EUnetHTA 21

EUnetHTA 21 – Individual Practical Guideline Document

D4.5 – APPLICABILITY OF EVIDENCE – PRACTICAL GUIDELINE ON MULTIPLICITY, SUBGROUP, SENSITIVITY AND POST HOC ANALYSES

Version 0.3, 04/07/2022

Template version 1.0, 03/03/2022

DOCUMENT HISTORY AND CONTRIBUTORS

Version	Date	Description
V0.1	**23/03/2022**	First draft for CSCQ and NC-HTAb review
V0.2	**25/05/2022**	Second draft for CSCQ and NC-HTAb review
V0.3	**04/07/2022**	Third draft for public consultation

Disclaimer

This Practical Guideline was produced under the Third EU Health Programme through a service contract with the European Health and Digital Executive Agency (HaDEA) acting under mandate from the European Commission. The information and views set out in this Practical Guideline are those of the author(s) and do not necessarily reflect the official opinion of the Commission/Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission /Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained herein.

Participants

Hands-on Group	Gemeinsamer Bundesausschuss [G-BA], Germany Haute Autorité de Santé [HAS], France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWIG], Germany Norwegian Medicines Agency [NOMA], Norway
Project Management	Zorginstituut Nederland [ZIN], the Netherlands
CSCQ	Agencia Española de Medicamentos y Productos Sanitarios [AEMPS], Spain Austrian Institute for Health Technology Assessment [AIHTA], Austria Belgian Health Care Knowledge Centre [KCE], Belgium Gemeinsamer Bundesausschuss [G-BA], Germany Haute Autorité de Santé [HAS], France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, [IQWIG], Germany Italian Medicines Agency [AIFA], Italy National Authority of Medicines and Health Products [INFARMED], Portugal National Centre for Pharmacoeconomics [NCPE], Ireland National Institute of Pharmacy and Nutrition [NIPN], Hungary Norwegian Medicines Agency [NOMA], Norway The Dental and Pharmaceutical Benefits Agency [TLV], Sweden Zorginstituut Nederland [ZIN], The Netherlands
CEB	

The work in EUnetHTA 21 is a collaborative effort. While the agencies in the Hands-on Group will be actively writing the deliverable, the entire EUnetHTA 21 consortium is involved in its production throughout various stages. This means that the Committee for Scientific Consistency and Quality (CSCQ) will review and discuss several drafts of the deliverable before validation. The Consortium Executive Board (CEB) will then endorse the final deliverable before publication.

Copyright

TABLE OF CONTENTS

LIST OF ACRONYMS - INITIALISMS

CEB	Consortium executive board
CER	Comparisonwise error rate
CSCQ	Committee for scientific consistency and quality
EUnetHTA	European network of health technology assessment
FWER	Familywise error rate
HaDEA	European Health and Digital Executive Agency
HOG	Hands-on group
HTA	Health technology assessment
HTAb	Health technology assessment body
HTD	Health technology developer
ICE	Intercurrent event
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
JCA	Joint clinical assessment
NMA	Network meta-analysis
PICO	Population, intervention, comparator, outcome
RCT	Randomised controlled trial
SAP	Statistical analysis plan

1 INTRODUCTION

1.1 Problem statement, scope and objective of the guideline

The design and conduct of a clinical study such as a randomised controlled trial (RCT) and analysis of its results are aimed at answering specific research question(s) defined by a health technology developer (HTD). Guidelines on statistical principles that should be used in this context have been developed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) [1].

During health technology assessment (HTA) under Regulation (EU) 2021/2282 (the EU HTA regulation), joint clinical assessment (JCA) starts with the assessment scope (Article 8(6) of the regulation; see EUnetHTA 21 Practical Guideline D.4.2.1 *Scoping process*). The assessment scope, expressed as one or more PICO (population, intervention, comparator, and outcome) questions, defines the needs of member states in terms of evidence that should be submitted by the HTD. Thus, these PICO questions are research questions defined by the member states for HTA purposes.

When drafting a JCA report, the assessor and co-assessor should not include any value judgement or conclusion on the clinical added value of the health technology assessed (Article 9(1)). The report should be limited to a factual assessment of the effectiveness and the certainty of results, considering the strengths and limitations of the available evidence submitted by the HTD. The outcome of the JCA should not affect the discretion of member states to draw conclusions regarding the clinical added value of the health technology assessed (Article 9).

When assessing the clinical added value of a health technology at a national level, member states are required to give due consideration to the JCA reports published (Article 13(1)). However, different member states can consider certain methodological aspects differently, especially because they can approach consistency or mismatches between the research question(s) as investigated by the HTD and the PICO question(s) differently. These differences in approach can be related to methodological issues pertaining to multiple hypothesis testing, subgroup, sensitivity, or post hoc analyses. Therefore, adequate assessment and reporting of the necessary elements of methods and results within the JCA report are needed to allow member states to draw their own conclusions regarding the clinical added value of a health technology. Moreover, studies submitted by HTDs as evidence can be individual clinical studies and/or evidence synthesis studies. The way in which different member states consider the aforementioned methodological aspects can be impacted by this distinction. Therefore, multiplicity, subgroup, sensitivity, and post hoc analyses are discussed in this guideline from the perspective of assessing individual clinical studies and assessing evidence synthesis studies in separate sections.

The intent for this practical guideline is not to endorse a particular approach regarding appraisal of the aforementioned methodological issues when member states draw their own conclusions at the national level. Rather, the objective is to guide assessors and co-assessors in assessing and reporting all the necessary elements that member states need to carry out national assessment of the clinical added value of the health technology regarding multiple hypothesis testing and subgroup, sensitivity, and post hoc analyses. Thus, all the requirements for assessment and reporting mentioned in this guideline assume that HTDs present the necessary elements in the submission dossier.

Multiplicity, subgroup, sensitivity, and post hoc analyses are not the only methodological aspects to consider when assessing the clinical added value of a health technology. Complementary elements in the reporting and assessment of the certainty of results (internal validity, statistical precision (e.g., confidence intervals), applicability) for individual clinical studies are described in EUnetHTA 21 Practical Guideline D4.6.1 *Validity of clinical studies*, while the validity of evidence synthesis studies is covered in EUnetHTA 21 Practical Guideline D4.3.1 *Direct and indirect comparisons* (along with EUnetHTA 21 Methodological Guideline D4.3.2 *Direct and indirect comparisons*). Additional considerations regarding the definition of clinically relevant outcomes and assessment of their validity, reliability and interpretability are covered in EUnetHTA 21 Practical Guideline D4.4.1 *Endpoints*.

This guideline predominantly deals with methodological issues related to inferential statistical analyses. RCTs are the gold standard for answering clinical research questions in accordance with a hypothetico-deductive approach. Therefore, while recommendations in this guideline may be better suited for RCTs, they can apply to various study designs. For simplicity, effectiveness is the common term used here to describe efficacy or effectiveness throughout the rest of the document. Effectiveness also includes safety within the context of this document. Furthermore, treatment is used as a common term for any health technology that can be assessed.

1.2 Relevant articles in Regulation (EU) 2021/2282

Articles from Regulation (EU) 2021/2282 directly relevant to the content of this practical guideline are:

- Article 8: initiation of joint clinical assessments,
- Article 9: joint clinical assessment reports and the dossier of the health technology developer,
- Article 13: member states’ rights and obligations.

2 DEFINITIONS

In the context of this document, the terms “planned” and “prespecified” refer to a given statistical analysis as planned according to a study protocol and/or statistical analysis plan (SAP) of a study submitted as evidence by a HTD.

Mirroring the previous definition, the term “post hoc analysis” can be understood, unless stated otherwise, as a synonym for any statistical analysis that was not planned according to a study protocol and/or SAP of a study submitted as evidence by a HTD. More details are provided in Sections 9 and 10 of this document.

3 MULTIPLE STATISTICAL HYPOTHESIS TESTING IN INDIVIDUAL CLINICAL STUDIES

3.1 Purposes, definitions and general methodological considerations

In general, statistical analyses are performed for sample(s) of patients from a population of interest, as collection of data for all patients in the population is usually not feasible. Thus, the statistics that are produced are estimates and not true values for the population. Therefore, even if a clinical study is free from bias, a difference observed for an **endpoint** of interest between groups (e.g., a difference in mortality observed in an RCT) does not necessarily equate to a true difference in the population of interest because of the sampling hazard (i.e., a form of random error). Thus, the risk of wrongly claiming the existence of treatment effectiveness needs to be controlled at an acceptable level, which is achieved via **statistical hypothesis testing**.

Under the **frequentist approach**, statistical hypothesis testing involves testing of two competing hypotheses: the **null hypothesis** (usually denoted H_0) and the **alternative hypothesis** (usually denoted H_1). In a superiority setting, the null hypothesis usually involves postulating the true absence of difference in the population of interest, while the alternative hypothesis involves postulating a true difference (two-sided tests) or true superiority or inferiority (one-sided tests). Statistical hypothesis testing relies on estimating the **p value**, which is the probability of the occurrence of a difference at least as large as the one observed if the null hypothesis is true. In RCTs, statistical test results are usually interpreted under the **Neyman-Pearson approach**: the p value of a test is compared to a prespecified risk level – the **α level** – and if the p value is less than the α level, the alternative hypothesis is accepted. In biomedical research, the consensus is usually to set the α level of a (two-sided) single test at 0.05 (5%). Any statistical test can lead to two errors: rejecting the null hypothesis when it is actually true (i.e., the **type-1-error**, or false positive) or not rejecting the null hypothesis when it is actually false (i.e., the

type 2 error, or false negative) [2]. The probability α of a type 1 error for one significant test is the **comparisonwise error rate** (CER) [3].

Situations in individual clinical studies occur for which multiple statistical tests are performed, which increases the risk of at least one false-positive test. If k independent tests are performed, the probability of rejecting at least one of the k independent null hypotheses when all null hypotheses are actually true is called the global **familywise error rate** (FWER, considering the **family** of k tests as one experiment under the complete null hypothesis). When considering independent tests, FWER is equal to $1–(1–α)^k$ [3][1]. When performing multiple tests, it is not usually expected that all null hypotheses will be true simultaneously. Therefore, multiple tests procedures mentioned below usually control **FWER in a strong sense** (also called multiple level): the probability of erroneously rejecting at least one true null hypothesis, irrespective of which and how many of the individual null hypotheses are true [3]. A multiple test procedure that controls FWER in a strong sense also controls the global FWER (but not vice versa) [4]. Most of the usual procedures described below control FWER in a strong sense [4]. Thus, for the rest of the document, we use the term FWER to mean “FWER in a strong sense”.

Three main situations for which multiple statistical tests arise for individual clinical studies are considered in this guideline. First, because most diseases have more than one consequence, many clinical studies are designed to estimate the effectiveness of a treatment for more than one endpoint [5,6]. Situations in which, as part of the final analysis of an individual clinical study, the same consequence is assessed at different time points (e.g., clinical remission at 6 months and at 12 months after inclusion) or the same endpoint is assessed in different populations (e.g., intention-to-treat population and a subpopulation) are considered to be related. Indeed, these lead to multiplicity issues that can be considered similar from a methodological perspective.

Second, multiplicity issues can arise in the context of **interim analysis**. An interim analysis is any analysis used to compare treatment groups with respect to effectiveness at any time before formal completion of a trial [1]. Thus, for a given endpoint, multiple analyses can be performed at different times. Interim analyses can be planned for making decisions on whether to stop the trial early. Reasons for early stopping may include clearly established superiority of the treatment(s) of interest(s), confirmation that superiority is unlikely to occur and unacceptable adverse effects. Stopping could apply to the entire trial or to a subset (e.g., ending a treatment group or discontinuing accrual of a subgroup of patients). The benefits of interim analyses (ethical, scientific) can be opposed by methodological disadvantages such as overestimation of the treatment effectiveness, lower precision or credibility, a potential increase in type 1 errors if not appropriately managed and lower power for accounting for the effects of prognostic factors [7].

Third, clinical trials can be conducted with more than two treatment groups (this situation is called “**multiple groups**” in the rest of the document). Broadly speaking, multiple-group trials can be used to compare different treatments between each other (“all pairwise comparisons” situation) or different treatments to a reference treatment (“many to one” situation) [8].

As already mentioned, FWER increases with the number of independent tests. Hence, numerous valid **multiplicity procedures**, such as the Bonferroni method, multiple-step procedures (e.g., the Holm procedure), parametric multiple testing procedures (e.g., the Dunnett procedure), hierarchical test sequences, α allocation, gatekeeping strategies and α spending functions, were developed to control FWER at an acceptable level (e.g., at a global level of 5%) [9]. These procedures differ in the way in which the algorithm for decision-making (i.e., rejecting or accepting the null hypotheses) is defined, the purpose of the analyses (e.g., interim analyses, analysis of multiple endpoints) and the balance they achieve between FWER control and loss of power (i.e., the probability of rejecting a false null hypothesis).

[1] If multiple tests are dependent (e.g., they assess correlated endpoints using the same data), there is no easy way to compute the theoretical FWER as it depends on the correlation structure between the different tests, but a high FWER can be expected anyway if many tests are performed.

With the **Bayesian inference** approach, decision-making does not rely on proving whether a null hypothesis is false and is instead guided by estimation of the distribution of treatment effect for the endpoint(s) of interest(s). This distribution, called the posterior distribution, is estimated by combining data from previous knowledge about the endpoint of interest (operationalised as the prior distribution) and data obtained by conducting the clinical study (operationalised as the likelihood). Decision-making, or stopping rules for interim analyses (i.e., claiming that a meaningful effect does or does not occur), can then be made by estimating whether the posterior probability is higher or lower than a prespecified threshold, which can have multiple definitions. For example, the posterior distribution of a risk ratio can be used to estimate if its real value has high probability (e.g., 97.5%) of being less than 1, or less than a value that is considered clinically meaningful [10]. Thus, the relevance of concepts such as the type 1 error for Bayesian inference is a matter of debate [11]. Nonetheless, as RCTs are designed to answer a specific research question in a binary manner (i.e., concluding if there is a true effect or not), some consider that controlling for the risk of false-positive conclusions still applies even if data are analysed using Bayesian inference [12]. Thus, methods for adjusting the threshold for interpreting the results of a Bayesian RCT while controlling for a desired FWER level when multiple hypotheses are tested have been proposed [13].

3.2 Requirements for appropriate reporting of methods and results in a JCA

3.2.1 Multiple outcomes

Prospective specification of all data analyses that are performed to test hypotheses about the prespecified endpoints, including the choice of multiplicity procedures, either before initiation of a clinical study, or at least before database lock (i.e., planned analyses), is considered an essential element of an adequate hypothetico-deductive approach. This helps in avoiding data dredging and ultimately helps in controlling the type 1 error rate.

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time points, measurement, method of assessment).
- Null and alternative hypotheses that are tested.
- How the endpoints were tested (statistical methods), including, if relevant, the multiplicity procedure that was used, the order of testing (if a hierarchical test procedure was used), the desired FWER level and which FWER was controlled (global level or multiple level).
- The α level used to determine if the study was a success.
- The CER level for each statistical test (i.e., the significance level required for each test).
- The results, with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).
- For the results for a given test, whether the test was appropriately controlled for multiplicity and if it was a planned analysis or not.

3.2.2 Interim analyses

As already mentioned, the design of planned analyses, including interim analyses, is considered an essential element of an adequate hypothetico-deductive approach.

Interim analyses are conducted with a definite cutoff date, usually defined as the occurrence of a specific number of events of interest (e.g., number of deaths for overall survival analyses). Statistical analyses of the corresponding database cannot be initiated before timely quality control and data management. Data and corresponding interim analyses become available with a time lag between the cutoff date and the clinical study report date (i.e., date of validation of the report of statistical analyses). However, the appropriate date to report when assessing an interim analysis should be its appropriate cutoff date and not the clinical study report date. Several interim analyses can be reported in the same clinical study report.

Interim analyses can lead to early stopping of a clinical study. Results for interim analysis and the decisions regarding clinical study continuation should be reported.

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time points, measurement, method of assessment).
- Schedule of all interim analyses.
- Respective cutoff date, with corresponding follow-up (the clinical study report date should not be used when reporting interim analyses in a JCA report).
- How the endpoints were tested (statistical methods), including the method chosen for controlling for multiplicity, for example, a triangular design, a group sequential design and its boundaries (e.g., Pocock, O’Brien-Fleming), α spending designs and their boundaries, and the desired FWER level.
- The α level used to determine if the study was a success.
- The CER level for each statistical test (i.e., the significance level required for each test).
- Implications (or not) and recommendations from an independent committee (e.g., data and safety monitoring board, data and safety monitoring committee).
- Any consequence of interim analyses for the conduct of the clinical trial (modification of study protocol, continuing or early stopping, no change, data release).
- The results, with appropriate statistics (position and dispersion indices in the intervention and comparator group, appropriate effect measure, p-value of the corresponding test and appropriate measures of statistical precision).
- For the results for a given test, if it was appropriately controlled for multiplicity and if it was a planned analysis or not.
- When unplanned interim analyses were conducted, why they were deemed necessary and by whom (sponsor or regulatory body).

3.2.3 More than two treatment groups

As already mentioned, planned analyses, including multiple-group comparison, are considered an essential element of an adequate hypothetico-deductive approach.

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time points, measurement, method of assessment).
- Null and alternative hypotheses that are tested, with an unambiguous definition of the comparator.
- How the endpoints were tested (statistical methods), including, if relevant, the multiplicity procedure that was used, the desired FWER level, and which FWER was controlled (global or multiple level).
- The α level used to determine if the study was a success.
- The CER level for each statistical test (i.e., the significance level required for each test).
- The results, with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).
- For the results for a given test, if it was appropriately controlled or not for multiplicity and if it was a planned analysis or not.

4 MULTIPLE STATISTICAL HYPOTHESIS TESTING IN EVIDENCE SYNTHESIS STUDIES

4.1 *Purposes, definitions and general methodological considerations*

When conducting evidence synthesis studies such as pairwise meta-analysis (i.e., synthesis of direct evidence (multiple head-to-head RCTs) for when exactly two treatments are compared) or analysis of more complex evidence networks (see EUnetHTA 21 Methodological Guideline D4.3.2 *Direct and indirect comparisons*) via network meta-analysis (NMA), multiplicity issues can arise in a multiple of ways. These issues can be similar to those encountered when dealing with individual clinical studies or they can be specific to the design of an evidence synthesis study [14]. However, the possibilities and necessities to deal with multiplicity in evidence synthesis studies are limited because the data are already observed. Therefore, it is not possible to plan for multiplicity adjustments in a strong confirmatory sense.

4.2 *Requirements for appropriate reporting of methods and results in a JCA*

4.2.1 Multiple outcomes

As evidence synthesis can concern a wide range of outcomes, it is usually recommended to prespecify before data extraction the outcome(s) that will be of interest to collect information about these outcomes only [14,15]. However, because users of evidence synthesis studies have heterogeneous interests in the consequence of a medical condition, evidence synthesis studies are frequently performed with the inclusion of all outcomes that are likely to be of importance. Moreover, decisions on including outcomes encountered during the data extraction process are frequently made. Thus, an unambiguous definition of what constitutes a family of tests in the context of evidence synthesis studies is difficult to achieve. Therefore, while the procedures mentioned in Section 3.1 for dealing with multiple statistical hypothesis testing can theoretically be used (some require access to individual patient data, while those based on p values can be performed using aggregated data only), in practice this is almost never the case [14].

Requirements for JCA reporting

- For evidence synthesis studies, control for multiplicity for dealing with multiple outcomes is a possibility but should not be expected.
- Accurate and unambiguous endpoint definitions (time point, measurement, method of assessment).
- Null and alternative hypotheses that were tested.
- How the endpoints were tested (statistical methods).
- If control for multiplicity was performed, how it was performed and the desired FWER level, and which FWER was controlled (global or multiple level).
- The CER level for each statistical test.
- The results (overall estimation at the evidence synthesis level) with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).
- For the results for a given test, whether the outcome was prespecified before data extraction or not.
- If control for multiplicity was performed, if it was appropriately conducted or not.

4.2.2 More than two treatment groups

Evidence synthesis comparing the relative effectiveness of more than two interventions using aggregated data are performed under the general framework for NMA. This framework can allow, if an appropriate network of evidence can be constructed, simultaneous estimation of the relative effectiveness of all pairwise comparisons of treatments included in the network, or at least for multiple two-by-two comparisons. Thus, an issue with multiple hypothesis testing may arise in NMA as several groups are observed in such a framework. Methodologically, how the potential issue of multiple hypothesis testing should be addressed when this type of evidence synthesis is performed is currently a matter of debate, and usually no multiplicity procedures are applied when estimating the relative effectiveness of the different treatments compared using an NMA [16]. Nonetheless, in the context of JCA, the assessment scope, expressed by the PICO question(s), defines the relevant comparator(s) for relative effectiveness assessment. Therefore, in a JCA, multiplicity due to multiple groups in evidence synthesis studies is not the main problem if the relevant comparison is the new intervention versus one control. When only one effect estimate of an NMA is of interest (and all the other effect estimates are only reported for completeness), multiplicity problems are therefore not an issue.

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time point, measurement, method of assessment).
- Null and alternative hypotheses that were tested.
- How the endpoints were tested (statistical methods).
- The CER level for each statistical test.
- The results (overall estimation at the evidence synthesis level) with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).

4.2.3 Multiple time points

When performing an evidence synthesis study, it is possible that outcomes are measured at different time points depending on the time points defined or the follow-up duration in the original studies that are pooled. It is also possible to perform multiple evidence syntheses for the multiple time points available. Thus, multiplicity issues due to multiple time points for analysis can arise in evidence syntheses.

A solution to limit the issue of multiple time points can be to choose a single time point for the analysis. However, this is only feasible if comparable time points are available from the studies included. Whether time points are comparable strongly depends on the clinical indication and the treatment assessed in the evidence synthesis. A different solution to the problem of different time points is to use a summary effect measure over time, such as repeated-measures analysis of variance for continuous outcomes or Cox regression for time-to-event data in the single studies [14]. The evidence synthesis can then be performed by using the estimates of the summary effect measure (e.g., the hazard ratio), avoiding multiplicity due to multiple time points. If individual patient data are available, methods for dealing with multiple time points can be used directly in the evidence synthesis, such as NMA for survival data with fractional polynomials to estimate, for example, the difference in restricted mean survival time for a selected time point (see EUnetHTA 21 Practical Guideline D.4.3.1 *Direct and indirect comparisons* and EUnetHTA 21 Methodological Guideline D.4.3.2 *Direct and indirect comparisons*).

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time point, measurement, method of assessment).
- If one common time point has been chosen for analysis for an endpoint, whether it was prespecified or not before data extraction and if choice of this common time point was justified (with its justification).
- Null and alternative hypotheses that were tested.
- How the endpoints were tested (statistical methods), especially methods that were used to estimate summary effect measures based on multiple time points.
- The CER level for each statistical test.
- The results (overall estimation at the evidence synthesis level) with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).

4.2.4 Multiple operationalisations and multiple effect measures

Multiplicity might arise in evidence synthesis because of the various ways available for analysing an endpoint and operationalising an endpoint. In evidence synthesis studies, different methods for analysing the same endpoint may be used to assess the robustness of a result, for example, via sensitivity analyses (see Section 8). However, in such cases the effect measure or operationalisation primarily used for the assessment should be prespecified [14].

For continuous data, a common effect measure is the mean difference. If the studies included in the evidence synthesis measure the same endpoint using a different operationalisation (e.g., level of depression using different depression scales), the effect measure can be standardised on a common metric. If individual patient data are available, another option could be to dichotomise continuous data, for example, as responders versus non-responders. However, the threshold for assignment as a responder or non-responder must be scrutinised regarding whether it was prespecified before data extraction and if it corresponds to validated and consensus cutoff values. For binary data, common effect measures are the risk ratio, odds ratio and absolute risk difference. If analysis of data with different effect measures leads to different results and the conclusion from the evidence synthesis would differ depending on the effect measure used, multiplicity becomes problematic. Therefore, the effect measure should be prespecified before data extraction and appropriate for the type of data being analysed [14]. Nevertheless, as member states may have different requirements regarding the effect measure for their national assessment of the health technology, it cannot be ruled out that multiple effect measures for an outcome need to be reported in the JCA.

Likewise, the analysis of several operationalisations for an endpoint (e.g., number or patients with event, time-to-event, event rate) may lead to issues with multiple hypothesis testing. Similar to the situation for multiple effect measures, member states may have different requirements regarding the operationalisation for their national assessment and several operationalisations for an endpoint may need to be reported in the JCA report.

Requirements for JCA reporting

- Accurate and unambiguous endpoint definitions (time points, measurement, method of assessment).
- If one operationalisation and/or effect measure was chosen for a specific endpoint, whether it was prespecified before data extraction or not and if it was justified (along with its justification).
- If an endpoint that was continuous in the original individual studies has been dichotomised for analysis in an evidence synthesis study, whether the threshold for dichotomisation was planned before data extraction or not, along with the justification for choosing this threshold.
- Null and alternative hypotheses that were tested.
- How the endpoints were tested (statistical methods) and, if performed, the multiplicity procedure that was used and the desired FWER level.
- The CER level for each statistical test.
- The results (overall estimation at the evidence synthesis level), with appropriate statistics (position and dispersion indices in the intervention and comparator groups, appropriate effect measure, p value for the corresponding test and appropriate measures of statistical precision).

5 SUBGROUP ANALYSES IN INDIVIDUAL CLINICAL STUDIES

5.1 Purposes, definitions and general methodological considerations

Patients may respond differently to treatments because of demographic factors, disease characteristics, comorbidities, environmental aspects, or characteristics related to other treatments, such as pre-treatment or concomitant treatment. To examine whether an estimated overall effect in a single study is driven by a specific patient group, subgroup analyses are conducted.

The term **subgroup** refers to a subset of the clinical trial population defined by one or more specific patient characteristics measured at baseline. Postbaseline factors are not appropriate for defining subgroups for the investigation of a treatment effect as they may be affected by the treatment itself received by patients during the study. The term subgroup is not to be confounded with the term **subpopulation**, which is defined as a subset of the patient population targeted as described in the therapeutic indication. Subpopulations of interest may be specified during the assessment scope (see EUnetHTA Practical Guideline D4.2.1 *Scoping process*) and are analysed as separate PICOs.

An **effect modifier** is a variable that modifies a treatment effect, that is, a variable that alters the relative effectiveness between two treatments. Effect modifiers may be patient characteristics as listed above, for example. Variables that represent methodological characteristics of a study are not regarded as potential effect modifiers and therefore their potential impact on estimating treatment effectiveness should be analysed in sensitivity analyses. Subgroup analyses refer to the comparison of treatment effects in the (disjoint) subgroups of a potential effect modifier. In statistical terms, an evident effect modification is referred to as an **interaction** between a treatment and the relevant variable.

<u>A priori planning of subgroup analyses</u>

Prespecification of subgroups is being encouraged in the planning of individual clinical studies as it can lend credibility to positive or negative subgroup findings. However, a priori planned subgroup analyses are often limited to the primary endpoint. From the perspective of assessment of an individual clinical study, all other subgroup analyses, such as analyses of subgroups or subgroup analyses for further endpoints not prespecified in the SAP, are unplanned analyses. These are not controlled for multiple hypothesis testing and lack statistical robustness.

Nevertheless, member states may require further subgroup analyses than those planned at the single study level for assessment at a national level (see EUnetHTA Practical Guideline D4.2.1 *Scoping* *process*). Precise investigation of these subgroups depends on the use of results from unplanned analyses at the single study level.

A particular point of attention is also the choice of cutoff value(s) for performing subgroup analyses when the characteristic that defines the subgroup is initially a continuous variable. Indeed, to comply with an adequate hypothetico-deductive approach, cutoff value(s) should be prespecified and the choice of the value should be justified with an adequate rationale. In the case of subgroup analyses performed because of the assessment scope, justification for the choice of cutoff value(s) pertains to the member state(s) that require specific subgroup analyses.

An interaction test is a requirement

When interpreting subgroup analyses, it should be considered that a statistically significant effect in one subgroup and a lack of effect or the reverse effect in another subgroup cannot be interpreted as the existence of different treatment effects between subgroups on its own. Instead, demonstration of different effects between different subgroups should be conducted using an appropriate interaction test (e.g., adequate regression or analysis-of-variance model). Within an individual clinical study, interaction can be tested on the basis of individual patient data. Different homogeneity and interaction tests have been discussed in the literature [17–20]. For this guideline, the term “interaction test” refers to all of these tests.

It should be kept in mind that owing to potential small sample sizes for subgroups, the power of interaction tests for detecting heterogeneity can be low. Furthermore, in very small sample sizes, prognostic variables (i.e., a patient characteristic that affects the outcome of interest irrespective of which treatment is received) may be unbalanced within subgroups between treatment groups if randomisation is not stratified according to the subgroup characteristic analysed [21,22]. In such cases, the unbalanced prognostic variable is therefore a confounder (i.e., a variable that affects both the treatment received and the outcome). Thus, the effect estimates within the subgroups may be biased due to confounding, and this bias can lead to different results in the different subgroups. Therefore, in the case of very small sample sizes, it cannot be ruled out that any differences detected between subgroups are caused by systematic errors such as confounding.

If for one outcome there is a difference, for example, between two age groups as well as between men and women, separate analyses would theoretically be required for each age group and for men and women (i.e., analyses of four subgroups) to interpret the results. However, such analyses are rarely available and may result in subgroups with rather small sample sizes.

5.2 Requirements for appropriate reporting of methods and results in a JCA

In the JCA report, information regarding a priori planning of subgroup analyses, consideration of multiplicity and definitions of subgroups in the protocol and SAP of the clinical studies assessed must be provided.

Requirements for JCA reporting

- Accurate and unambiguous definitions of subgroups and endpoints.
- Null and alternative hypotheses that are tested.
- How the subgroup analysis was performed (statistical methods), including the multiplicity procedure that was used, if performed, and the desired FWER level.
- When cutoff value(s) for a continuous variable were chosen for defining subgroups, whether they were prespecified and how the choice of these values was justified.
- The CER level for each subgroup analysis.
- The results (p values) of an appropriate interaction test for all subgroup analyses conducted.
- The results (appropriate estimates and effect measures for each subgroup, with a corresponding measure of statistical precision and p values for the effect in each subgroup).
- Whether each statistical test for subgroup analysis was appropriately controlled for multiplicity or not, and if it was a planned analysis or not.
- Visual presentation of the results using a forest plot is strongly encouraged.

6 SUBGROUP ANALYSES IN EVIDENCE SYNTHESIS STUDIES

6.1 Purposes, definitions and general methodological considerations

The purpose of subgroup analyses and the definition of an effect modification (an interaction) described for subgroup analyses in individual clinical studies also apply to evidence synthesis studies.

The term subgroup should refer to a subset of the patient population included in the evidence synthesis defined by one or more specific patient characteristics measured at baseline. In an evidence synthesis study, the individual studies included may represent subgroups if they included patients with specific characteristics of the respective subgroup.

A priori planning of subgroup analyses

In an evidence synthesis, subgroup analyses that were planned in each of the studies included are mostly not available. Nonetheless, subgroup analyses can be planned within the study protocol and/or SAP of a specific evidence synthesis study. In addition, within the assessment scope, subgroups may be defined together with the PICO framework.

An interaction test is a requirement

Within an evidence synthesis, the results from several studies can be summarised via meta-analyses. To investigate whether an estimated overall effect in meta-analysis is driven by a specific patient group, common tests for heterogeneity (in this case, heterogeneity between subgroups rather than studies) or meta-regression may be considered. In the case of evidence synthesis performed using individual patient data, for example, an adequate regression or analysis-of-variance model with a corresponding interaction term can be used. When only aggregated data are available, a Q test in a meta-analysis and an F test in a meta-regression are examples of appropriate tests for interaction. As for subgroup analyses in single studies, statistical tests for interaction may have low power and may not be sufficient to exclude the possibility of meaningful subgroup interactions.

In a meta-regression, the statistical association between the effect sizes in individual studies and the study characteristics is investigated, so that study characteristics can possibly be identified that explain the different effect sizes, that is, the heterogeneity. However, it is important that the limitations of such analyses are considered when interpreting any results. Meta-regressions that attempt to show an association between the different effect sizes and the average patient characteristics in individual studies are subject to the same limitations as the results from ecological studies in epidemiology [23].

The high risk of bias in such analyses based on aggregated data cannot be balanced by adjustment. An alternative approach is therefore the use of individual patient data, as meta-analyses that include individual patient data generally provide greater certainty of results, that is, more precise results not affected by ecological bias [24,25]. If heterogeneity is plausible, it can threaten the certainty of results associated with an evidence synthesis study. More information on this issue can be found in EUnetHTA 21 Practical Guideline D4.3.1 *Direct and indirect comparisons*.

6.2 Requirements for appropriate reporting of methods and results in a JCA

The requirements for appropriate reporting of methods and results from evidence syntheses are similar to those for single studies. In general, a priori planned subgroup analyses should not be replaced by unplanned analyses. All analyses should be reported.

Requirements for JCA reporting

- Accurate and unambiguous definitions of the subgroups and endpoints are required, as well as the null and alternative hypotheses that were tested.
- How the subgroup analysis was performed (statistical methods), including the multiplicity procedure that was used, if performed, and the desired FWER level.
- When cutoff value(s) for a continuous variable were chosen for defining subgroups, whether they were prespecified and how the choice of these values is justified.
- The CER level against which each subgroup analysis was performed.
- The results (p values) of an appropriate interaction test for all subgroup analyses conducted.
- The results (appropriate estimates and effect measures in each subgroup, with a corresponding measure of statistical precision and p values for the effect in each subgroup).
- Whether each statistical test for subgroup analysis was appropriately controlled for multiplicity or not, and if it was a planned analysis or not.
- Visual presentation of the results using a forest plot is strongly encouraged.

7 SENSITIVITY ANALYSES IN INDIVIDUAL STUDIES

7.1 Purposes, definitions, and general methodological considerations

Sensitivity analyses are an integral part of the reporting of clinical study results and are essential in investigating the robustness of the effect observed in the clinical study to variations in the assumptions and their impact.

In any clinical trial, the primary **estimand** should be defined according to the principles outlined in ICH E9 and its addendum (E9(R1)) [1,26]. The aim of the estimand framework is to define "*a precise description of the treatment effect reflecting the clinical question posed by the trial objective"*. It summarises at a population level what the outcomes would be in the same patients under different treatment conditions being compared. The statistical analyses should be aligned to the estimand (not vice versa) and sensitivity analyses should be planned in the study protocol to "*explore the robustness of inference from the main estimators to deviations from its underlying modelling assumptions and limitations of the data*" [26].

The estimand is defined by its five attributes: (1) population, (2) treatment, (3) variable (endpoint), (4) **intercurrent events** (ICEs) and (5) the summary measure. ICEs (events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest) should be addressed when describing the clinical question of interest to precisely define the treatment effect that is to be estimated. ICEs are context-dependent; the same event can be defined as **missing data** in one setting and as an ICE in another. For examples see the ICH

E9(R1). Sensitivity analyses (i.e., a series of analyses conducted with the intent of exploring the robustness of inferences from the main estimator to deviations from its underlying modelling assumptions and limitations in the data) should be used to explore the impact that changes to the assumptions for any or all of these elements might have on the primary outcome of a study [26]. In addition, the ICH E9 Addendum differentiates between sensitivity analyses and supplementary analyses (i.e., a general description for analyses that are conducted in addition to the main and sensitivity analysis with the intent of providing additional insights into understanding the treatment effect) [26]. The JCA should indicate clearly which analyses are primary, sensitivity or supplementary analyses.

Focus should be given to the difference between missing data (i.e., data that would be meaningful for analysis of a given estimand but were not collected) and ICEs and their handling in analyses. Indeed, missing data should be distinguished from data that do not exist or data that are not considered meaningful because of an ICE. Guidelines on handling of missing data are available [26,27] and describe appropriate sensitivity analysis strategies (see the definitions above). Handling of missing data (e.g., missing laboratory assessments) is considered a statistical problem that needs to be addressed via appropriate statistical analyses with the aim of explore the impact of the level of missing data on the basis of certain assumptions [27]. Avoiding missing data is considered of utmost importance, although some degree of missing data should be anticipated in any clinical study. Therefore, appropriate handling of this issue should be predefined. The methodologies chosen should be reported in the JCA.

Continued collection of data even after ICEs (e.g., treatment discontinuation or initiation of a rescue medication) to support assessment of their impact on the clinical questions is highly supported and different strategies to do so are described in the ICH E9 addendum [26]. Again, the JCA should report on the strategy for the primary and any additional estimands and the strategies chosen for handling of ICEs.

7.2 Requirements for appropriate reporting of methods and results in a JCA

The study protocol and SAP should always be submitted to allow assessment of the estimand strategy. Results should be presented according to the prespecified analyses based on the estimand framework in the study protocol as well as the strategies for handling missing data and accompanying analyses, and this should be reflected in the JCA.

There is no rule for the amount of missing data that is considered acceptable. Therefore, reports should highlight the uncertainty with respect to the amount as well as the handling of missing data. The acceptability of missing data is subject to member state differences in interpretation of their relevance within their respective decision-making process.

The primary estimand describes the objective of the study via definition of the five attributes. The objectives of studies might therefore be aligned with a certain PICO or not. If a secondary estimand better aligns with other relevant PICO(s), this should be highlighted in the report and the JCA should be clear in distinguishing the different estimands. Because estimands describe the treatment in the context of the attributes, it is possible that different HTAs could also prefer different estimands. This situation might be rare but is addressed in the PICO scoping process and should then be reflected in the report. If secondary estimands have less statistical rigour (because they are based on outcomes not included in the inferential testing strategy), this should be clearly highlighted in the report.

The acceptability of sensitivity analyses is subject to member state differences in interpretation of their relevance within their respective decision-making process.

Requirements for JCA reporting

- The JCA should present the relevance of the chosen estimand with respect to the original trial protocol as well as relevant PICO(s).
- There should be a detailed description of the chosen estimand(s), with a focus on the five attributes as well as the ICE strategy.
- Strategies for handling of ICEs are distinct from strategies to handle missing data and these differences should be clearly conveyed.
- Sensitivity and supplementary analyses should be distinguished from the primary estimand(s) and its analyses.
- There should be a clear definition of the purpose and the underlying assumption for each sensitivity analysis.
- All sensitivity analyses should be presented in the report, preferably as a summary table. Such table(s) should include the attribute(s) that the sensitivity analyses address as well as the analysis method used for each individual analysis and the results.
- When the results of a sensitivity analysis are not of the same directionality as for the results of the primary analysis, this should be highlighted.

8 SENSITIVITY ANALYSES IN EVIDENCE SYNTHESIS STUDIES

8.1 Purposes, definitions, and general methodological considerations

Evidence synthesis results are sensitive to the inclusion/exclusion of individual studies and sensitivity analyses can help to explore the impact of individual studies on the overall conclusions, assess the robustness of the analyses in general and confirm assumptions underlying the evidence synthesis.

Sensitivity analyses are a set of analyses estimating the same effect but with different methodology to assess the impact of different decisions compared to the primary assumptions on the analysis. These alternative decisions can pertain to the inclusion of studies (size, population and outcomes, among others), certain groups of the patient population (in range/out of range), the risk of bias or the use of fixed-effect versus random-effect models.

8.2 Requirements for appropriate reporting of methods and results in a JCA

Requirements for JCA reporting

- It should be stated whether the analysis was prespecified in the study protocol and/or SAP, was identified during the assessment process or is the result of the PICO process.
- There should be a clear definition of the purpose and underlying assumption for each sensitivity analysis.
- All sensitivity analyses should be presented in the report, preferably as a summary table. Such table(s) should include the elements the sensitivity analyses address, such as the evidence included, the eligibility criteria, the data used with the underlying assumptions and the analysis method used for each individual analysis, and the results. Sensitivity analyses can be conducted not only for single factors but also for multifactorial situations, so the report should be clear on what type of analysis has been performed.
- When the results of a sensitivity analysis are not of the same directionality as for the results of the primary analysis, this should be highlighted.

9 POST HOC ANALYSES IN INDIVIDUAL CLINICAL STUDIES

9.1 Purposes, definitions, and general methodological considerations

The term post hoc is derived from the Latin phrase *post hoc ergo propter hoc*, meaning “after this, therefore because of this”. Thus, in the strictest sense, post hoc analyses are all analyses that are performed because of the results of a previous analysis. Therefore, there can exist post hoc analyses that can be planned, such as a statistical hypothesis test performed for a particular outcome because of the results of the previous test when hierarchical test sequence procedures for controlling multiplicity issues are used. However, as mentioned earlier, the scope of this document mainly addresses unplanned post hoc analyses, as these are the ones that can be considered problematic in terms of deviation from an adequate hypothetico-deductive approach. Indeed, while planned analyses are acceptable when appropriate measures are taken regarding emerging multiplicity issues, unplanned post hoc analyses violate the principles of inferential hypothesis testing. Both the power of a study and the certainty for correctly rejecting the null hypothesis are built on the principle of defining the parameters of the hypothesis to be tested before the real data are observed.

However, during a HTA it might be desirable to obtain data for a patient subset that, for example, reflects a PICO more closely than the strategy pursued by the applicant. In principle, post hoc analyses can address all elements of the trial and not just subgroups of the population, as well as different outcome measures or statistical methods.

In such situations an explorative investigation based on post hoc–defined subgroups might be considered, reflective of the known methodological caveats. Post hoc analyses should be clearly identified as such to distinguish them from the primary analyses in the JCA.

9.2 Requirements for appropriate reporting of methods and results in a JCA

Reporting of post hoc analyses should follow the principles outlined in the European Medicines Agency guideline on the investigation of subgroups in confirmatory trials [28]. Subgroup analyses need to reflect on the heterogeneity of the overall population versus any subgroups, the consistency of results across the subgroups and the credibility of any subgroup, which is directly linked to the biological plausibility and support for the findings from external sources.

If analyses derived from unplanned post hoc assessment of data are presented, they should preferably be reported using descriptive statistics with clear identification that they have not been generated within the inferential framework of the trial (p values must be clearly marked as nominal, i.e., as unplanned analyses and not controlled for multiplicity).

HTDs have to provide all information available on the characteristics of the subgroups, substantiate any claims regarding balance in terms of randomisation, provide evidence that no interactions with other prognostic or predictive factors might be the underlying cause of any differences observed and provide a strong biological rationale if a specific subgroup performs better or worse than the overall trial population.

Requirements for JCA reporting

- Planned post hoc analyses, such as a procedure for control of multiplicity issues, are to be reported according to the requirements described in the corresponding sections of the document (e.g., Section 3 if these analyses deal with controlling for multiplicity).
- Unplanned post hoc analyses, such as those requested by a HTA as a consequence of the PICO process, should be clearly flagged as unplanned.

10 POST HOC ANALYSES IN EVIDENCE SYNTHESIS STUDIES

10.1 Purposes, definitions and general methodological considerations

Evidence generation should follow a planned protocol to reduce the likelihood of drawing biased conclusions. Full prespecification is difficult and often not possible for systematic reviews because knowledge is already available for the underlying studies. Therefore, if an important aspect was not addressed in the planning stage (PICO scoping) but proves to be of importance for the assessment, additional post hoc analyses might be required.

10.2 Requirements for appropriate reporting of methods and results in a JCA

The JCA should report post hoc analyses but highlight them to distinguish them from other planned analyses.

Requirements for JCA reporting

- A report of the protocol-defined analyses and their relevance to the PICO.
- The report should clearly distinguish between planned analyses and unplanned post hoc analyses.

11 RELATED EUNETHTA DOCUMENTS

- **EUnetHTA 21 Practical Guideline D4.3.1: Direct and indirect comparisons**

A practical guideline for assessors and co-assessors that describes possible approaches and specific instructions for reporting and assessing the evidence from evidence synthesis studies (pairwise meta-analyses and indirect comparisons).

- **EUnetHTA 21 Methodological Guideline D4.3.2: Direct and indirect comparisons**

A methodological guideline for assessors and co-assessors that provides an understanding of the basic methodological principles behind conducting evidence synthesis studies (pairwise meta-analyses and indirect comparisons).

- **EUnetHTA 21 Practical Guideline D4.2.1: Scoping process**

A practical guideline for assessors and co-assessors that describes the methods and principal steps for the scoping process.

- **EUnetHTA 21 Practical Guideline D4.4.1: Endpoints**

A practical guideline for assessors and co-assessors that describes how to deal with several issues encountered around the assessment of endpoints, and guides member states on defining relevant endpoints during the scoping process.

- **EUnetHTA 21 Practical Guideline D4.6.1: Validity of clinical studies**

A practical guideline for assessors and co-assessors that defines the main aspects of the validity of individual clinical studies, defines and classifies the different types of clinical studies that can be conducted, and describes how to report and assess the validity of individual clinical studies whether they are RCTs or not.

12 REFERENCES

1. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Harmonised Tripartite Guideline. Statistical Principles for Clinical Trials E9*. Geneva: ICH; 1998.

2. Neyman J. “Inductive behavior” as a basic concept of philosophy of science. *Rev Int Stat Inst* 1957;25(1/3):7.

3. Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol* 2001;54(4):343–9.

4. Bauer P. Multiple testing in clinical trials. *Stat Med* 1991;10(6):871–90.

5. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997;18(6):530–45.

6. Zhang J, Quan H, Ng J, et al. Some statistical methods for multiple endpoints in clinical trials. *Control Clin Trials* 1997;18(3):204–21.

7. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc Ser Gen* 1969;132(2):235–44.

8. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. New York, NY: Chapman and Hall/CRC; 2016.

9. Dmitrienko A, Tamhane AC, Bretz F, editors. *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton, FL: Chapman & Hall/CRC; 2010.

10. Lesaffre E, Lawson A. *Bayesian Biostatistics*. Chichester: Wiley; 2012.

11. Ryan EG, Brock K, Gates S, et al. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? *BMC Med Res Methodol* 2020;20(1):150.

12. Kapur J, Elm J, Chamberlain JM, et al. Randomized trial of three anticonvulsant medications for status epilepticus. *N Engl J Med* 2019;381(22):2103–13.

13. Guo M, Heitjan DF. Multiplicity-calibrated Bayesian hypothesis tests. *Biostatistics* 2010;11(3):473–83.

14. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL, et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol. 2008;61(9):857–65.

15. Li T, Higgins JPT, Deeks JJ. Collecting data. In: Higgins JPT, Thomas J, Chandler J, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions version 6.3*. London: Cochrane Collaboration; 2022. Chapter 5.

16. Efthimiou O, White IR. The dark side of the force: multiplicity issues in network meta-analysis and how to address them. *Res Synth Methods* 2020;11(1):105–22.

17. Christensen R, Bours MJL, Nielsen SM. Effect modifiers and statistical tests for interaction in randomized trials. *J Clin Epidemiol* 2021;134:174–7.

18. Tanniou J, van der Tweel I, Teerenstra S, et al. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol* 2016;16:20.

19. Dmitrienko A, Muysers C, Fritsch A, et al. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat* 2016;26(1):71–98.

20. Alosh M, Huque MF, Bretz F, et al. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med* 2017;36(8):1334–60.

21. Cui L, Hung HM, Wang SJ, et al. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat* 2002;12(3):347–58.

22. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.

23. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989;18(1):269–74.

24. Simmons LA. Self-perceived burden in cancer patients: validation of the Self-perceived Burden Scale. *Cancer Nurs* 2007;30(5):405–11.

25. Berlin JA, Santanna J, Schmid CH, et al, Anti-Lymphocyte Antibody Induction Therapy Study G. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;21(3):371–87.

26. International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. *Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials*. E9(R1). Geneva: ICH; 2019.

27. European Medicines Agency. *Guideline on Missing Data in Confirmatory Clinical Trials*. London: EMA; 2010.

28. European Medicines Agency. *Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials*. London: EMA; 2019.