HTA CG | MEMBER STATE COORDINATION GROUP ON HEALTH TECHNOLOGY ASSESSMENT

Guidance on reporting requirements for multiplicity issues and subgroup, sensitivity and post hoc analyses in joint clinical assessments

Adopted on 10 June 2024 by the HTA CG pursuant to Article 3(7), point (d), of Regulation (EU) 2021/2282 on Health Technology Assessment

Contents

List of abbreviations

Abbreviation	Definition
CG	Coordination Group
CSR	Clinical study report
CWER	Comparisonwise error rate
EC	European Commission
EU	European Union
EUnetHTA	European Network of Health Technology Assessment
FWER	Familywise error rate
HaDEA	European Health and Digital Executive Agency
HTA	Health technology assessment
HTD	Health technology developer
ICE	Intercurrent event
ICEMAN	Instrument to assess the Credibility of Effect Modification Analyses
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
JCA	Joint clinical assessment
MPG	Methodological and Procedural Guidance
MS	Member State
NMA	Network meta-analysis
PICO	Population, intervention, comparator, outcome
SAP	Statistical analysis plan

1 Scope and objective

This guidance describes how to deal in practice with multiplicity issues and complementary analyses (like specific subgroup analyses, specific post hoc analyses and sensitivity analyses) in joint clinical assessment (JCA) reports. Guidance is provided for assessors, co-assessors and other members of the assessment team (henceforth referred to collectively as *assessors*) on how to report these analyses. Chapter 3 of this guidance contains a list of reporting requirements pertaining to multiple hypothesis testing or complementary analyses which should be included in a JCA.

Member States (MSs) may take the approach of assessing evidence from individual studies within the framework of the original studies' statistical analysis plan (SAP) and/or of assessing a statistical summary (or evidence synthesis) of one or several studies within the framework of a systematic review. These approaches impact the way in which different MSs consider specific methodological issues such as multiplicity and subgroup, sensitivity and post hoc analyses. It is not the intent of this guidance to endorse a particular approach but to enable MSs to draw their own conclusions at the national level.

Of note, the analysis and reporting recommendations for assessors are made with the implicit assumptions that appropriate analyses and information is provided by the health technology developer (HTD). As such, this guidance also has practical implications for the submission dossier which should be taken into account in the preparation of this document.

2 Definitions

Alpha level: the significance threshold set by researchers for statistical significance tests to reject the null hypothesis in favour of the alternative hypothesis.

Comparisonwise error rate (CWER): the probability of rejecting one null hypothesis, which in fact is true.

Effect modifier: a variable that modifies a treatment effect, that is, a variable that alters the relative effectiveness between two treatments.

Evidence: empirical information from scientific studies regarding a specific question.

Evidence synthesis: a broad term for combining the results of multiple studies investigating the same topic.

Effectiveness: how well a treatment works; includes efficacy and safety.

Familywise error rate (FWER) in the weak sense: the probability of rejecting at least one of k null hypotheses when in fact all null hypotheses are true; also called global level.

Familywise error rate (FWER) in the strong sense: the probability of rejecting at least one of k null hypotheses when in fact this null hypothesis is true, irrespective of which and how many of the other null hypotheses are true; also called multiple level.

Intercurrent event (ICE): an event occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest.

Interim analysis: any analysis of study data with respect to effectiveness at a time point before formal completion of a trial (henceforth referred to collectively as the *final* analysis).

Multiplicity: the performance of multiple statistical analyses for a research question.

***p*-value**: the conditional probability of obtaining a result equal to or more extreme than what was actually observed under the condition that the null hypothesis is true.

Pairwise meta-analysis: the synthesis of two or more head-to-head studies with a common intervention and comparator, to produce a pooled estimate of the relative treatment effect.

Planned or **prespecified**: a statistical analysis as planned according to a study protocol or SAP.

Post hoc analysis: a statistical analysis that was not planned according to a study protocol or SAP.

Power: the conditional probability that a significance test detects an effect under the condition that the alternative hypothesis is true.

Predictive factor: a patient characteristic that alters the relative effectiveness between two treatments (is the same as effect modifier).

Prognostic factor: a patient characteristic that affects the outcome of interest irrespective of which treatment is received.

Sensitivity analysis: a set of analyses estimating the same effect but with different methodology to assess the impact of different decisions compared to the primary assumptions of the analysis.

Subgroup: a subset of the study/patient population defined by one or more specific patient characteristics (e.g., age, sex) measured at baseline. Subgroup analyses in the context of a JCA are performed within a given PICO.

Subgroup analysis: the estimation and comparison of treatment effects in the (disjoint) subgroups of a potential effect modifier.

Subpopulation: a subset of the study/patient population covered by the therapeutic indication. Subpopulations in the context of a JCA result in separate PICOs for each subpopulation.

Treatment: a health technology to be assessed. Health technologies encompass medicinal products, medical devices, *in vitro* diagnostic medical devices and medical procedures.

3 Reporting requirements for multiple hypothesis testing and complementary analyses in a JCA

3.1 General considerations

In all analyses described in the JCA it should be clear whether this analysis represents the primary analysis regarding the PICO (population, intervention, comparator, outcome) or a complementary analysis (like specific subgroup analyses, specific post hoc analyses and sensitivity analyses). Additionally, it should be clear which statistical method was used, and if performed, which adjustments for multiple testing were used.

In any clinical trial, the primary estimand should be defined according to the principles outlined in International Conference for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) E9 and its addendum (E9(R1)) [9,14]. 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” [9]. The estimand is defined by the same attributes of the PICO, but supplemented with a specification of the effect measure and the handling of intercurrent events (ICEs). The statistical analyses should be aligned to the estimand (not vice versa) and sensitivity analyses should be prespecified in the study protocol and SAP to “explore the robustness of inference from the main estimators to deviations from its underlying modelling assumptions and limitations of the data” [9].

As regulatory assessments are likely to be based, at least partly, on the same data that are used for JCAs, a certain amount of overlap can be expected. However, the regulatory agencies often do not address the same research questions and therefore have a different focus in their assessment reports. Where detailed analysis of aspects such as multiplicity has been undertaken by the regulatory agency in their assessment it may be sufficient to report from this assessment. However, where additional analyses are required to address the JCA PICO question(s), for example outcomes or subgroups that may not have been assessed fully at regulatory level, there may be a need to examine the methodological aspects of these analyses in more detail. It is advised that assessors should be concise whenever possible to reduce overlap, while keeping in mind that a JCA report needs to be readable without the regulatory assessment report as adjunct. Further guidance on how regulatory assessments can be used in the context of JCA will be developed by the MPG subgroup.

While some recommendations in this guidance are more important for randomised controlled trials, most of the general principles also apply to other (non-randomised) study designs. Below is a list of general reporting requirements which apply to multiple testing and complementary analysis. Other reporting requirements only apply to specific analyses; they are listed below as “additional requirements”.

3.2 Reporting requirements

3.2.1 General reporting requirements

- A clear description must be provided of the applied estimands and null and alternative hypotheses that were tested.
- The α level of each significance test must be provided, as well as information on whether the comparisonwise (CWER) or familywise error rate (FWER) in the weak (global level) or in the strong sense (multiple level) was applied.
- For each estimand a description of the statistical method used must be provided and, if performed, the multiplicity procedure.
- For each analysis it must be clearly stated whether the analysis was prespecified (i.e., in the study protocol or SAP) or not. If an unplanned analysis was conducted, a clear justification must be provided and it should be reported by whom they were deemed necessary (e.g., sponsor, regulatory body, HTA body).
- The results of all relevant analyses with appropriate statistics must be provided, in most cases with effect estimates, confidence intervals and corresponding *p*-values (with corresponding multiplicity adjustments, if appropriate).
- In the case of analyses of several time points, the assessors can select relevant time point(s) from the submission dossier (e.g., the last time point or an adequate summary measure over time), but then a justification for their choice must be provided.
- If for a requested outcome for which analyses were planned no results are available, the assessors must provide the reason why no results can be presented.
- The assessors should distinguish complementary analyses from the primary estimand(s) and its analyses.

3.2.2 Additional requirements for multiple data cuts

- In general, in the case of multiple data cuts, the assessors must report the results of the last prespecified data cut. For this data cut, results for all outcomes must be reported. If the results of the last pre-specified data cut are of insufficient quality (e.g., due to a high proportion of missing data or treatment switching) the assessors may report the data of an earlier data cut. An earlier data cut may also be warranted if the length of follow-up has to be standardised across studies in evidence synthesis. For such reasons, assessors can deviate from the requirement to report the results of the last prespecified data cut, but then a justification must be provided.

- In addition, for multiplicity controlled outcomes in the original studies, the results of the first data cut for which the null hypothesis was rejected must be provided in case it does not correspond to the latest prespecified data cut.
- In addition, for the outcome ‘overall survival', the result of the last data cut must be provided in case it does not correspond to the last prespecified data cut.

3.2.3 Additional requirements for interim analyses

- The cutoff date for each interim analysis must be provided.
- If an interim analysis had an impact on the conduct on the clinical study (e.g., modification of study protocol, early stopping, data release), the impact must be reported, as well as the reason(s) why a change was deemed necessary and by whom (e.g., data and safety monitoring board, sponsor). In addition, the stopping rule must be provided if the study was stopped early.
- If the final (prespecified) analysis for an outcome is available, only a reference to the tables in the clinical study report (CSR) with the results of all interim analyses must be provided.
- If the final (prespecified) analysis for an outcome is not (yet) available, the results of the last prespecified interim analysis must be provided, as well as a reference to the tables in the CSR with the results of all interim analyses. If needed, assessors can deviate from this requirement but then a justification must be provided.

3.2.4 Additional requirements for subgroup analyses

- A description of the subgroup analyses performed must be provided, as well as references to the tables in the submission dossier with the results of all subgroup analyses.
- If the interaction test is significant (according to the SAP or, if unknown, a two-sided *p*-value < 0.05), the results of subgroup analyses (including the *p*-values of the interaction tests) must be reported. In addition, if a MS requested a specific subgroup analysis, the results must be reported. If needed, assessors can deviate from these requirements but then a justification must be provided.

3.2.5 Additional requirements for sensitivity analyses

- A clear definition of the purpose and underlying assumption of each sensitivity analysis must be provided.
- A description of the sensitivity analyses performed must be provided, as well as references to the tables in the submission dossier with the results of all sensitivity analyses.

- If in an evidence synthesis an operationalisation for an outcome (i.e., measurement of the outcome or effect measure) was chosen which deviates from the primary one in the original study, a justification must be provided.
- The results of relevant sensitivity analyses with regards to the primary analysis must be reported, namely those with
  - another directionality (e.g., a relative risk of 1.3 versus 0.8 or a mean difference of +1 vs −1); OR
  - different results regarding statistical significance (e.g., a *p*-value ≤ 0.05 versus > 0.05); OR
  - large differences in effect estimates.

  If needed, assessors can deviate from this requirement but then a justification must be provided.
- If the results of the primary analysis differ from the (reported) sensitivity analyses, explanations for these differences must be provided, if possible.

4 Multiple statistical hypothesis testing

4.1 Purpose and general methodological considerations

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 in general 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 outcome of interest between groups 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).

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)) [16]. Statistical hypothesis testing controls the risk of wrongly claiming the existence of treatment effectiveness at an acceptable level and it relies on calculating the *p*-value. The *p*-value of a test is compared to a prespecified significance level (i.e., the α level). If the *p*-value is less than the α level, the null hypothesis is rejected. In biomedical research, the consensus is usually to set the α level of a (two-sided) single test at 0.05 (5%). The probability α of a type 1 error for one significant test is the CWER [4].

Performing multiple statistical significance tests increases the risk of at least one false-positive test. Thus, if k significance tests are performed, the probability of rejecting at least one of the k null hypotheses when all null hypotheses are actually true is called the familywise error rate (FWER, considering the family of k tests as one experiment under the complete null hypothesis) in the weak sense. When considering k independent tests, FWER is equal to $1–(1–\alpha)^k$ [4]. If multiple tests are dependent (e.g., they assess correlated outcomes 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 [4].

In many cases, however, it is not realistic that all considered null hypotheses are true simultaneously. The FWER in the strong sense (also called multiple level) is the probability of erroneously rejecting at least one true null hypothesis, irrespective of which and how many of the other individual null hypotheses are true [4]. A multiple test procedure that controls FWER in the strong sense also controls the FWER in the weak sense (but not vice versa) [2]. Thus, for the rest of the document, we use the term FWER to mean “FWER in the strong sense”. Numerous procedures, such as the Bonferroni method, multiple-step procedures (e.g., the Holm procedure) and the principle of closed testing or resampling-based procedures, were developed to control the FWER at an acceptable level [2,4,12]. The procedures for dealing with multiple statistical hypothesis testing based on *p*-values are generally applicable, while others

require access to individual patient data. The latter are therefore not applicable in evidence syntheses based upon aggregated data [3].

When conducting evidence synthesis studies such as pairwise meta-analysis or analysis of more complex evidence networks via network meta-analysis (NMA), multiplicity issues can arise in multiple ways. These issues can be similar to those encountered when dealing with original clinical studies or they can be specific to the design of an evidence synthesis study [3]. As users of evidence synthesis analyses have heterogeneous interests in the consequence of a medical condition, evidence synthesis analyses are frequently performed with the inclusion of all outcomes that are likely to be of importance. Because decisions on which outcomes to include are also frequently encountered during the data extraction process, an unambiguous definition of what constitutes a family of tests in the context of evidence synthesis is difficult to achieve. Furthermore, the possibilities and necessities to deal with multiplicity in evidence synthesis are limited because the data are already observed. This makes it not possible to plan for multiplicity adjustments in a strong confirmatory sense. The only practical way to deal with multiplicity issues in evidence synthesis is to take multiplicity into account when interpreting the results [3].

4.2 Multiple statistical hypothesis testing in original clinical studies

Multiplicity can arise in original studies because of the various ways available for analysing an outcome and the operationalisation of an outcome. Outcomes can also be assessed at different time points (e.g., clinical remission at 6 months and at 12 months) or in different populations (e.g., full population or a subpopulation). These analyses lead to multiple testing issues which can be considered similar from a methodological perspective. In this guidance three main situations for which multiple statistical tests arise are discussed in more detail, namely multiple outcomes, interim analyses and more than two treatment groups.

4.2.1 Multiple outcomes

Clinical studies are in general designed to estimate the effectiveness of a treatment for more than one outcome, because most diseases have more than one consequence [17,21]. To avoid data dredging and control the type 1 error rate, prospective specification of all outcomes and all data analyses that are performed to test hypotheses about the prespecified outcomes (including the choice of multiplicity procedures) either before initiation of a clinical study or at least before database lock (i.e., planned analyses) should be performed.

4.2.2 Interim analyses

Interim analyses are performed before the completion of a trial and can lead to several analyses for a given outcome before formal completion of the study. The results of these analyses may be used 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. The ethical and scientific benefits of interim analyses may be opposed by methodological disadvantages, including a potential increase in type 1 errors if not appropriately managed [1].

4.2.3 Multiple treatment groups

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) [6]. These analyses can also lead to multiplicity analyses. To reduce the number of analyses, data only need to be reported for the intervention and comparator(s) of interest (see the guidance on the scoping process).

4.3 Multiple statistical hypothesis testing in evidence syntheses

Evidence synthesis usually includes multiple outcomes, which can be analysed and operationalised in various ways. It is recommended that the HTD prespecifies, before data extraction, how an outcome will be synthesized together to avoid the appearance of data dredging [3,15].

A solution to limit the use of different effect measures or operationalisations of an outcome is to prespecify and justify in the SAP the primary operationalisation before data extraction [3]. Different methods for analysing the same outcome may then be used to assess the robustness of a result, for example, via sensitivity analyses (see Chapter 6).

In general, outcomes are measured at different time points. It is possible to perform multiple evidence syntheses for multiple time points available. A solution to limit the issue of analysing multiple time points can be to choose a single time point for the analysis. This is, however, 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 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 [3]. The evidence synthesis can then be performed by using the estimates of the summary effect measure (e.g., the hazard ratio). 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. More information regarding these analyses is provided in the guidances for quantitative evidence synthesis: direct and indirect comparisons.

5 Subgroup analyses

5.1 Purpose 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 pretreatment or concomitant treatment. This can result in different effect sizes. The aim of subgroup analyses is to identify either consistency or large differences in the magnitude of treatment effect among different patient groups. Demonstration of different effects between different subgroups should be conducted using an appropriate interaction test (e.g., adequate regression or analysis-of-variance model) for one factor at a time. Different homogeneity and interaction tests have been discussed in the literature [7,11,20]. In this guidance, the term “interaction test” refers to all of these tests. The Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) criteria [18] may be used to interpret and assess the results of submitted subgroup analyses.

It should be kept in mind that due to potential small sample sizes for subgroups, the power of interaction tests for detecting heterogeneity between subgroups can be low. It can also not be ruled out that any differences detected between subgroups are caused by imbalances in patient characteristics. In very small sample sizes, prognostic and predictive factors may be unbalanced within the respective subgroups if randomisation is not stratified according to the subgroup characteristic analysed [10,19]. In such cases, the unbalanced prognostic factor can affect both the treatment received and the outcome. Thus, the effect estimates within the subgroups may be biased due to imbalances and this bias can lead to different results in the different subgroups.

5.2 Subgroup analyses in original clinical studies

Prespecification of subgroups in a study protocol or SAP (of the original study or evidence synthesis), as well as the definition of the purpose and underlying assumption of each sensitivity analysis, is needed as it can lend credibility to positive or negative subgroup findings. However, if MSs request to explore potential effect modifiers (e.g., age and sex) an HTD might have to perform unplanned subgroup analyses (see for more information the guidance on the scoping process). Therefore, it is important to state in an JCA who requested an unplanned analysis (e.g., HTA body, sponsor, regulatory body).

HTDs have to provide all relevant information available on the characteristics of the subgroups in the original studies, as well as the definition of the purpose and underlying assumption of each analysis. Furthermore, they need to substantiate statements 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.

5.3 Subgroup analyses in evidence syntheses

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 between subgroups or meta-regression can be considered. 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 the guidances for quantitative evidence synthesis on direct and indirect comparisons.

When 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. In a meta-regression, the statistical association between the effect sizes in original studies and the study characteristics is investigated, so that study characteristics can possibly be identified that explain 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 original studies are subject to the same limitations as the results from ecological studies in epidemiology [13]. 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 [5]. In the case of an evidence synthesis with individual patient data, an adequate regression or analysis-of-variance model with a corresponding interaction term can be used.

6 Sensitivity analyses

6.1 Purpose and general methodological considerations

Sensitivity analyses are an integral part of the reporting of statistical analyses and are essential in investigating the robustness of the estimated effects to variations in the assumptions and their impact.

Sensitivity analyses are important for example to explore the impact of missing data because they can lead to serious bias. Handling of missing data is considered a statistical problem that needs to be addressed via appropriate statistical analyses with the aim to explore the impact of the level of missing data on the basis of certain assumptions [8]. Guidelines on handling of missing data are available [8,9] and describe appropriate sensitivity analysis strategies. There is no rule for the amount of missing data that is considered acceptable. Thus, the acceptability of missing data is subject to MSs differences in interpretation of their relevance within their respective decision-making process. Therefore, reports should highlight the uncertainty with respect to the amount, as well as the handling of missing data.

Within the context of a JCA, it is not expected that sensitivity analyses have to be conducted for every PICO question, especially for every outcome. It is the responsibility of the HTD to provide as many sensitivity analyses as appropriate, according to good clinical and statistical practices, along with a clear definition of their purpose, underlying assumption(s) and attribute(s) they address.

6.2 Sensitivity analyses in original clinical studies

In sensitivity analyses focus should also be given to ICEs (i.e., events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest). ICEs 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 should be used to explore the impact that changes to the assumptions for any or all of these elements might have on the outcomes of interest [9].

6.3 Sensitivity analyses in evidence syntheses

There are many potential sources for missing data in evidence synthesis, e.g., missing complete studies, missing outcomes in some studies, missing summary data for some outcomes in some study, or missing study data for some study participants. The amount of and reasoning for missing data should be carefully described and the potential impact of the missing data on the findings of the evidence synthesis should be addressed in sensitivity analyses.

7 Post hoc analyses

7.1 Purpose 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 (planned and unplanned) analyses that are performed because of the results of a previous analysis. However, this chapter will only address unplanned post hoc analyses, as these can be considered problematic in terms of deviation from an adequate hypothetico-deductive approach. 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. Therefore, unplanned post hoc analyses should be clearly identified as such to distinguish them from the planned analyses of the original clinical studies.

In the context of JCA, the assessment scope, expressed by the PICO question(s), defines the relevant comparator(s) for relative effectiveness assessment. Therefore, it might be necessary to obtain data for a subpopulation that, for example, reflects the PICO more closely than the study population of the original clinical study. This requires a post hoc analysis. In this case, for some MSs, the unplanned post hoc analysis has more importance for the PICO than the planned analysis in the original clinical study and represents the primary analysis in the JCA (for that specific PICO).

7.2 Post hoc analyses in original clinical studies

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 original inferential framework of the trial. This means that *p*-values must be clearly marked as nominal, i.e., as results that come from unplanned analyses which have not been controlled for multiplicity.

7.3 Post hoc analyses in evidence syntheses

Evidence syntheses should follow a planned SAP before data extraction to reduce the likelihood of drawing biased conclusions. But 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.

8 References

1. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc Ser A 1969; 132: 235-244. https://dx.doi.org/10.2307/2343787.
2. Bauer P. Multiple testing in clinical trials. *Stat Med* 1991; 10(6): 871-890. https://dx.doi.org/10.1016/S0895-4356(00)00314-0.
3. Bender R, Bunce C, Clarke MJ et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008; 61(9): 857-865. https://dx.doi.org/10.1016/j.jclinepi.2008.03.004.
4. Bender R, Lange S. Adjusting for multiple testing – when and how ? *J Clin Epidemiol* 2001; 54(4): 343-349. https://dx.doi.org/10.1016/S0895-4356(00)00314-0.
5. Berlin JA, Santanna J, Schmid CH et al. 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): 371-387. https://dx.doi.org/10.1002/sim.1023.
6. Bretz F, Hothorn T, Westfall P. *Multiple Comparisons Using R*. New York, NY: Chapman and Hall/CRC; 2016.
7. Christensen R, Bours MJL, Nielsen SM. Effect modifiers and statistical tests for interaction in randomized trials. *J Clin Epidemiol* 2021; 134: 174-177. https://dx.doi.org/10.1016/j.jclinepi.2021.03.009.
8. Committee for Medicinal Products for Human Use (CHMP). *Guideline on missing data in confirmatory clinical trials*. EMA; 2010.
9. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.
10. Committee for Medicinal Products for Human Use (CHMP). *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (Step 5) (EMA/CHMP/ICH/436221/2017)*. EMA; 2020. https://www.ema.europa.eu/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf.
11. Cui L, Hung HM, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. J Biopharm Stat 2002; 12(3): 347-358. https://dx.doi.org/10.1081/bip-120014565.
12. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat* 2016; 26(1): 71-98. https://dx.doi.org/10.1080/10543406.2015.1092033.

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

14. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989; 18(1): 269-274. https://dx.doi.org/10.1093/ije/18.1.269.

15. ICH Expert Working Group. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials. *Stat Med* 1999; 18(15): 1905-1942. https://dx.doi.org/10.1002/(SICI)1097-0258(19990815)18:15<1903::AID-SIM188>3.0.CO;2-F.

16. Li T, Higgins JPT, Deeks JJ. Chapter 5: Collecting data. In: Higgins JPT, Thomas J, Chandler J et al. (Ed). *Cochrane Handbook for Systematic Reviews of Interventions, 2nd Edition*. Hoboken, NJ: Wiley; 2019. S. 109-141.

17. Neyman J. “Inductive behavior” as a basic concept of philosophy of science. *Rev Int Stat Inst* 1957; 25(1/3): 7. https://dx.doi.org/10.2307/1401671.

18. Pocock SJ. Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Control Clin Trials* 1997; 18(6): 530-545. https://dx.doi.org/10.1016/s0197-2456(97)00008-1.

19. Schandelmaier S, Briel M, Varadhan R et al. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ* 2020; 192(32): E901-E906. https://dx.doi.org/10.1503/cmaj.200077.

20. 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: 850-854. https://dx.doi.org/10.1136/bmj.c117.

21. Tanniou J, van der Tweel I, Teerenstra S, Roes KC. Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes. *BMC Med Res Methodol* 2016; 16: 20. https://dx.doi.org/10.1186/s12874-016-0122-6.

22. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. *Control Clin Trials* 1997; 18(3): 204-221. https://dx.doi.org/10.1016/S0197-2456(96)00129-8.

23.