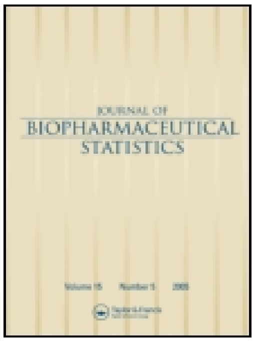

Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: www.tandfonline.com/journals/lbps20

General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials

Alex Dmitrienko, Christoph Muysers, Arno Fritsch & Ilya Lipkovich

To cite this article: Alex Dmitrienko, Christoph Muysers, Arno Fritsch & Ilya Lipkovich (2016) General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials, Journal of Biopharmaceutical Statistics, 26:1, 71-98, DOI: 10.1080/10543406.2015.1092033

To link to this article: https://doi.org/10.1080/10543406.2015.1092033

Published online: 20 Jan 2016.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2016, VOL. 26, NO. 1, 71–98
http://dx.doi.org/10.1080/10543406.2015.1092033

General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials

Alex Dmitrienko[a], Christoph Muysers[b], Arno Fritsch[c], and Ilya Lipkovich[a]

[a]Center for Statistics in Drug Development, Quintiles, Overland Park, Kansas, USA; [b]Clinical Statistics, Bayer HealthCare, Berlin, Germany; [c]Clinical Statistics, Bayer HealthCare, Wuppertal, Germany

ABSTRACT

This article focuses on a broad class of statistical and clinical considerations related to the assessment of treatment effects across patient subgroups in late-stage clinical trials. This article begins with a comprehensive review of clinical trial literature and regulatory guidelines to help define scientifically sound approaches to evaluating subgroup effects in clinical trials. All commonly used types of subgroup analysis are considered in the article, including different variations of prospectively defined and post-hoc subgroup investigations. In the context of confirmatory subgroup analysis, key design and analysis options are presented, which includes conventional and innovative trial designs that support multi-population tailoring approaches. A detailed summary of exploratory subgroup analysis (with the purpose of either consistency assessment or subgroup identification) is also provided. The article promotes a more disciplined approach to post-hoc subgroup identification and formulates key principles that support reliable evaluation of subgroup effects in this setting.

ARTICLE HISTORY
Received 14 August 2015
Accepted 14 August 2015

KEYWORDS
Biomarker evaluation; enrichment designs; regulatory guidance; subgroup analysis; subgroup identification

1. Introduction

It is broadly recognized that patient populations studied in clinical trials cannot be *a priori* considered homogeneous, and the treatment effect is suspected to vary across different subsets of the overall population. For this reason, clinical trial sponsors and regulatory authorities have been actively interested in examining treatment effect heterogeneity and investigating patient subgroups with desirable properties. There are, however, multiple potential pitfalls in subgroup assessments and, as emphasized in the Wiley Encyclopedia of Clinical Trials by Simon (2008), “subgroup analyses have been long criticized by statisticians and clinical trialists. The criticisms have generally focused on investigators attempting to transform negative studies into positive studies” (p. 341).

This article deals with the assessment of patient subgroups in late-stage confirmatory clinical trials. Despite the general focus on confirmatory trials, subgroup assessment strategies considered in this article can be exploratory or confirmatory in nature. Due to this, most subgroup analysis methods can also be applied to early phases of clinical development to gain a better understanding of relevant patient characteristics and subpopulations. It is important to bear in mind that the later the decision to enrich a patient population is made (if this is the ultimate goal), the more convincing the data and arguments need to be. In other words, while data from early-stage trials are typically used in a rather informal manner to define the target population for a clinical development program, adjustments of the patient population at later stages require convincing biological arguments combined with strict adherence to key scientific and statistical principles. For a long time, clinical trial sponsors and regulatory authorities have subscribed to the view that the patient population in late-stage trials should closely mirror the future users on the market (see, for example, the ICH E9

CONTACT Alex Dmitrienko ✉ Alex.Dmitrienko@gmail.com Center for Statistics in Drug Development, Quintiles, 6700 W 115th St., Overland Park, KS 66209, USA.

guidance, ICH, 1999). Remarkable progress in areas related to innovative trial design and analysis has improved sponsors' ability to select the most appropriate population of future patients in late-stage trials.

Regulatory agencies place increasingly more emphasis on different types of routine subgroup assessments in all late-stage trials. These assessments are conducted to examine potential interactions and provide a better description of the general characteristics of the target population. These approaches need to rely on scientifically sound methods and support reliable evaluation of subgroup effects in clinical trials, especially in the context of post-hoc (exploratory) subgroup analysis.

The main goal of this article is to provide practical guidance for a broad set of statistical and operational problems that arise in the context of subgroup analysis in late-phase clinical trials. This includes the review of clinical trial literature with emphasis on approaches to the analysis of prespecified patient subgroups as well as identification and confirmation of subgroups in post-hoc assessments. The article emphasizes the importance of using scientifically sound strategies in subgroup analysis and promotes a more disciplined approach to subgroup investigation. Furthermore, a discussion of applicable regulatory guidance documents will be provided, including the Food and Drug Administration (FDA) enrichment guidance (FDA, 2012), Asian guidelines that deal with subgroup investigation and the draft European Medical Agency (EMA) guidance on subgroup analysis (EMA, 2014). The statistical and clinical considerations presented in these guidance documents might have a direct impact on the definition of the target population for a particular therapy (to be specified in the product label) even if subgroup assessments in the Phase III trials were exploratory rather than confirmatory. It is, however, very important to discuss specific approaches to conducting confirmatory subgroup analysis that can be employed by trial sponsors.

The article is organized as follows. Section 2 introduces the general topic of subgroup analysis in late-stage clinical trials. Regulatory guidelines for subgroup evaluation, including the recently published FDA and EMA guidelines, are discussed in Section 3. Sections 4 and 5 provide an overview of confirmatory and exploratory subgroup analysis settings. Two important classes of exploratory methods (consistency assessments and post-hoc subgroup exploration) are discussed in Sections 6 and 7. Section 8 reviews other important considerations in subgroup analysis such as the use of post-baseline patient characteristics or subgroup investigations in clinical trials with a noninferiority assessment. A summary of the recommendations presented in the article is included in Section 9.

2. Subgroup analysis approaches in clinical trials

Proper consideration of patient subgroups is required at multiple stages of trial development. For the sake of illustration, we will focus on Stage 1 (trial planning), Stage 2 (analysis planning), and Stage 3 (reporting and interpretation). Beginning with the trial planning stage, it is critical to define the patient population in the trial. In a broader sense, the definition of the trial population is embedded into the discussion of the trial's main estimand which is currently under discussion as an addendum to the ICH E9 guidance entitled "*Statistical Principles for Clinical Trials*" (ICH, 1999, 2014). Unlike other important elements such as the trial objective, the population is a relevant aspect of the estimand determination. Due to major implications for the final patient population, it is challenging to identify the right balance in this context. For example, it is recommended in the ICH E9 guidance to "… relax the inclusion and exclusion criteria as much as possible within the target population, while maintaining sufficient homogeneity to permit precise estimation … ." The definition of the trial population should be accompanied by a thorough discussion of the biological and clinical plausibility of relevant covariates/biomarkers (i.e., clinical or demographic variables) as requested, for instance, by the draft EMA subgroup analysis guidance (EMA, 2014, Section 5.1) and their impact on the treatment effect. This discussion will have a direct impact on the trial's analysis plan in Stage 2.

The analysis plan specifies the purpose of subgroup evaluations, e.g., subgroup analysis performed to support an efficacy claim or consistency assessments, and identifies key covariates which need special considerations according to the EMA guidance. Analysis of other covariates that are associated with hypothesis generation for future studies are purely exploratory in nature. These analyses can be planned and conducted at a later stage or in a post-hoc manner.

Stage 3 deals with formulating conclusions related to therapeutic efficacy and safety based on the knowledge gained in the trial. This information should provide the basis for the final decision-making, e.g., concluding homogeneity of treatment effects across relevant covariates.

2.1. *Classification of subgroup evaluation methods*

Collaboration with experts in the corresponding therapeutic area plays a key role in the trial planning and reporting/interpretation stages. The experts' feedback helps guide the selection of covariates and define the general goals of subgroup analysis. Analysis planning in Stage 2 is mostly driven by the trial's statisticians and will be discussed in detail in this article.

The demand for subgroup analysis methodology has led to the development of an impressive portfolio of methods. Some of the subgroup evaluation methods used in clinical trials have a fairly long history and are well known, e.g., tests for interaction. Others are new and more sophisticated, including modern tools for performing multiplicity adjustment and constructing adaptive designs. It is also worth mentioning new proposals that rely on basic rules that are easy to apply; however, their utility is under discussion. Examples include the least beneficial subgroup approach (e.g., Alosh and Huque, 2013) or the EMA's recommendation to flag inconsistent results (EMA, 2014, Section 6.1). Since it is crucial to ensure that the statistical methodology is aligned with the goals of subgroup evaluation, we will describe the pros and cons of available statistical methods based on a classification scheme presented in Table 1.

The high-level overview provided in Table 1 can also be used as a decision tree which facilitates the selection of appropriate subgroup analysis methods. The selection process relies on a series of questions pertaining to the specification or timing of subgroup investigation, i.e., if the subgroup analysis is prospectively planned or post-hoc, what is the purpose of subgroup assessment, what is the analysis type, etc. Another important consideration is whether or not a particular method accounts for multiplicity and applies alpha adjustments.

Confirmatory subgroup analyses labeled as Category A analyses in Table 1 involve one or more prospectively defined subpopulations of patients. This setting is commonly used in the development of targeted therapies (see, for example, Millen et al., 2012; Millen et al., 2014b; Dmitrienko et al., 2015b). Such approaches are described in more detail in Section 4 and are generally aimed at evaluating the efficacy profile of a new treatment in a specified trial population. Section 4 reviews relevant trial design considerations, state-of-the-art statistical methodology, and decision-making processes.

Category B analyses comprise consistency assessments that focus on characterizing the heterogeneity of a population's response to the experimental treatment. It is well known that treatment effect is potentially affected on different levels by a number of patient characteristics/covariates due to known or less understood aspects of the treatment's mechanism of action. Traditional exploratory subgroup analysis often specified in clinical trial protocols falls into this category. This analysis may include around 10–20 candidate covariates and, most commonly, one-at-a-time interaction tests are carried out on each covariate. Due to the increasing interest in sensitivity analysis and assessment of the robustness of subgroup analysis results, several innovative subgroup analysis methods have been introduced over the recent years. Despite the exploratory character of sensitivity analysis approaches, the results might be used for formal decision-making in terms of approval or labeling in the regulatory process. Commonly used methods in Category B will be described and discussed in Sections 6.1 and 6.2. We will differentiate between purely descriptive methods, e.g., forest plots, and inferential approaches, e.g., interaction tests. While exploratory methods in Category B may utilize multiplicity adjustments, it is important to note that a less stringent approach to multiplicity control is normally applied in this setting, which distinguishes it from strict multiplicity adjustments employed in Category A. Further, one needs to bear in mind that p-values and confidence intervals

Table 1. Classification of subgroup analysis methods.

Specification	Prospective			Post-hoc	
Primary purpose	Efficacy claim	Consistency assessment		Hypothesis generation/Analysis of failed studies	
Analysis	Confirmatory	Exploratory		Exploratory	
Alpha-adjustment	Yes	Yes	No	Yes	No
Typical approaches/ methods	Established methods for controlling familywise error rate (FWER), including multiple comparison procedures (MCPs), enrichment and adaptive designs aimed at FWER control	Alpha adaptation while maintaining the overall power (Koch and Schwartz, 2014) Criterion based on the “least benefited” subgroup (Alosh and Huque, 2013)	Interaction tests and their extensions, e.g., the interaction-to-overall-effects ratio (Wang and Hung, 2014) Graphical approaches (e.g., forest plots) Shrinkage estimation (Berger et al., 2014) Flags for inconsistency (EMA, 2014)	Subgroup search methods, e.g., SIDES (Dmitrienko et al., 2015a) or Virtual Twins (Foster et al., 2011) Biomarker discovery (Lipkovich and Dmitrienko, 2014)	Data mining to examine data patterns Graphical approaches to visualize effects of continuous covariates (Royston and Sauerbrei, 2013)
Categorization and reference in this article	Category A analyses in Section 4	Category B analyses in Section 6		Category C analyses in Section 7	

cannot be used for confirmatory conclusions if there is no strict multiplicity control in combination with a proper pre-planning. As stressed in the Wiley Encyclopedia of Clinical Trials (2008), “the use of statistical significance tests, confidence intervals, multiplicity corrections, or Bayesian modeling can confuse the fact that no valid inference can really be claimed” (p. 342). Consequently, even if, for instance, the use of confidence intervals and *p*-values in a forest plot facilitates the inspection of consistency, it is premature to draw formal inferences regarding positive/negative treatment effects or inconsistency across subpopulations.

Post-hoc subgroup analysis labeled as Category C analysis in Table 1 refers to a broad class of approaches aimed at evaluating treatment effects across multiple subgroups of patients defined using baseline patient characteristics in a data-mining manner. Two general types of exploratory post-hoc subgroup analysis are presented in Table 1 and will be further described in Section 7. Considering post-hoc subgroup analyses in Category C, it is important to make a distinction between disciplined approaches such as structured subgroup identification and purely exploratory analyses that may deal, for example, with visualizing subgroup patterns. Methods in the first class focus on biomarker and subgroup discovery from a large pool of candidate covariates, employing complex data-mining/machine-learning algorithms, often with the idea of informing future trial designs. The second class of post-hoc subgroup analysis helps address questions about the experimental treatment after the end of the development program, e.g., responses to regulatory inquiries, assessment of safety issues, post-marketing activities in Phase IV trials, assessment of treatment heterogeneity in multiregional trials, and hypothesis generation for future development programs. Due to currently available computing resources, extensive subgroup searches within the data-mining framework can even enumerate and explore all potential subgroups in a given dataset. The question remains here how the enormous amount of information can be reasonably interpreted without the risk of detecting too many false-positive signals. Despite the awareness of multiplicity issues and limited value of post-hoc analysis, each signal tends to trigger concerns or discussions of their reliability and consequences. Application of sound statistical methodology, e.g., permutation-based multiplicity adjustments, shrinkage estimation, or alpha adaptation methods defined in Table 1, helps reduce the probability of incorrectly discovering noninformative subgroups in exploratory settings. In any case, all results from post-hoc subgroup analysis must be considered very cautiously and cannot be used to support any inferential conclusions.

It is worth mentioning several review papers and tutorials on the general topic of subgroup investigation in late-stage clinical trials. Scientific and regulatory considerations arising in confirmatory and exploratory subgroup analysis settings were discussed in Dmitrienko et al. (2015b). Ondra et al. (2015) presented a systematic review of trial design and analysis strategies employed in clinical trials with target subgroups. Lipkovich et al. (2015) provided a comprehensive review of statistical methods used in post-hoc subgroup analysis.

3. Regulatory guidelines

An overview of applicable regulatory guidance documents that consider subgroup analysis issues will be provided in this section. The guidance documents have fostered the development of a significant number of statistical methods for subgroup analysis. From a high-level perspective, it is reasonable to subdivide this section according to three geographical regions covered by the FDA and EMA as well as the Chinese and Japanese regions represented by the Chinese Food and Drug Administration (CFDA) and Pharmaceuticals and Medical Devices Agency (PMDA), respectively. These will be presented in the following subsections.

Beginning with ICH guidelines, only a few aspects of subgroup analysis are mentioned in the ICH E9 guidance. A general statement made in this document, i.e., “the subjects in confirmatory trials should more closely mirror the target population,” is certainly reasonable. Nevertheless, when more sophisticated trial designs with adaptive approaches and enrichment strategies are considered, it raises the question of how closely the initial trial population should match the final prescribed label

Table 2. Outcomes and risks in subgroup analysis.

	Efficacy analyses in subpopulation		Safety analyses in subpopulation	
	Treatment is effective	Treatment is not effective	Treatment is harmful	Treatment is not harmful
Label restriction based on subgroup analysis	Deprive subpopulation of treatment	Correct decision	Correct decision	Deprive subpopulation of treatment
No label restriction	Correct decision	Unnecessary exposure	Harm	Correct decision

population because of adaptations that take place after study planning. In addition, as stated above, the ICH E9 guideline requests to "… relax the inclusion and exclusion criteria as much as possible within the target population, while maintaining sufficient homogeneity to permit precise estimation …." As in the discussions surrounding the precise definition of consistency across multiple patient subgroups (see, for example, in Section 3.2), it is natural to ask what level of homogeneity translates into "sufficient homogeneity." In practice, decisions regarding homogeneity should be made on a case-by-case basis with an interdisciplinary discussion of relevant factors and considerations.

An important issue which triggers a remarkable amount of discussion pertains to determining the right balance between the following two risks in the approval and labeling process. On the one hand, there is a risk of overlooking a subgroup of patients who do not benefit from the treatment, which leads to unnecessary drug exposure with an accompanying risk of adverse drug reactions. This might happen due to undetected treatment effect heterogeneity and an overwhelming effect in the most-benefitting subgroup. The other risk is based on an erroneous conclusion of treatment effect heterogeneity, which results in excluding a patient subpopulation from the product label due to the apparent lack of efficacy. This conclusion may be reached in clinical trials with a large set of prospective and post-hoc analyses without an appropriate multiplicity adjustment. In this case, the subpopulation will be deprived of an available beneficial treatment. The arguments presented above focus on subgroup analyses in the context of efficacy assessment. If safety signals are taken into account, the risks are reversed. The resulting decision-making framework for a given subpopulation is presented in Table 2. Despite the straightforward implications in the table, it is worthwhile carefully examining the different outcomes and risks.

Table 2 shows that the patient subpopulation is deprived of treatment due to a decision to restrict the product label in two scenarios. This erroneous decision is due to lack of appropriate statistical adjustments in subgroup assessments or due to chance findings. Further, the label may be restricted if relevant side effects are detected in the patient subpopulation. A thorough benefit–risk assessment is required in this case even if the treatment has been correctly identified as harmful. If the treatment is not harmful, patients in the subpopulation are deprived of a beneficial treatment and the benefit–risk assessment becomes irrelevant. Naturally, it is unknown whether or not the benefit–risk assessment is useful due to uncertainties related to the usually descriptive nature of subgroup analysis. Despite the fact that uncertainties arise due to small subgroups and the descriptive nature of the analyses in the corresponding benefit–risk assessment, one should keep in mind that ultimately exactly one decision is to be made either including or excluding the corresponding subgroup of patients.

With regard to benefit–risk assessments, a specific situation is described in the draft EMA subgroup analysis guidance (EMA, 2014, Section 6.4). In this setting, the overall treatment effect is statistically persuasive; however, therapeutic efficacy is borderline. It is possible to identify a subgroup of patients in a post-hoc manner where the efficacy and risk benefit would be convincing. However, this scenario is deemed to be applicable only in rare occasions.

3.1. *FDA regulations*

There is currently no dedicated guideline on subgroup analysis available from the FDA. However, the FDA has established a working group within the Office of Biostatistics which is preparing a white paper to provide guidelines for subgroup analysis in clinical trials. The current plan for the white paper has a clear focus on Category A analyses defined in Table 1 as well as consistency assessments. Remarkably, consistency assessments are considered mostly in the context of post-hoc analysis (Category C analysis methods) rather than prospective planning (Category B analyses). This appears to be a major difference compared to the EMA approach which is described in Section 3.2.

The plan for the FDA white paper comprises the possible impact on the population for treatment use on completed trials, i.e. based on post-hoc analyses. This will be discussed in light of chance findings and testing for subgroup-by-treatment interaction along with relevant statistical considerations. When designing clinical trials prospectively to establish an efficacy claim, attention needs to be given to powering such trials appropriately (see also Section 4). As an outlook within the white paper, the following additional topics were mentioned: Bayesian perspectives, noninferiority trials, personalized medicine, and impact from error in subgroup classifier.

The FDA enrichment guidance (FDA, 2012) should be mentioned in the context of Category A analyses. This guidance focuses on the composition of the trial population with respect to prognostic and predictive factors to support personalized medicine approaches in clinical trials.

The published FDA enrichment guidance and planned FDA white paper on subgroup analysis deal mostly with the prospective confirmatory setting (Category A analyses) or post-hoc setting (Category C analyses), respectively. In addition, several recent publications by the authors of the white paper contribute to the Category B analysis framework. This includes Alosh and Huque (2013) and Wang and Hung (2014).

Within the framework of prospective confirmatory subgroup analysis, Alosh and Huque (2013) defined a criterion for concluding that the treatment effect in the least-benefitted (complementary) subgroup exceeds a certain minimum threshold. The criterion is based on testing the effect in the complementary subgroup, but at an alpha level that is potentially higher than the usual two-sided 0.05. The authors recommended to choose the alpha level based on safety considerations, where a safe drug would allow for a higher alpha. The resulting alpha might go up to 0.5, so that it would only be required that the estimated treatment effect in the complementary subgroup is in the right direction. This approach has the advantage of giving a more precise criterion for assessing consistency. On the other hand, the exact choice of the consistency alpha might still be challenging. In addition, the approach would be difficult to apply outside the specific situation of the complementary subgroup in a confirmatory subgroup analysis.

Wang and Hung (2014) introduced an approach based on the interaction-to-overall effects ratio, which might lead in certain cases to a recommendation for a label restriction. See Section 6.2 for more details.

3.2. *EMA regulations*

The EMA has issued the draft guidance entitled "*Guideline on the investigation of subgroups in confirmatory clinical trials*" (EMA, 2014) which formulates general principles that are broadly applicable to different classes of subgroup assessments. Consequently, this triggers the need for more practical guidance and discussion on subgroup analysis implementation. In particular, it is stated in the guidance that it "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" (EMA, 2014, Section 2). This guideline refers to several previously issued guidance documents. The most relevant guidance documents will be discussed later in this subsection.

It is worth noting that the EMA subgroup analysis guidance overlaps in timing and also some textual parts with the EMA guidance entitled "*Guideline on adjustment for baseline covariates in clinical trials*" (EMA, 2015), previously known as "Points to consider on adjustment for baseline covariates" (EMA, 2003). The overlap deals with the need for stratified randomization of important covariates. With regard to the inclusion of interaction terms in the analysis model, it is stated that the "… primary model should not include treatment by covariate interactions," assuming that no different subgroup effects are present. However, where a priori substantial interactions can be expected, "the trial should be designed to allow separate estimates of the treatment effects in specific subgroups." The analysis of interaction terms is recommended in the framework of sensitivity analysis in both guidance documents, but it is stressed that the sole reporting and interpretation of interaction p-values is not adequate. The EMA subgroup analysis guidance emphasizes qualitative assessments, for example, it is stated in the context of exploratory and sensitivity analysis that, if the interaction term "… is particularly large in size or qualitative in nature, then interpretation of the overall results of the trial may become impossible." (EMA, 2015, Section 7.3).

When dealing with Category A approaches to subgroup analysis, it is important to mention two other EMA guidance documents, namely, "*Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*" (EMA, 2007) and "*Points to consider on multiplicity issues in clinical trials*" (EMA, 2002). Note that EMA (2002) is currently under revision and a revised guidance document is expected to be released as a draft for public commenting in 2015. The former guidance pleads for a cautious implementation of adaptive designs despite the advantages for subgroup analysis as described in Section 4. Interestingly, the term "subgroup" is mentioned only once in this guidance document and only in the context of exploratory consistency assessment. Further, EMA (2002) describes the impact of unexplained heterogeneity in important subpopulations on potential license restriction. On the other hand, if a trial is conducted to evaluate the treatment effect in a particular subgroup, the analysis strategy must be prespecified based, for example, on Category A approaches. It is also mentioned in EMA (2002) that important patient subgroups should be included in exploratory evaluation of treatment effects in confirmatory trials.

As mentioned in Section 3.1, the major difference between the FDA and EMA approaches is that the latter focuses on consistency assessments within a sensitivity analysis framework, which should be prospectively planned according to EMA (2014). Consequently, Category B approaches are of particular importance in this setting. Furthermore, the guidance cannot be viewed as a summary of statistical subgroup analysis methodology but is rather "intended to provide assessors in European regulatory agencies with guidance on assessment of subgroup analyses in confirmatory clinical trials" (EMA, 2014, Section 2). Based on this scope, the EMA subgroup analysis guidance provides straightforward tools for consistency assessments, e.g., forest plots with subsequent visual inspection or flags for inconsistency described in Section 6.2.

EMA (2014) introduced three scenarios for subgroup assessment in late-stage clinical trials. Starting with the less frequent third scenario, a strong and clear medical need has to be shown when a benefitting subgroup is intended to be established in failed trial based on a post-hoc evaluation. The second scenario is more relevant and deals with settings where the overall population results are statistically positive but with therapeutic efficacy or benefit/risk which is borderline or unconvincing. It is of interest to identify a patient subgroup that has not been prespecified in a confirmatory sense, where efficacy and risk–benefit would be convincing. For this scenario and the first scenario (described below), a decision tree is provided in the guidance to support the assessor's adjudication. The most relevant scenario, labeled as Scenario 1, is embedded into the key aspects of the guidance. It describes how credibility can be established in the context of consistency assessment. Despite the exploratory character of the approaches considered, many nodes in the decision tree include concrete advice, i.e., "Need to pursue. Precautionary principle may dictate regulatory action" (EMA, 2014, Section Annex 1, Scenario 1). Since the decision tree may be applied to several key covariates in a trial, it is important to recognize that the probability of an incorrect conclusion increases with the number of covariates. To help address this issue, the EMA guideline recommends that trial sponsors should predefine key covariates for which biological plausibility of an interaction is discussed in advance, whereas the remaining candidate covariates would be treated as "truly exploratory" (EMA, 2014, Section 5.2).

To summarize, the EMA guidance on subgroup analysis places much emphasis on prospective considerations related to biological plausibility and consistency assessments. Evaluation of biological plausibility requires a thorough interdisciplinary discussion of all potential relevant covariates ideally based on historical data. This discussion is requested to be part of the trial protocol as well as of the trial report (the role of biological plausibility in subgroup evaluation is discussed further in Section 7.3). Consistency assessments based on Category B approaches are crucial for the European decision-making process. Despite the importance of consistency assessments, a stringent pathway through a statistical process similar to the structured Category A framework defined in Table 1 is not available.

It is worth mentioning two papers that were published in the special issue of Journal of Biopharmaceutical Statistics on subgroup analysis in clinical trials guest-edited by Alex Dmitrienko and Sue-Jane Wang (January 2014). The two papers, Hemmings (2014) and Koch and Framke (2014), focus on regulatory considerations in late-stage trials with post-hoc subgroup assessments. The ideas presented in these papers are consistent with the general framework presented in EMA (2014). Hemmings (2014) provided an overview of statistical and regulatory issues with respect to the three stages mentioned in Section 2, i.e., the trial planning, analysis planning, and reporting/interpretation stages. As a result, this article covered a broad spectrum of topics and provided interesting discussion and useful recommendations. The background information presented in this article sheds additional light on the key principles of the EMA guidance on subgroup analysis and helps explain why consistency assessments play a predominant role in that guidance document. Koch and Framke (2014) highlighted further aspects of subgroup assessments such as analysis of treatment effect across regions in multiregional clinical trials.

Finally, there is an extension to the ICH E9 guidance under preparation. Even though this initiative is primarily triggered by the recent developments in handling and prevention of missing data, it supports some ideas of the EMA approach of gaining more information from sensitivity analyses. The Final Concept Paper E9 (R1) Addendum states that “following ICH E9, it has become standard in all regions to prespecify a primary statistical analysis for efficacy, but it has also been common practice to investigate the extent to which the outcomes of other approaches to the analysis lead to consistent findings.”

3.3. *CFDA and PMDA regulations*

Asian regulatory guidelines predominately stem from the CFDA and the Japanese Pharmaceuticals and Medical Devices Agency (PMDA). When considering subgroup analyses, CFDA and PMDA guidance documents focus on the evaluation of ethnic differences only. Dedicated guidelines on subgroup analysis apart from the assessment of ethnic issues are not yet available. It would be especially important to provide guidance on the situations where the ethnic differences are evaluated in a framework of subgroup analysis.

Analysis of ethnic differences in clinical drug development is discussed in ICH guidance documents. The ICH E5 guideline (ICH, 1998) entitled “*Ethnic Factors in the Acceptability of Foreign Clinical Data*” regulates the bridging process which addresses intrinsic and extrinsic ethnic differences. However, the bridging concept is based on separate studies rather than subgroup analysis in a single trial of corresponding Asian and non-Asian populations. In contrast to this approach and in light of large multiregional studies that are commonly included in Phase III programs, it is becoming increasingly important to address subgroup analysis issues even if restricted to ethnic consistency assessments. Furthermore, all analysis plans that deal with potential regional heterogeneity should be discussed beforehand with regulators to achieve a consensus.

The PMDA released the reference cases for supplemental explanation of global trials in East Asia in “Basic Principles on Global Clinical Trials” (PMDA, 2012) to encourage Japanese subject participation in global trials. A reasonable size of the subgroup of Japanese subjects is expected in these trial to support subgroup evaluations to demonstrate consistency with the overall trial population. However, a global trial with Japanese subjects is typically conducted only if homogeneity/consistency can be assumed. When a global trial is planned, two criteria are defined. The first criterion requires a sufficient number of Japanese subjects to show that at least half of the overall effect is retained in the Japanese subgroup. The second criterion requires a sufficient number of Japanese subjects to show the same positive trend as in the overall population. Important statistical considerations such as sample size requirements as well alternative criteria are discussed in Carroll and Le Maulf (2011). Apart from the standard ethnic factors, there is a trend to consider other background factors or influential factors that need to be evaluated and discussed as part of subgroup investigation.

The CFDA requests local data in different categories of their registration types in separate studies. Replication of the overall treatment effect in ethnic subgroups is described in "Provisions for Drug Registration" (CFDA, 2007). It is interesting that, more recently, the CFDA advice often includes references to "consistency assessments" as in the EMA guidance on subgroup analysis. In 2015 the CFDA released a guideline on International Multicenter Clinical Trials (IMCTs) (CFDA, 2015). The intention was to define the requirements for IMCTs involving Chinese study sites and subsequent analysis of subgroups of Chinese subjects in comparison to the overall trial population to investigate the consistency of treatment effects. However, the exact definition of "consistency" or "similarity" is still lacking, which complicates the development of quantitative approaches to evaluating consistency in multiregional clinical trials with Chinese subjects.

4. Confirmatory subgroup analysis setting

As indicated in Section 2, confirmatory subgroup analysis is performed in Phase III clinical trials with prospectively defined patient populations (overall population and target subpopulations). Each subpopulation is defined using a binary classifier based on one or more biomarkers. Confirmatory subgroup analysis is conducted by the sponsor and has important labeling implications. The outcome of subgroup evaluations has a direct impact on the product label, e.g., the label can be potentially enhanced or restricted.

The main statistical and clinical components of the confirmatory subgroup analysis setting include

- Trial design (e.g., subpopulation-only design versus multi-population design, fixed design versus adaptive design, oversampling strategy, etc.).
- Statistical analysis methods, including multiplicity and oversampling adjustments.
- Decision-making framework for selecting the most relevant patient populations in the product label.

A detailed review of these components is provided Sections 4.1–4.3.

Despite the use of sophisticated statistical methods, confirmatory subgroup analysis approaches in a standard fixed-design setting are fully accepted from a regulatory perspective. An agreement on the individual elements of the trial design and analysis methodology is to be achieved among the stakeholders before the study start. The outcome of the confirmatory subgroup evaluation is highly credible if the prespecified decision rules are followed. As indicated below, additional evidence may be required to achieve an acceptable level of credibility if an adaptive design with data-driven rules for population selection or enrichment is proposed by the sponsor.

4.1. *Trial designs*

This subsection discusses candidate options for designing clinical trials with prospectively specified subpopulations. Multiple factors impact the choice of the most appropriate trial design, including the available information on predictive strength of classifiers used to define the subpopulations, relative size of the subpopulations, etc. For example, if there is reliable evidence that the new treatment will benefit mostly classifier-positive patients, the trial's sponsor may consider an adaptive design with an option to discontinue the enrollment of classifier-negative patients at an interim analysis.

4.1.1. *Fixed designs*

Beginning with a fixed-design setting, assume that all elements of the trial's design are predefined and consider two common approaches to selecting the trial's population and objective (Millen et al., 2012). With the first approach, known as a *subpopulation-only* or *biomarker-enriched* design, patient enrollment is restricted to the subpopulation of classifier-positive patients, and thus the treatment effect is not studied in the complementary subpopulation. By contrast, both classifier-positive and classifier-negative patients are enrolled in the trial, when the second approach is considered (*multi-population* or *biomarker-stratified* designs). Multi-population designs are more efficient than single-population designs and support investigation of the treatment effect in several subgroups within a single trial.

An important feature of multi-population designs is that the proportion of classifier-positive patients can be increased to better characterize the efficacy and safety profile of the treatment in this subset. This can be accomplished via *oversampling* (Zhao et al., 2010). A simple approach to oversampling is based on the following two-stage randomization scheme. In the first stage, all patients are enrolled until a prespecified proportion of classifier-negative patients is reached. After that, the randomization is modified to enroll classifier-positive patients only. Alternatively, the randomization algorithm can be customized to enroll all classifier-positive patients and randomly reject patients with a classifier-negative status.

When comparing subpopulation-only and multi-population designs in the development of targeted therapies, clinical trial sponsors need to be aware of potential limitations of excluding classifier-negative patients. As an illustration, consider the tailoring strategy employed in the trastuzumab development program in breast cancer (trastuzumab is widely marketed as Herceptin). The patient population in this program was limited to a subset of classifier-positive patients (patients with HER2-positive breast cancer) and trastuzumab was shown to be effective in this subpopulation. Subsequent clinical trials demonstrated that trastuzumab may be also effective in other types of breast cancer. This means that patients who were excluded from the population studied in the original development program could have benefited from the new treatment. The decision to utilize a subpopulation-only design prevented the sponsor from performing a comprehensive risk–benefit assessment in this program and potentially deprived a relevant subgroup of patients from a beneficial therapy.

In addition, it was emphasized in Freidlin et al. (2012) that subpopulation-only designs implicitly rely on the assumption that the selected classifier reliably identifies patients who are most likely to benefit from a given therapy. However, classifiers used in Phase III trials are typically developed based on rather small sets of data, and it may be challenging to accurately estimate their operating characteristics (sensitivity and specificity). Finally, when discussing the pros and cons of subpopulation-only designs, the FDA enrichment guidance highlighted the importance of investigating effects of new treatments in broader populations, including subpopulations of classifier-negative patients (FDA, 2012, Section VII.B). This implies that multi-population designs should generally be preferred to subpopulation-only designs.

4.1.2. *Adaptive designs*

The fixed-design setting considered above can be extended by considering adaptive approaches to designing confirmatory trials with predefined patient subgroups that have been discussed in multiple publications. Adaptive approaches provide a flexible alternative to conventional multi-population trials because they rely on multistage designs with an option to modify the trial population based on the data available at one or more interim looks. Most common types of adaptive designs include *adaptive population selection* or *adaptive population enrichment* designs that support decisions to update the patient enrollment and/or analysis strategy at the final analysis. Returning to the example given at the beginning of this subsection, interim data may suggest that the experimental treatment is not effective in classifier-negative patients, in which case it is no longer worthwhile to recruit these patients. Different types of adaptive population selection designs were examined in Brannath et al. (2009), Jenkins et al. (2011), Friede et al. (2012), and Stallard et al. (2014).

Brannath et al. (2009) highlighted the advantages of an adaptive approach to population selection over traditional approaches. Within the traditional development framework, the following three-step process may be employed:

- Step 1. Multiple hypotheses regarding the most appropriate choice of the patient population are generated in a Phase II trial and a promising subpopulation is selected.
- Step 2. The treatment effect in this subpopulation is evaluated in another Phase II trial.
- Step 3. The subpopulation is confirmed in a Phase III trial.

Adaptive population selection designs enable the sponsor to combine Steps 2 and 3 into a single trial and the initial population selection (Step 1) can be performed in a parallel exploratory (biomarker) trial. Along the same line, the FDA enrichment guidance mentioned adaptive designs as a way to facilitate the assessment of treatment effects in target subgroups (FDA, 2012, Section VI.D).

Another example of an adaptive approach to setting up multi-population designs deals with threshold selection for a continuous biomarker. As stated earlier in this section, a binary classifier needs to be specified to define a subpopulation of patients. Clinical trial sponsors may encounter situations where a continuous biomarker with strong predictive properties is selected from multiple biomarkers examined in a Phase II development program. However, the available data may be limited to reliably estimate the threshold/cutoff point to set up a binary classifier for the Phase III program. In this case, the program's sponsor may consider a two-stage adaptive design with an interim analysis. The data collected in the first stage of the trial can be used to compute the threshold, which is then utilized in the second-stage analysis.

With virtually all adaptive designs and especially with more experimental approaches such as designs with an adaptive threshold selection, it is important to ensure that the overall methodology is scientifically sound and credible. It was emphasized in multiple publications and regulatory guidelines (see, for example, EMA, 2007) that operating characteristics of trial designs with adaptive elements ought to be well understood. It is highly recommended to perform a comprehensive evaluation of key characteristics of candidate adaptive designs such as the Type I error rate via simulations.

4.2. *Statistical analysis methods*

As stated in Section 4.1, a clinical trial with a multi-population design (e.g., overall population and a single-target subpopulation) enables its sponsor to examine the effect of the experimental treatment in two groups of patients. Since there are several opportunities to claim overall success, an important concern in the analysis of multi-population trials is protection of the Type I error rate. There are numerous multiplicity adjustment methods that can be applied to control the error rate at the nominal level, including basic non-parametric procedures (Bonferroni procedure) and more powerful parametric procedures (parametric chain and feedback procedures). In addition, if multiple objectives are pursued in a clinical trial with several patient populations, e.g., the efficacy of the experimental treatment is evaluated using several clinical endpoints, more advanced methods such as gatekeeping procedures need to be considered. For a review of traditional and advanced multiplicity adjustment methods, see Dmitrienko et al. (2013) and Dmitrienko and D'Agostino (2013).

When selecting an appropriate multiplicity adjustment for a multi-population trial, especially in the case of multiple objectives, it is important to utilize all available clinical information. Specifically, the trial's sponsor needs to account for the clinically relevant logical restrictions among the null hypotheses of interest. For example, in a clinical trial with a single prespecified subpopulation (target subpopulation), the null hypotheses of no treatment effect in the overall population and target subpopulation are interchangeable since each one of them is associated with an independent regulatory claim. It will be inappropriate to apply a sequentially rejective method in this particular problem and test the null hypothesis in the target subpopulation only if a significant treatment effect is established in the overall population. Second, relevant statistical information, e.g., information on the joint distribution of the test statistics, needs to be taken into account. Since prespecified subpopulations are subsets of the overall population and are likely to overlap, the test statistics are positively correlated. Using the known joint distribution with positive correlation coefficients, the trial's sponsor can set up parametric multiple testing procedures that provide a power advantage over basic procedure such as the Bonferroni.

It is worth briefly mentioning additional considerations that are relevant in the context of adaptive population selection designs. Since any adaptive design offers multiple opportunities to claim success, it is critical to employ an analytical method to protect the overall Type I error rate (recent publications on this broad topic include, for example, Friede et al., 2012; Stallard et al., 2014). It is also recommended to perform additional adjustments such as the computation of bias-adjusted estimates of the treatment effect to determine the true magnitude of the treatment difference in the populations of interest.

If an oversampling strategy is employed in a multi-population trial design to increase the relative size of the target subgroup, the conventional test statistic for the treatment effect in the overall population is no longer meaningful. This test statistic needs to be modified to account for the fact that classifier-positive patients are overrepresented in the trial population, which biases the inferences. Adjustments for oversampling in a fixed-design setting were developed in Zhao et al. (2010). This methodology can be extended to multistage adaptive designs with several patient populations.

4.3. *Decision-making framework*

In addition to statistical adjustments used in multi-population trials such as a multiplicity adjustment, it is critical to ensure that the conclusions drawn at the final analysis are clinically meaningful and support the formulation of specific regulatory claims. A general framework was developed in Millen et al. (2012, 2014b) to streamline the decision-making process in a multi-population setting. This framework is based on the *influence* and *interaction* conditions (see Table 3).

To introduce the two conditions, consider a clinical trial with a predefined subpopulation (target subpopulation). Using an appropriate multiplicity adjustment, the trial's sponsor can perform inferences in the overall patient population and target subpopulation, which results in three distinct outcomes and associated regulatory claims:

- The treatment is shown to be effective in the overall population only (broad effect claim).
- The treatment is shown to be effective in the target subpopulation only (tailored effect claim).
- The treatment is shown to be effective in the overall population and target subpopulation (enhanced effect claim).

When considering the *broad effect claim* in the overall population, it is important to bear in mind that the statistically significant overall treatment effect may be caused by a highly significant effect in the target subpopulation, while no benefit is observed in the complementary subpopulation. If this is indeed the case, it will be inappropriate to consider the broad effect claim. An alternative claim (*tailored effect claim*) can be justified by applying the influence condition as the first step of the decision-making process in Table 3. This condition requires a certain amount of evidence of a beneficial effect in the complementary subgroup to support the conclusion that the beneficial

Table 3. Decision-making framework is based on the influence and interaction conditions.

Decision-making process	Condition is met	Condition is not met
Step 1. Evaluate the influence condition	Continue to Step 2	Tailored effect claim
Step 2. Evaluate the interaction condition	Enhanced effect claim	Broad effect claim

treatment effect is applicable to the overall patient population. The tailored effect claim is considered if the influence condition is not satisfied.

If the influence condition is met in Step 1, the next step involves the interaction condition which focuses on the magnitude of differential treatment effect between the target and complementary subpopulations. If the interaction condition is satisfied in Step 2 (i.e., there is a meaningful difference between the two subpopulations), the *enhanced effect claim* is recommended in the clinical trial. On the other hand, if the interaction condition is not met (i.e., a treatment-by-subgroup interaction is not established), the trial's sponsor can consider the *broad effect claim*.

Millen et al. (2012, 2014b) proposed simple frequentist rules based on treatment effect estimates in the target and complementary subpopulations for evaluation of the influence and interaction conditions (see also Dmitrienko et al., 2015b). These rules were extended in Millen et al. (2014a) using Bayesian methodology. The general decision-making framework defined above was extended in several directions. For example, Alosh et al. (2015) focused on the evaluation of the influence error rate (it was referred to as the error rate of the complementary subgroup).

The decision-making framework defined above is equally applicable to multi-population trials with adaptive designs. The influence and interaction conditions can be applied to multistage adaptive designs to facilitate the selection of the most relevant patient populations for a given treatment as well as the hypothesis tests to be carried out at the final analysis. As a quick example, consider again a clinical trial with a target subpopulation and suppose that a two-stage design with a single interim analysis will be utilized in the trial. If there is evidence of a meaningful treatment effect in the complementary subpopulation (influence condition is met) as well as differential effect between the target and complementary subpopulations (interaction condition is met) at the interim analysis, it will be sensible to make the following choices in the second stage of the trial:

- Patient population in the second stage: Enroll all patients.
- Hypothesis tests at the final analysis: Null hypothesis of no effect in the overall population and null hypothesis of no effect in the target subpopulation.

On the other hand, if the influence condition is satisfied but there is no appreciable difference between the target and complementary subpopulations (interaction condition is not satisfied), it will be most appropriate to consider the following patient population and analysis strategy:

- Patient population in the second stage: Enroll all patients.
- Hypothesis tests at the final analysis: Null hypothesis of no effect in the overall population (hypothesis test in the target subpopulation is no longer relevant since the interaction condition is not met).

5. Exploratory subgroup analysis setting

A key feature of the confirmatory subgroup analysis setting presented in Section 4 is that it assumes that all relevant patient subgroups are prospectively defined. However, it is impossible to prespecify all potentially important subgroups and, as a result, plausible subgroups must be examined in an exploratory manner. This section as well as Sections 6 and 7 focuses on exploratory aspects of subgroup analysis in Phase III clinical trials.

To begin with general goals of exploratory subgroup analysis, it is important to point out that subgroup exploration is commonly performed by sponsors of Phase III development programs as well as regulatory reviewers and can potentially lead to labeling changes. Examples of the impact of exploratory subgroup investigation are provided below:

- Scenario 1: Restricted product label due to lack of efficacy. A subgroup of patients may be excluded from the product label due to lack of efficacy and, as a consequence, the product label will be restricted to the complementary subgroup. Scenario 1 was discussed, for example, in Alosh et al. (2015) in the context of identifying a set of subgroups that benefit the most from a novel treatment in a clinical trial with a significant overall effect.
- Scenario 2: Restricted product label due to unacceptable safety. A subgroup of patients may be excluded from the product label due to unacceptable safety issues. This setting was discussed, for example, in Wang et al. (2006).
- Scenario 3: Enhanced product label due to superior efficacy/improved risk–benefit. A subgroup with enhanced treatment effect may be highlighted in the product label in addition to presenting a beneficial effect in the overall population of patients.
- Scenario 4: Subgroup exploration to support a positive product label. A subgroup with a beneficial treatment effect may be identified in a failed clinical trial or program (overall treatment effect is not significant).

Scenarios 1, 2, and 4 were discussed in the EMA subgroup analysis guidance (EMA, 2014, Section 6). However, Scenario 3 did not receive sufficient attention in the guidance (this possibility was only briefly mentioned in Section 6.4 in the Appendix of this EMA guidance). Traditionally, exploratory subgroup analysis in a clinical trial with a significant overall effect is viewed mainly as a tool for restricting the product label to a subgroup rather than potentially enriching the product label by emphasizing an enhanced treatment effect in a subgroup. It is vital to share information on treatment options with the prescribing physicians, including the efficacy in the general population and, if applicable, key subpopulations. This approach to presenting the results of exploratory subgroup investigations will ultimately help improve patient care.

Further, when considering Scenario 4, it is important to remember that “subgroup analyses will not usually rescue failed trials” (EMA, 2014, Executive summary). Positive findings in a subgroup must be replicated in one or more subsequent confirmatory trials. The program’s sponsor needs to provide strong reasons why a subgroup is to be analyzed to rescue a failed clinical trial. General reasons include an unmet medical need or a poorly planned clinical trial, e.g., inclusion criteria were too broad to demonstrate a meaningful effect. Subgroup analysis findings from a clinical development program with a negative outcome carry more weight if the selected patient subgroup is defined based on an important prespecified covariate such as a stratification factor. In addition, it is helpful if a priori arguments of the biological plausibility of a beneficial effect in the subgroup are presented.

For the purposes of this article, we will focus on two key types of exploratory subgroup analysis mentioned in Table 1:

- Consistency assessment.
- Post-hoc subgroup identification.

Consistency assessments play a key role in determining whether or not the overall conclusions regarding treatment effectiveness apply across all important subgroups of patients in a given trial. Conventional approaches to consistency checks as well as new methods aimed at facilitating consistency assessments in Phase III trials are discussed in Section 6.

While consistency assessments are normally aimed at confirming treatment effect homogeneity across relevant patient subgroups, the main goal of post-hoc subgroup identification is to provide a reliable characterization of treatment effect heterogeneity. This characterization helps discover subsets of patients with desirable features such as an improved benefit or reduced side effects. For example, exploratory methods can help identify a subgroup of patients who may experience a beneficial treatment effect despite a negative result in the overall population in Scenario 4. This can be accomplished through a systematic subgroup search which focuses on subgroups of patients with a strong treatment effect compared to patients in the corresponding complementary subgroup.

Section 7 provides a review of subgroup identification approaches in positive and failed Phase III clinical trials.

6. Exploratory subgroup analysis: Consistency assessment

Consistency assessments provide insights into the homogeneity of the treatment effect across multiple subsets of the trial population based on key patient characteristics (for example, demographic and clinical factors such as patient's gender or disease characteristics). If the treatment effects are shown to be consistent across the levels of all key characteristics, this finding strengthens the evidence of treatment effectiveness in the proposed patient population.

This section discusses traditional and recently proposed approaches to the assessment of treatment effect homogeneity which is conducted as part of post-hoc subgroup investigation in all Phase III trials.

6.1. *Conventional approaches to consistency assessment*

Two types of consistency/heterogeneity assessments are typically conducted in a clinical trial or integrated clinical trial database. The first one is *restricted heterogeneity assessment* which is based on a small number of predefined factors/patient characteristics such as stratification factors. The second one is *unrestricted subgroup exploration* which utilizes all relevant covariates.

The investigation of consistency of a beneficial treatment effect in the overall population across important subgroups of patients is commonly performed using graphical tools such as forest plots. Forest plots present a series of unadjusted confidence intervals for the treatment difference within selected subgroups of patients.

Statistical assessment of the heterogeneity of treatment effects across subgroups relies on appropriate treatment-by-subgroup interaction tests. It is well known that interaction tests are not very sensitive and power to detect a significant treatment-by-subgroup interaction tends to be low. A useful rule-of-thumb was mentioned in Wang and Hung (2014) which sheds light on the expected sensitivity of interaction tests. The interaction effect (i.e., difference in the treatment effect between subgroups) needs to be at least twice as large as the treatment difference for which the trial was powered to achieve the same level of power to detect a significant interaction. Equivalently, the trial needs to be at least four times as large to detect an interaction effect with the same probability as the treatment effect in the overall population of the original trial. This simple rule-of-thumb assumes a balanced trial design and power of interaction tests is even lower in unbalanced settings. A non-significant interaction clearly does not generally indicate that the treatment effect is in fact homogeneous across the subgroups of interest.

Most interaction tests are defined based on a simple contrast between the treatment effects in a subgroup and its complement, e.g., the difference between the sample estimates of the treatment effects in the two subgroups normalized by the pooled standard error. An important property of standard interaction tests is that the relative sizes of the subgroups affect the pooled standard error but not the between-subgroup difference. As a result, standard interaction tests may behave erratically if one of the subgroups is small. Tests of a differential treatment effect present a viable alternative to commonly used interaction tests. These tests are based on the difference between the treatment effect test statistics in the two subgroups and perform better in smaller subgroups. Lipkovich and Dmitrienko (2014) provided a comparison of conventional interaction tests and differential effect tests in the context of subgroup evaluation and concluded that differential effect tests tend to be less sensitive to the subgroup sizes.

6.2. *New proposals for consistency assessment*

A number of alternative approaches to assessing the homogeneity of treatments effects have been recently proposed. We will first review methods that rely on simple decision rules and univariate analyses and then discuss analytical approaches that focus on a synergistic effect of multiple patient characteristics used in subgroup investigations. A well-known weakness of basic univariate analyses of this type is that they do not protect the probability of incorrect conclusions. The sponsor or regulatory agency may mistakenly conclude that the new treatment is effective in the trial population while, in truth, the treatment is beneficial only in a certain subgroup of patients. On the other hand, the treatment benefit may only be established in a subgroup while it actually applies across the overall trial population. The likelihood of drawing incorrect conclusions pertaining to apparent homogeneity or heterogeneity of the treatment across multiple patient subgroups tends to be quite high when multiple patient characteristics are examined. For example, when discussing the use of forest plots, the EMA subgroup analysis guidance points to their limitations: "when interpreting forest plots it is tempting to find reassurance in directional consistence of estimated effects" (EMA, 2014, Section 4.3).

The EMA subgroup analysis guidance introduced the following easy-to-apply rule which can be effectively combined with a forest plot to identify inconsistent findings:

"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 2× or 3× 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" (EMA, 2014, Section 6.1)

While such a rule can easily be applied to any forest plot without even having access to the underlying data, its operating characteristics have not been studied, e.g., it is not clear how often this rule would indicate inconsistency of the results and if there are indeed no differences in the treatment effect between subgroups.

Koch and Schwartz (2014) introduced criteria for assessing whether the results observed in prespecified supportive subgroups are in "harmony" with a statistical significant result in the overall population. The criteria are based on keeping the power for detecting a subgroup effect at the same level as for the overall population. This can be accomplished by using a higher Type I error rate or, alternatively, a less stringent noninferiority margin δ (note that this approach supports an option to use a noninferiority margin in the subgroups if a superiority test is carried out in the overall population). The rationale for keeping the power at a high level is that, if a subgroup fails such a criterion, it can be interpreted as potentially inconsistent with the findings in the overall population. The downside of this approach to consistency assessment is that, especially for smaller subgroups with the relative size below 30%, it results in a remarkably high Type I error or very wide margins. For example, if the relative size of a subgroup is 25%, the Type I error rates needs to be increased to 0.34 to achieve 90% power.

Wang and Hung (2014) introduced the interaction-to-overall effects ratio (denoted by ξ). It was defined as the ratio of the difference in the treatment effect between two subgroups of interest and the treatment effect in the overall population. It was shown that, if $\xi > 2$, a trial is generally adequately powered to detect an interaction effect. Also, ξ is greater than 2 if there is no beneficial effect in one subgroup and at most 50% of patients are in the effective subgroup. Based on these arguments, Wang and Hung (2014) recommended using a sample estimate of the ratio ξ for regulatory decision-making purposes. Suppose, for example, that there is a statistically significant treatment effect in the overall population and a significant interaction is present. If the estimated ratio exceeds 2, it is likely that there is no effect in one of the subgroups, which might then lead to a label restriction. The authors stated that a priori criteria are likely to be required to conclude a "subgroup-only" effect. Therefore, they recommended the joint consideration of a decision tree which is described in the article, subgroup size, and observed interaction-to-overall effects ratio, as a statistical criterion to guide regulatory decision-making.

Finally, a number of more sophisticated methods can be considered to arrive at a more reliable assessment of treatment effect heterogeneity across multiple patient subgroups defined using baseline patient characteristics. As indicated above, key operating characteristics of consistency checks based on simple univariate methods are typically unknown, and it is quite difficult to judge how reliable they are in a particular setting. Statistical approaches that focus on a synergistic effect of multiple patient characteristics can be applied to improve the credibility of conventional consistency checks. This includes methods that utilize analytical or resampling-based adjustments to properly account for multiplicity issues inherent in post-hoc subgroup analysis. This class of adjustments was studied in several recent publications, including Bonetti and Gelber (2004) and Hothorn and Zeileis (2008).

7. Exploratory subgroup analysis: Post-hoc subgroup investigation

It is sometimes stated that exploratory subgroup analysis is mainly a regulatory decision-making problem rather than a statistical problem. This implies that statistical principles are unlikely to play a key role in a post-hoc investigation of subgroups within a confirmatory clinical trial. However, if our goal is to maximize patient benefit and minimize patient risks, it will be important to rely on analytical methods rather than ad-hoc rules. As explained below, even though specific subgroups to be examined cannot be prespecified as in the general confirmatory setting considered in Section 3, specific algorithms for subgroup identification and confirmation can be prospectively defined. A principled approach of this type will enable trial sponsors and regulators to fully examine the operating characteristics of the decision rules used in exploratory subgroup analysis and select methods that perform best in each particular setting.

In what follows, we will discuss analytical strategies in exploratory subgroup investigation, including both subgroup identification and subgroup confirmation, which have been successfully used in Phase III clinical trials.

There is a general concern in statistical and regulatory communities that data-driven subgroup analysis is prone to producing spurious results that are "usually false positive and not replicable" (Huque and Röhmel, 2009). The EMA subgroup analysis guidance on subgroup analyses states that well-implemented drug development programs "minimise the need for data-driven investigations, relying instead on a well-reasoned pre-specified strategy" (EMA, 2014). This section emphasizes that a well-reasoned prespecified subgroup assessment strategy should also contain "data-driven" components and propose to contrast haphazard or ad-hoc "data-driven investigations" with principled data-driven subgroup analysis strategies. The latter recognizes the subgroup analysis as a special case of model selection and thrives on statistical methodologies for model selection developed within the machine/statistical learning and related fields, including multiple comparisons and causal inference.

7.1. *Disciplined subgroup search*

The ultimate goal of disciplined subgroup search (Ruberg and Shen, 2015) is to determine the degree of true treatment heterogeneity across clinically relevant subgroups of patients. Clinically relevant subgroups are defined using candidate biomarkers with strong predictive properties, i.e., the ability to predict the treatment effect assuming it is heterogeneous in the overall population.

The common thread in modern data-driven subgroup identification and analysis methods is the prespecification of the subgroup selection and evaluation strategy rather than the final set of patient subgroups. It is important to recognize that this strategy is data-driven in two aspects. First of all, the strategy is similar to model selection methods in that it focuses on selecting a few final models from a large set of candidate models within the model space. This approach to subgroup identification supports the goal of identifying best subgroups as members of a broadly defined collection of candidate subgroups. Second, subgroup selection methods often include certain meta-parameters that need to be estimated from the data. This includes, for example, parameters that control model space complexity. However, the entire strategy is prespecified in the sense that the model space and methods for estimating the meta-parameters are prospectively defined. Subgroup selection and evaluation strategies may and often involve multiple steps where selection of subsequent steps may depend on what has been learned from previous steps.

Key principles that separate disciplined or principled subgroup search strategies from various ad-hoc strategies are defined below.

It is critical to predefine the entire strategy as a single procedure that can be programmed and applied to a dataset in an algorithmic (i.e., completely automated) fashion without the need for "human intervention." This principle requires prespecification of (i) the "search space," (ii) methods for determining meta-parameters of the procedure, and (iii) rules for defining whether the search returns no subgroups. The "search space" or scope is defined by the datasets used, biomarkers included, and allowable subgroup structure, e.g., whether subgroups can be defined based on a single biomarker, at most two biomarkers, etc. The importance of other two components will be discussed in greater detail later in this section. Here we emphasize that subgroup identification procedures are inherently multistage, and it is important to account for the uncertainty associated with the entire subgroup search process. Subgroup identification and analysis efforts are often presented as a collection of heuristic procedures where the dependence of certain elements of the strategy on the results of previous steps and expert judgment may be obscured. Preliminary data-driven steps that narrowed down the scope of the search may be omitted or not considered part of the strategy. This practice prevents formal evaluation of operating characteristics of the subgroup search procedure in question. In particular, lack of clear definition of "futility rules" when no subgroup would have been claimed makes it impossible to evaluate the Type I error (or false-positive) rates. Without this information, it is challenging to support informed decisions, e.g., whether a new trial which focuses on identified patient subpopulation(s) should be conducted.

Indeed, there is increasing demand for multiplicity control with respect to the entire subgroup identification strategy. While evaluating statistical significance was seen as a rather unusual or unnecessary feature in data-mining/machine-learning applications in the past, many modern data-mining procedures include statistical significance as its core element, perhaps combined with other concepts such as complexity and reproducibility. See, for example, Meinshausen and Bühlmann (2010) and Gunter et al. (2011). While strict multiplicity control based on strong control of the familywise error rate may be impossible to achieve owing to the complexity of the model space and difficulty of enumerating all null hypotheses, weak control of the probability of incorrect subgroup selection associated with a subgroup identification strategy can be implemented by using resampling methods (Lipkovich et al., 2011).

Another and perhaps even more important principle is the principle of complexity control. Complexity control relies on the fundamental idea of achieving a tradeoff between bias and variance to prevent data overfitting. Since model spaces are very large in subgroup identification problems, "uncontrolled" or greedy search is likely to result in data overfitting. In other words, multiple patient subgroups that look very promising may be selected; however, the probability of confirming these subgroups with future data may be quite low. It is worth noting that applying multiplicity adjustments following subgroup selection does not directly address this problem since it does not help find biomarkers with desirable properties "after the fact." As it is commonly done in data mining/machine learning, appropriate complexity control should be built into the model selection or subgroup selection strategy. On the other hand, when complexity control is built into the model selection process (e.g., via penalized likelihood), it alleviates multiplicity burden and reduces multiplicity adjustments compared to a greedy selection method with no complexity control. Multiplicity and complexity control in subgroup search are closely related to each other and should be used in combination.

As part of complexity control and more specific to subgroup identification is the need for controlling (reducing) selection bias when evaluating different candidate subgroups. The model space in subgroup search problems is often defined by considering all possible partitions based on nominal or numerical biomarkers. The number of partitions depends on the number of biomarker's levels, which directly affects the numbers of candidate subgroups. As a result, biomarkers with a larger number of values have an advantage over biomarkers with a few values. It is important to ensure that the probability of falsely selecting a noninformative biomarker does not depend on the number of levels. The problem of selection bias was studied in the context of recursive partitioning algorithms by Loh and Shih (1997) and Hothorn et al. (2006).

The principle of complexity control is also related to the reproducibility principle. Note that lack of constraints on the search space typically leads to selecting patient subgroups based on noninformative biomarkers that have little chance of being replicated in future trials. In general, reproducibility assessment focuses on the probability of reproducing the subgroups identified on training data with future data. Reproducibility assessments are commonly conducted using methods based on the bootstrap and cross-validation followed by replicating the entire subgroup search strategy. Stable patient subgroups that appear often in multiple resamples are more likely to be reproduced in the future. An example of reproducibility assessments based on the learn-and-confirm method is given in Section 7.3.

Finally, a very challenging task in subgroup identification is derivation of reliable estimates of treatment effects in the identified subgroups of patients (the estimates are known as bias-corrected or "honest" estimates). Reliable estimates obtained in early-stage trials play a key role in designing future confirmatory studies. It is well known that biased estimates of the treatment benefit may lead to wrong decisions, resulting in wasted resources and/or lost opportunities. Additional independent datasets are commonly required to compute honest estimates. If no independent datasets are available, resampling-based methods are often applied. In this case, it is important to note that the entire search strategy, including estimation of meta-parameters, ought to be implemented anew on each dataset. A detailed discussion of methods for estimating treatment effects within patient subgroups is provided in Section 7.4.

7.2. *Subgroup identification methods*

To facilitate the discussion and understanding of different subgroup identification methods used in late-stage clinical trials, Lipkovich and Dmitrienko (2014) proposed a general taxonomy of different methods, including global outcome modeling, global treatment effect modeling, and local modeling approaches to subgroup identification. A shortened version of this classification scheme is presented in this section.

Global outcome modeling methods focus on modeling the outcome function, which allows the trial's sponsor to predict the treatment outcome given a patient's biomarker profile and treatment assignment. It is important to note that outcome models support predicting potential or counterfactual outcomes. For example, using the estimated outcome model, the sponsor can predict what outcome the patient with a specific covariate profile would have achieved if assigned to the active treatment even if the patient had been in reality assigned to the control arm. Examples include global outcome modeling methods based on a parametric approach that utilize multivariable fractional polynomials (Royston and Altman, 1994) for modeling treatment-by-covariate interactions (Royston and Sauerbrei, 2004) and a penalized regression method known as FindIT (Imai and Ratkovic, 2013). The Virtual Twins method (Foster et al., 2011) is an example of a non-parametric global outcome modeling approach based on Random forests.

Global treatment effect modeling methods directly model the unobservable treatment contrast at the individual patient's level. This results in predicting the difference in potential outcomes if treated or untreated for a patient with a given biomarker profile. This general approach obviates the need to estimate the "main effect" (i.e., effects of prognostic factors) and focuses on estimating predictive effects (i.e., treatment-by-covariate interactions). The modified covariate method (Tian et al., 2012) is a parametric method, and non-parametric methods include the STEPP method (Bonetti and Gelber, 2004) and several tree-based methods such as the Interaction trees method (Su et al., 2008) and several methods proposed by Loh et al. (2015). As a special case of methods aimed at directly modeling the treatment effect, we can consider the case of predicting only the sign (rather than the magnitude) of the treatment difference. This leads to identifying predictive biomarkers that contribute to qualitative rather than quantitative treatment-by-covariate interactions. Promising methods in this class include outcome-weighted methods (Zhao et al., 2012) and a method developed by Zhang et al. (2012). Both approaches reduce the problem of identifying optimal treatment regimes to a weighted classification problem within the predictive learning framework. As a related note, it is often argued that the reversal of treatment effect (qualitative interactions) is uncommon, and regulators and sponsors should focus on cases of quantitative interactions, for example, on scenarios when there is a zero effect in one subgroup. These settings can be also covered by the methods aimed at identifying optimal treatment regimes, for example, by introducing an appropriate utility function that takes into account safety signals or treatment cost. The subgroup with a zero or small positive treatment effect may often receive a negative utility, after factoring in safety and evaluating the efficacy–safety tradeoff.

Local modeling methods support a direct search for subgroups with a beneficial treatment effect. Methods in this class bypass the problem of fitting outcome or treatment contrast functions over the entire covariate space. Examples of local modeling approaches include methods inspired by "bump hinting" (or patient rule induction) methods introduced by Friedman and Fisher (1999). Other notable examples are the SIDES (Lipkovich et al., 2011) and SIDEScreen methods (Lipkovich and Dmitrienko, 2014).

7.3. *Subgroup confirmation*

As emphasized in Section 7.2, a very important aspect of subgroup investigation is reproducibility assessment which is aimed at evaluating credibility of the identified subgroups of patients. It was stated in the EMA subgroup analysis guidance that "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" (EMA, 2014, p. 3). These factors include biological plausibility, strength of beneficial effect in patient subgroups, and replication of evidence.

Beginning with biological possibility, it is well known that post-hoc assessment of biological possibility is unreliable. In general, "plausibility" may be an unfortunate choice of terminology since almost any finding appears to be clinically plausible after the fact. Patient subgroups are commonly defined in terms of important prognostic factors, and any combination of these variables is likely to appear biologically meaningful. As a result, biological plausibility may be viewed as the weakest factor in establishing post-hoc credibility of subgroup findings. When discussing biological possibility, it is important to focus on the underlying biological mechanisms of the treatment being studied and, as emphasized in the EMA subgroup guidance, include a discussion of factors that are expected to predict treatment response along with relevant information such the probable direction of effect, in the trial protocol.

Second, it was stated earlier in this section that the assessment of strength of evidence in a given patient subgroup is virtually impossible without accounting for multiplicity or, in general, selection bias inherent in any post-hoc subgroup search. It is recognized in the clinical trial literature that failure to account for selection bias increases the chances of finding spurious subgroup effects (multiple references are available beginning with Yusuf et al., 1991). The EMA subgroup analysis guidance recognized the problem of selection bias but generally downplayed the importance of explicit multiplicity adjustments:

"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." (EMA, 2014, Section 4.3)

Further discussion of treatment effect assessment within individual patient subgroups, with emphasis on treatment effect estimates corrected for selection bias, is provided in Section 7.3.

The EMA subgroup analysis guidance emphasized the importance of the reproducibility assessments in subgroup exploration. In fact, reproducibility (replication of evidence) appears to be the only objective tool for assessing the credibility of the apparent treatment effect observed within a subgroup. When performing reproducibility assessments, it is important to remember that replication of positive findings in a subgroup or multiple subgroups is not a simple yes/no procedure. That results of reproducibility assessments are more likely to range from "the treatment effects in the same subgroup are unlikely to be consistent between the trials" to "consistency is highly plausible." The observed effect sizes in a given subgroup may vary a lot across multiple trials simply by chance, and thus sampling variability needs to be taken into account when reproducibility assessments are performed.

Subgroup confirmation is performed in a very straightforward manner when an independent dataset exists. For example, when two or more trials are included in a Phase III development program, one of the trial databases can be used for the purpose of subgroup identification and the other trial databases can be examined to verify whether or not the selected subgroups are confirmed on independent data.

When no independent datasets are available, an efficient approach to reproducibility assessment is the learn-and-confirm method which employs a cross-validation approach commonly used in data mining and other applications (Lipkovich et al., 2011). In its simplest form, the approach relies on splitting the dataset of interest (e.g., a clinical trial database) into two subsets known as the training and test subsets. It is critical to ensure that each subset is balanced with respect to key baseline characteristics, e.g., the balanced allocation procedure introduce in Lipkovich et al. (2011) can be utilized. Further, a set of promising subgroups is identified using the training subset, and the probability of confirming these subgroups in the test subset is evaluated. Examples of the learn-and-confirm approach to subgroup confirmation are provided in Dmitrienko et al. (2015a).

As indicated in Dmitrienko et al. (2015a), a simple version of the learn-and-confirm approach defined above may not produce reliable results since it uses an arbitrarily chosen single split of the original clinical trial database. It is more sensible to utilize "repeated cross-validation" in the context of the learn-and-confirm method. Repeated cross-validation is based on examining a large number of random splits that generate pairs of training and test subsets. Subgroup confirmation assessments are then performed for each pair of subsets, and the results are averaged across the pairs to reliably estimate the probability of replicating the findings in an independent dataset and bias involved in post-hoc subgroup identification. This approach is easily applied to assess reproducibility of subgroup findings in any Phase III development program, including programs with a single pivotal trial. Note that the latter case was identified as especially challenging for performing replications in the EMA subgroup analysis guidance (EMA, 2014, Section 6.4).

7.4. *Treatment effect estimation*

Multiplicity adjustments that are applied to treatment effect p-values within the individual patient subgroups can be viewed as a special case of general adjustments for selection bias in subgroup search. It was emphasized in Section 7.2 that another important aspect of selection bias adjustment is the derivation of bias-corrected estimates of the treatment effect within exploratory subgroups. Presenting subgroup results based on bias-corrected estimates of the treatment effect is very helpful for putting extreme chance findings in smaller subgroups into perspective.

The EMA subgroup analysis guidance emphasized the importance of obtaining unbiased subgroup effect estimates:

"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." (EMA, 2014, Section 4.3)

Several approaches can be applied to obtain bias-corrected estimates of treatment effects within patient subgroups. This includes Bayesian shrinkage-based methods that were mentioned in the EMA subgroup analysis guidance (EMA, 2014, Section 4.3) as well as statistical methods developed in the context of data mining and machine learning.

Shrinkage methods can be applied to derive reliable treatment effect estimates within identified subgroups. As stated in Yusuf et al. (1991), the main justification for this class of methods is that the treatment effect estimate in the overall population is usually a better guide to the effect in a subgroup than the subgroup-specific estimate. A related argument was given by Senn (2007). There is probably some truth to this, at least for exploratory subgroups without biological plausibility of a differential treatment effect. Shrinkage estimation methods combine the overall effect estimate with the estimate within in a given subgroup. They thus offer a compromise between the assumption that there is no difference across the subgroups and assumption that the subgroup effects are completely unrelated.

Shrinkage estimation is usually performed within a Bayesian framework; examples include White et al. (2005), Jones et al. (2011), and Berger et al. (2014). The EMA guidance highlighted the importance of clearly stating the assumptions underlying shrinkage-based estimation methods (EMA, 2014, Section 4.3). Indeed, these methods assume some form of prior distribution for the interaction effect. The prior distribution is usually centered at zero, meaning no difference in subgroup effects. As a result, the amount of shrinkage depends mainly on the prior variance. A smaller prior variance causes the shrinkage estimate of the treatment effect within a subgroup to be closer to the overall population estimate, and the choice of the prior variance is thus most important. Especially for labeling decisions, a good understanding of the influence of the prior will be required.

Another popular approach to computing bias-corrected estimates of the treatment effect in patient subgroups identified using an appropriate subgroup search algorithm originated in the data-mining literature. This approach relies on cross-validation as well as parametric and non-parametric bootstrap procedures. For example, Foster et al. (2011) proposed a number of bias-correction methods for the Virtual Twins subgroup identification procedure. See Lipkovich et al. (2015) for a discussion of bias-corrected treatment effect estimates based on several subgroup identification methods.

8. Other important considerations in subgroup analysis

This section reviews a number of important issues arising in subgroup investigations based on post-baseline covariates, subgroup analysis in clinical trials with noninferiority objectives, treatment of unbalanced subgroups, and assessment of patient subgroups in the context of meta-analysis.

8.1. *Patient subgroups based on post-baseline patient characteristics*

An important practical issue concerns the definition of patient subgroups based on patient characteristics measured after the start of study drug administration. The EMA subgroup analysis guidance gives the following advice on this topic:

"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" (EMA, 2014, Section 4.1).

This advice is sensible and will be illustrated with an example. A typical case of a subgroup analysis based on post-baseline characteristics is the investigation of the potential effect of changes in a patient characteristic (biomarker) on certain clinical events. Such analyses can be useful to identify or confirm prognostic markers or surrogate endpoints, see also Burzykowski et al. (2005). However, this investigation should focus on the overall assessment of the biomarker and event of interest irrespective of the assigned treatment rather than on a comparison of treatment effects within specific subgroups. For example, consider a two-arm clinical trial (novel treatment versus control) and assume that a subgroup of responders and non-responders has been defined based on a change from baseline in a certain biomarker. Assume further that the biomarker is prognostic, i.e., a higher rate of undesirable clinical events is observed at a later time point in the subgroup of non-responders. In addition, there is a higher response rate in the experimental treatment arm compared to the control arm and a negative treatment effect is detected in the subgroup of non-responders, i.e., treated patients are doing worse than patients in the control arm. This could be due to either ineffective treatment or due to the fact that only very sick patients were non-responders in the experimental treatment arm. It is very easy to be misled and incorrectly interpret the results if such issues are not considered appropriately. In general, subgroups defined based on post-baseline characteristics might occasionally be useful, but they do not represent randomized treatment comparisons and should be interpreted with caution.

8.2. *Subgroup analysis in noninferiority trials*

It is important to briefly review issues that are specific to subgroup assessment in clinical trials with a noninferiority objective. The question under discussion is whether it is reasonable to adjust the noninferiority margin, which was prospectively defined for the overall population analysis based on clinical and statistical reasoning, in the analysis of subgroups of patients. The EMA subgroup analysis guidance emphasizes that heterogeneity of the treatment effect is very likely to be influenced by certain covariates and expects a thorough discussion of their biological plausibility (EMA, 2014). If heterogeneity is indeed expected and appears biologically plausible, it is difficult to justify a single noninferiority margin that applies to all possible subsets of the overall population. In fact, using a single universal margin is only justifiable when the treatment effect is homogeneous across the trial's population and/or additional conditions are met. For example, the placebo effect needs to be inferred in two-arm clinical trials with an active control. Subgroup-specific margins may be used in noninferiority tests if historical data suggest a differential placebo effect in certain subgroups or the effect of the active control is expected to be affected by baseline patient characteristics.

8.3. *Unbalanced patient subgroups*

Additional challenges arise if patient subgroups are unbalanced. Unbalanced subgroups may be caused by different factors, and it is relevant to distinguish between the situations where the number of treated patients varies across subgroups or subgroup sizes differ. The imbalanced treatment allocation can be minimized if stratified randomization is employed or if the individual subgroups are sufficiently large, which reduces the chances of an imbalanced treatment allocation. Consequently, stratified randomization should be considered for important factors even if the number of factors is limited for practical reasons. General recommendations related to handling subgroup imbalance and limitations of subgroup assessments in these settings are discussed in the EMA subgroup analysis guidance (EMA, 2014, Sections 5.1–5.3 and 6) and EMA guideline on adjustment for baseline covariates in clinical trials (EMA, 2015).

In other situations where the sizes of subgroups of interest differ, the imbalance is mostly likely caused by the presence of certain criteria in the overall population of patients from which the trial's population is selected. Even if this does not directly lead to bias, provided the treatment allocation is not unbalanced, the power of statistical tests is likely to be remarkably reduced, e.g., for the classical interaction test. However, the likelihood of an imbalanced treatment allocation is increased in smaller subgroups, which further complicates the interpretation of subgroup analysis findings and may require additional adjustments for hypothesis tests and other inferences. Unfortunately, it is impossible to predefine adjusted analyses for all potential small subgroups in advance. Furthermore, the validity of inferences from such analysis is questionable as discussed in Grouin et al. (2005) and Senn (1994).

8.4. *Meta-analysis*

A frequently discussed question is whether meta-analytical approaches can be utilized to correct chance findings from subgroup assessments in a single clinical trial. Since meta-analysis traditionally considers heterogeneity across trials, multiple approaches are available that can be easily applied to perform heterogeneity assessment across subsets of the patient population within a trial. An overview of different methodological approaches from the field of meta-analysis applied to subgroup analysis is provided in Borenstein et al. (2009, Chapter 19). However, it is stressed in this monograph that meta-analytical methods focus on statistical rather than clinical significance. Consequently, these methods can only confirm or weaken the signal from a single trial. The clinical relevance of the signal needs to be discussed on a case-by-case basis. Further, as mentioned in Section 7.3, replication of findings is especially important in the context of post-hoc subgroup analysis, and it is already quite reassuring if results in several trials go into the same direction.

9. Discussion

Even though most clinical trials are aimed at demonstrating effectiveness and safety of new treatments in the overall population, analysis of treatment defects in subsets of the overall population is playing an increasingly important role in late-stage trials. It is well known that the magnitude of the treatment effect across subsets of the trial population is likely to be affected by multiple baseline patient characteristics. The main goal of subgroup analysis is to provide a comprehensive investigation of patient characteristics in order to provide the most relevant information on the effectiveness as well as safety of a new treatment in the overall population as well as key subpopulations. It is therefore of paramount importance to apply scientifically sound approaches to subgroup investigation and confirmation, especially in the context of exploratory subgroup analysis.

This article reviewed the most important statistical issues related to the analysis of subgroup effects and their interpretation. It focused on reviewing relevant literature, including recently published regulatory guidelines, and defining practical guidelines for scientifically sound subgroup assessment in late-stage clinical trials.

To facilitate the discussion of commonly used subgroup analysis methods, a simple classification of different approaches to subgroup assessment, including exploratory and confirmatory approaches, was introduced. The classification scheme includes three main classes of statistical methods labeled Category A (prospectively defined inferential subgroup analysis), Category B (prospectively defined exploratory subgroup analysis), and Category C (post-hoc exploratory subgroup analysis).

In the context of confirmatory subgroup analysis, this article reviewed key trial design and analysis considerations in tailoring clinical trials. This includes a multi-population approach which supports investigation of treatment effects across several prospectively defined patient populations within a single trial. The multi-population approaches can be applied in a fixed-design and adaptive-design settings. Adaptive designs serve as an efficient tool for refining the patient population compared to a conventional approach of conducting a series of trials where the patient population may be modified based on the outcomes observed in the preceding trials.

Statistical methods in Category B include approaches aimed at assessing homogeneity of the treatment effect across subsets of the overall trial population. Consistency assessments are commonly performed based on clinically relevant factors that are expected to affect the outcome variable or treatment effect, e.g., tumor stage in oncology trials or gender in cardiovascular trials. Consistency evaluation supports the conclusion that the experimental treatment is effective in a broad patient population. If the treatment effect is shown to be inconsistent across levels of important factors, routine consistency assessments may lead to a comprehensive subgroup exploration based on an appropriate subgroup identification method.

An important class of subgroup analysis methods is post-hoc subgroup identification methods (Category C approaches). As emphasized in this article, disciplined approaches to subgroup discovery are based on a solid analytical foundation and have found multiple applications in late-stage clinical trials, e.g., they play a central role in the development of tailored therapies. An important feature of principled subgroup identification methods is that they minimize the probability of chance finding in post-hoc subgroup analysis. Popular subgroup identification methods often generate a set of promising patient subgroups with desirable characteristics. The subgroups can be ordered using clinically relevant criteria. This approach facilitates a benefit/risk assessment and enables the trial's sponsor to select a subgroup for further investigation which satisfies multiple criteria, for example, criteria based on the magnitude of treatment effect within the subgroup, subgroup's size, and safety profile.

References

Alosh, M., Huque, M. F. (2013). Multiplicity considerations for subgroup analysis subject to consistency constraint. *Biometrical Journal* 55:444–462.

Alosh, M., Huque, M. F., Koch, G. (2015). Statistical perspectives on subgroup analysis: Testing for heterogeneity and evaluating error rate for the complementary subgroup. *Journal of Biopharmaceutical Statistics* 25:1161–1178.

Berger, J., Wang, X., Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* 24:110–129.

Bonetti, M., Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5:465–481.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: Wiley.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy on oncology. *Statistics in Medicine* 28:1445–1463.

Burzykowski, T., Molenberghs, G., Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York, NY: Springer.

Carroll, K. J., Le Maulf, F. (2011). Japanese guideline on global clinical trials: Statistical implications and alternative criteria for assessing consistency. *Drug Information Journal* 45:657–667.

CFDA. (2007). Provisions for drug registration. State Food and Drug Administration (SFDA) Order No. 28.

CFDA. (2015). Guidance for international multicenter clinical trials (IMCT). China Food and Drug Administration Announcement No. 2 [2015] trial implementation.

Dmitrienko, A., D'Agostino, R. B., Huque, M. F. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine* 32:1079–1111.

Dmitrienko, A., D'Agostino, R. B. (2013). Tutorial in biostatistics: Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 32:5172–5218.

Dmitrienko, A., Lipkovich, I., Hopkins, A., Li, Y. P., Wang, W. (2015). Biomarker evaluation and subgroup identification in a pneumonia development program using SIDES. In *Applied Statistics in Biomedicine and Clinical Trials Design*, Chen, Z., Liu, A., Qu, Y., Tang, L., Ting, N., and Tsong, Y., eds, pp. 427–468. New York, NY: Springer.

Dmitrienko, A., Millen, B., Lipkovich, I. (2015). Statistical and regulatory considerations in subgroup analysis. *Statistics in Medicine*. To appear.

EMA. (2002). Points to consider on multiplicity issues in clinical trials. European Medicines Agency/Committee for Proprietary Medicinal Products. CHMP/EWP/908/99.

EMA. (2003). Points to consider on adjustment for baseline covariates. European Medicines Agency/Committee for Proprietary Medicinal Products. EMA/CPMP/EWP/2863/99.

EMA. (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency/Committee for Medicinal Products for Human Use. CHMP/EWP/2459/02.

EMA. (2014). Guideline on the investigation of subgroups in confirmatory clinical trials. Draft. European Medicines Agency/Committee for Medicinal Products for Human Use. EMA/CHMP/539146/2013.

EMA. (2015). Guideline on adjustment for baseline covariates in clinical trials. European Medicines Agency/Committee for Medicinal Products for Human Use. EMA/CHMP/295050/2013.

FDA. (2012). *Guidance for Industry: Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products*. U.S. Food and Drug Administration.

Foster, J. C., Taylor, J. M. C., Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30:2867–2880.

Freidlin, B., McShane, L. M., Polley, M. Y., Korn, E. L. (2012). Randomized Phase II trial designs with biomarkers. *Journal of Clinical Oncology* 30:3304–3309.

Friede, T., Parsons, N., Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 31:4309–4320.

Friedman, J. H., Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* 9:123–143.

Grouin, J. M., Coste, M., Lewis, J. (2005). Subgroup analyses in randomized clinical trials: Statistical and regulatory issues. *Journal of Biopharmaceutical Statistics* 15:869–882.

Gunter, L., Zhu, J., Murphy, S. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the familywise error rate. *Journal of Biopharmaceutical Statistics* 21:1063–1078.

Hemmings, R. (2014). An overview of statistical and regulatory issues in the planning, analysis, and interpretation of subgroup analyses in confirmatory clinical trials. *Journal of Biopharmaceutical Statistics* 24:4–18.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15:651–674.

Hothorn, T., Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics* 64:1263–1269.

Huque, M., Röhmel, J. (2009). Multiplicity problems in clinical trials: A regulatory perspective. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko, A., Tamhane, A. C., and Bretz, F., eds. New York, NY: Chapman and Hall/CRC Press.

ICH. (1998). Ethnic factor in the acceptability of foreign data. ICH E5 Expert Working Group. *The US Federal Register* 83:31790–31796.

ICH. (1999). Statistical principles for clinical trials: ICH harmonized tripartite guideline. ICH E9 Expert Working Group. CPMP/ICH/363/96.

ICH. (2014). *Final Concept Paper E9* (R1): *Addendum to Statistical Principles for Clinical Trials on Choosing Appropriate Estimands and Defining Sensitivity Analyses in Clinical Trials*. ICH Steering Committee.

Imai, K., Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7:443–470.

Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless Phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 10:347–356.

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 8:129–143.

Koch, A., Framke, T. (2014). Reliably basing conclusions on subgroups of randomized clinical trials. *Journal of Biopharmaceutical Statistics* 24:42–57.

Koch, G., Schwartz, T. A. (2014). An overview of statistical planning to address subgroups in confirmatory clinical trials. *Journal of Biopharmaceutical Statistics* 24:72–93.

Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G. (2011). Subgroup identification based on differential effect search (SIDES): A recursive partitioning method for establishing response to treatment in subject subpopulations. *Statistics in Medicine* 30:2601–2621.

Lipkovich, I., Dmitrienko, A. (2014). Biomarker identification in clinical trials. In *Clinical and Statistical Considerations in Personalized Medicine*, Carini, C., Menon, S., and Chang, M., eds. New York, NY: Chapman and Hall/CRC Press.

Lipkovich, I., Dmitrienko, A., D'Agostino, R. B. (2015). Tutorial in biostatistics: Exploratory subgroup analysis in clinical trials. *Statistics in Medicine*. To appear.

Loh, W.-Y., Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica* 7:815–840.

Loh, W.-Y., He, X., Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* 34:1818–1833.

Meinshausen, N., Bühlmann, P. (2010). Stability selection. *Journal of Royal Statistical Society. Series B* 72:417–473.

Millen, B., Dmitrienko, A., Ruberg, S., Shen, L. (2012). A statistical framework for decision making in confirmatory multi-population tailoring clinical trials. *Drug Information Journal* 46:647–656.

Millen, B., Dmitrienko, A., Song, G. (2014). Bayesian assessment of the influence and interaction conditions in multi-population tailoring clinical trials. *Journal of Biopharmaceutical Statistics* 24:94–109.

Millen, B., Dmitrienko, A., Mandrekar, S., Zhang, Z., Williams, D. (2014). Multi-population tailoring clinical trials: Design, analysis and inference considerations. *Therapeutic Innovation and Regulatory Science* 48:453–462.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., Posch, M. (2015). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics*. To appear.

PMDA. (2012). Basic principles on global clinical trials (Reference cases). Japanese Ministry of Health, Pharmaceuticals and Medical Devices Agency.

Royston, P., Altman, D. G. (1994). Regression using factional polynomials of continuous covariates: Parsimonious parametric modeling (with discussion). *Applied Statistics* 43:429–467.

Royston, P., Sauerbrei, W. (2004). A new approach to modelling interaction between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 23:2509–2525.

Royston, P., Sauerbrei, W. (2013). Interaction of treatment with a continuous variable: Simulation study of significance level for several methods of analysis. *Statistics in Medicine* 32:3788–3803.

Ruberg, S. J., Shen. L. (2015). Personalized medicine: Four perspectives for clinical drug development. *Statistics in Biopharmaceutical Research*. To appear.

Senn, S. (2007). *Statistical issues in drug development* (Second Edition). Chichester, UK: Wiley.

Senn, S. J. (1994). Testing for baseline imbalance in clinical trials. *Statistics in Medicine* 13:1715–1726.

Simon, R. (2008). *Subgroup Analysis. Wiley Encyclopedia of Clinical Trials*. New York: John Wiley and Sons, Inc.

Stallard, N., Hamborg, T., Parsons, N., Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* 24:168–187.

Su, X., Zhou, T., Yan, X., Fan, J., Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics* 4(1), Article 2.

Tian, L., Alizaden, A. A., Gentles, A. J., Tibshirani, R. (2012). A simple method for detecting interactions between a treatment and a large number of covariates. Available at http://arxiv.org/abs/1212.2995.

Wang, S. J., Cohen, N., Katz, D. A., Ruano, G., Shaw, P.M, Spear, B. (2006). Retrospective validation of genomic biomarkers-what are the questions, challenges and strategies for developing useful relationships to clinical outcomes – Workshop summary. *Pharmacogenomics Journal* 6:82–88.

Wang, S. J., Hung, H. M. J. (2014). A regulatory perspective on essential considerations in design and analysis of subgroups when correctly classified. *Journal of Biopharmaceutical Statistics* 24:19–41.

White, I. R., Pocock, S. J., Wang, D. (2005): Eliciting and using expert opinions about influence of patient characteristics on treatment effects: A Bayesian analysis of the CHARM trials. *Statistics in Medicine* 25:3805–3821.

Yusuf, S., Wittes, J., Probstfield, J., Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association* 266:93–98.

Zhang, B., Tsiatis, A. A., Laber, E. B., Davidian, M. (2012) A robust method for estimating optimal treatment regimes. *Biometrics* 68:1010–1018.

Zhao, Y. D., Dmitrienko, A., Tamura, R. (2010). Design and analysis considerations in clinical trials with a sensitive subpopulation. *Statistics in Biopharmaceutical Research* 2:72–83.

Zhao, Y., Zheng, D., Rush, A. J., Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107:1106–1118.