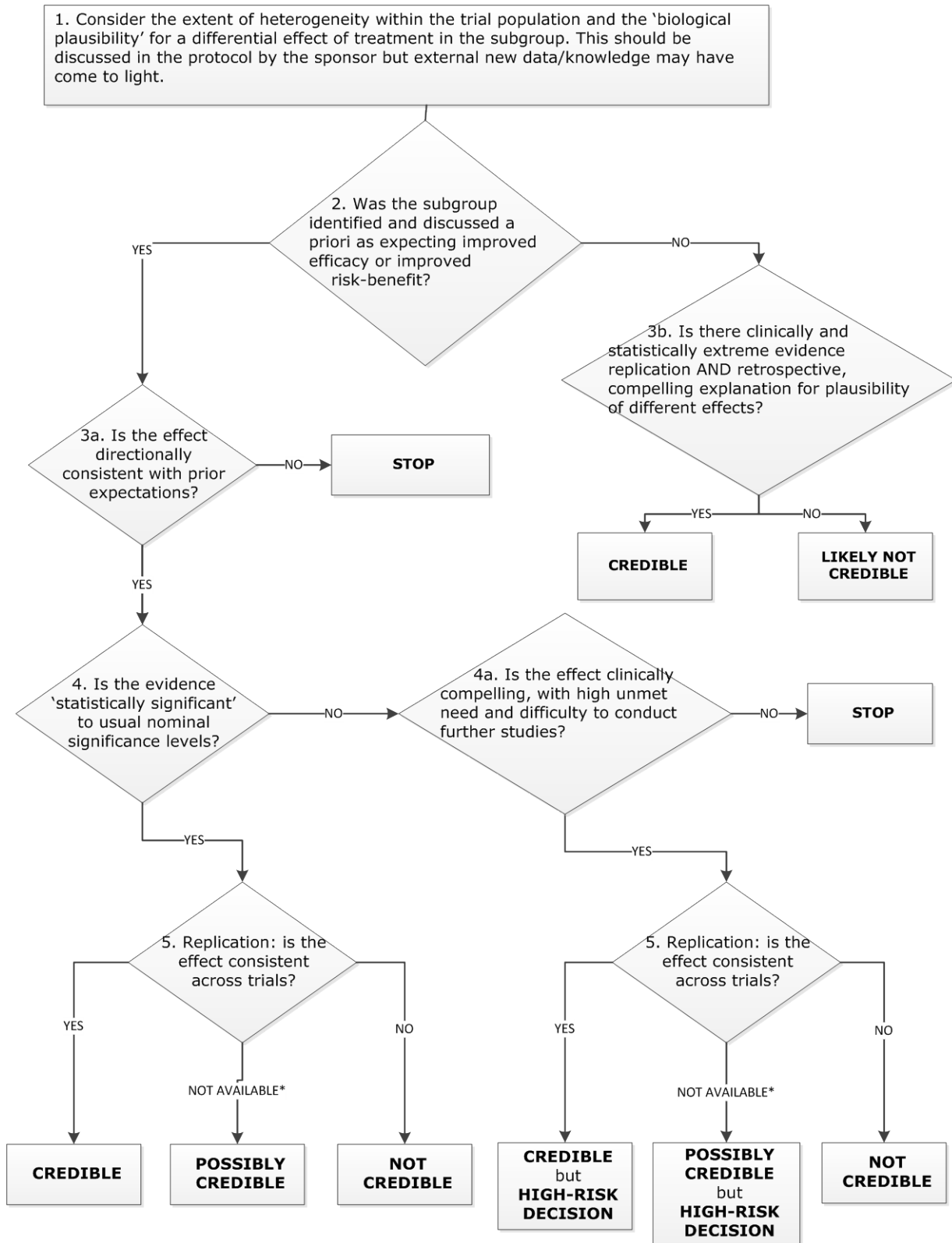
713 If nevertheless a positive licensing decision is, exceptionally, considered in this circumstance then  
714 Section 6.4 represents the minimum criteria that should be fulfilled. In addition, in such a situation, a  
715 clear rationale must exist as to why a properly planned trial has failed despite the drug being regarded  
716 as efficacious and why additional prospective studies to establish formal proof of efficacy are unfeasible  
717 or unwarranted.

718



\*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**



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