EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.5* – Applicability of evidence**

General answer:

EUnetHTA21 wishes to thank the many organizations and individuals who have responded to the public consultation of this practical guideline. We have taken all comments into consideration and provided individual answers to them. Given that many similar comments were made, we will try to address at least some of the main themes in the clarifying text below.

1) We received several comments suggesting that this guideline requires mandatory analyses (for example mandatory adjustment for multiplicity). As introduced in the guideline (114-120), this guideline only requests the reporting of some methodological elements, but not the mandatory application of them. As also stated in the introduction (101-113), MS appraise certain aspects differently. This guideline is not intended to endorse one unique approach and state that it is the 'mandatory' one. To allow MS drawing their own conclusions on the clinical added value of health technology, certain methodological aspects must be reported adequately in the JCA report.

2) We received several comments asking for a harmonized European methodology (i.e., methodology has to be the same for every MS). As clearly stated in the HTAR, the JCA is based on the chosen parameters during the assessment scope (Article 9). As also clearly stated in the Article 8 (6): 'the assessment shall be inclusive and reflect Member States' needs in terms of parameters and of the information, data, analysis and other evidence to be submitted by the health technology developer'. Considering the above, assessment scope is based on MS needs, and different MS needs could exist. There is therefore no limitation or requirement for a harmonized methodology.

Name organisation & abbreviation	**Country**
DVSV	Austria
European Union of General Practitioners/Family Physicians (UEMO)	Belgium
European Confederation of Pharmaceutical Entrepreneurs (EUCOPE)	Belgium
European Federation of Pharmaceutical Industries and Associations (EFPIA)	Belgium
Alliance for Regenerative Medicine (ARM)	Belgium
The European Society for Paediartic Oncology (SIOPE)	Belgium
Takeda Pharmaceuticals International AG	Brussels, Switzerland, local operating companies across the European Union
European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) HTA SIG	Europe
MedTech Europe (MTE)	Europe - Belgium
Lymphoma Coalition - Lymphoma Coalition Europe (LCE)	France
EHA	France

Please add extra rows as needed.

EURORDIS	France
Ecker + Ecker GmbH (E+E)	Germany
SKC Beratungsgesellschaft mbH (SKC)	Germany
Verband Forschender Arzneimittelhersteller (vfa) e.V	Germany
GKV-Spitzenverband (GKV-SV)	Germany
Advanced Medical Services GmbH (AMS)	Germany
Bayer AG & Bayer Vital GmbH	Germany
German Medicines Manufacturer´s Association (BAH)	Germany
Lumanity	Lumanity is a global company with several European entities, including in Ireland and the Netherlands.

Outside the EU

Name organisation & abbreviation	**Country**
Institut national d'excellence en santé et en services sociaux (INESSS)	Canada
AstraZeneca (AZ)	Global (UK based)
F. Hoffmann-La Roche Ltd (Roche)	Switzerland
Medtronic	Switzerland
GSK	UK
PHMR	UK
ISPOR	US Based

Please add extra rows as needed.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.5* – Applicability of evidence**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
DVSV	general		Many thanks to the authors for the excellent work.	Thank you!
Advanced Medical Services GmbH	general		With regard to the Submission Dossier Template we would like to propose a "design" with text boxes that contain specific descriptions and definitions of what is required and how it has to be presented (for an example see the submission dossier template for the German Benefit Assessment that has text boxes shadowed in grey). Thus, the health technology developer (HTD) will have a comprehensive instruction of what is expected in the Submission Dossier corresponding to the listings in the boxes of D4.5: "Requirements for JCA reporting".	The submission dossier template and JCA template are the purpose of other EUnetHTA 21 deliverables.
Daniel Widmer UEMO	general		Very good and clear technical paper with requirements for JCA reporting.	Thank you.
Matias Olsen, EUCOPE	general		It is difficult to comment on some of the methods without better understanding the communication between the joint HTAs and HTDs, and when expectations will be provided and discussed, e.g. a scoping meeting.	Interactions between HTAb and HTD are addressed in another guideline (D7.1), and are therefore out of scope.
Matias Olsen, EUCOPE	general		The CER is mentioned in all the requirements for JA reporting in Sections 3.2.1 – 3.2.3, however it is only applicable to section 3.2.3 for multiple comparisons.	We do not agree. The CER is the alpha level of a specific test.
Matias Olsen, EUCOPE	general		There is no mention of multiplicity control while performing confirmatory subgroup analysis where the intended labelling is for the whole population and/or sub-population represented by the subgroup. Here efficacy in at least one population provides foundation for registration. Closed testing (ref: Marcus et al., 1976, Biometrika) and related methods like fixed sequence procedures (ref: Bretz et al., 2009, Statistics in Medicine) can be used to control for type-I error inflation in this case. Multiplicity also needs to be controlled in adaptive enrichment trials where the population of interest may be enriched at the interim analysis based on a predictive or	The guideline is intended to allow factual reporting of elements within a JCA report from the perspective of the assessment scope. In the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			prognostic biomarker. Here multiplicity can be controlled with the use of closed testing and stage-wise weighing of p-values or test-statistics.	end, we think the requirements for reporting are covering various situations. The guideline is not intended to be a statistical textbook describing statistical possibilities for controlling for multiple hypothesis testing.
Mihai Rotaru - EFPIA	General		Duplication The "Requirements for JCA reporting" on the hypotheses tested in the trial and the adjustment for multiplicity used in the trial listed in chapters 3.2, 4.2, 5.2., 6.2, 7.2, 8.2., 9.2. and 10.2 are standard requirements for Good Scientific Practice and are therefore reported in the Clinical Study Reports (CSR) of the corresponding studies. As the CSR needs to be supplied as part of the submission, for the purpose of JCA reporting there should be references to the corresponding sections of the CSR to avoid "overloading" the JCA dossier and report with duplicate information.	The fact that some elements are already in the CSR do not justify to not include them in the JCA.
Mihai Rotaru - EFPIA	General		Hypothesis testing and adjustment for multiplicity Hypothesis testing and the Neyman-Pearson approach, as well as corresponding multiplicity adjustments is a necessary concept for regulatory decision making, as it gives the binary decision if a trial meets its primary objective with an overall prespecified error probability alpha, normally set to 5%. With a positive trial, the proof that the trial met its objectives is available, as the primary null hypothesis or	What is proposed within the guideline is a factual reporting of the elements member states

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			hypotheses was/were rejected, and the pre-specified alpha was spent for that decision point. The EU JCA should not be repeating analyses already undertaken and confirmed by the EMA decision, for which the CSR and EPAR are available for Member States. Prespecified hypotheses and multiplicity adjustment procedures are well described in the CSR. For purposes of EU JCA, defining a hierarchy of endpoints is, implicitly, a value judgement, as significance levels and acceptability of uncertainty are a national decision of each Member State. As stated in the Regulation1, 'It is necessary therefore that Union action is limited to those aspects of HTA that relate to the joint clinical assessment of a health technology and to ensure in particular that there are no value judgements in joint clinical assessments in order to respect the responsibilities of Member States pursuant to Article 168(7) TFEU.' Furthermore, it could be reasonably argued that presenting a hierarchy of endpoints in the JCA report is a form of ranking. The Regulation explicitly excludes such ranking: "…The joint clinical assessment report should be factual and should not contain any value judgement, ranking of health outcomes, conclusions on the overall benefit or clinical added value of the assessed health technology, any position on the target population in which the health technology should be used, or any position on the place the health technology should have in the therapeutic, diagnostic or preventive strategy." (Recital 28). Following on from this, no mandatory adjustment for multiplicity should be applied. As different countries may apply different significance levels for their decision making on the endpoints of the trial, conclusions on the significance of a given result should be avoided (for example, not meeting a prespecified significance level in the context of the clinical trial should not be flagged as a factor that affects certainty of the evidence.)	could need to perform their appraisal at the national level. Thus, reporting how outcomes were analysed in the original clinical studies performed by HTD does not constitute de facto a ranking of outcomes.
Mihai Rotaru - EFPIA	General		Risks of multiple analyses To avoid wrong conclusions and data dredging due to multiple analyses that can lead to deviating results, the scope of the JCA should be based as much as possible on the	HTAR allows MS to request what they need without

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			prespecified analyses in the trial and should be limited to analyses that are really necessary for the JCA. Complementary, sensitivity and subgroup analyses should not be a regular requirement in the JCA. The general approach should be to conduct meaningful sensitivity analyses for situations where the investigation of the robustness of the results is crucial. The decision to conduct additional analyses should be explained on an individual basis, and a strong scientific, clinical, biologic or regulatory rationale needs to be given by the requestor. Early involvement of the HTD in the scoping process and alignment on analyses needed for the JCA is therefore of utmost importance. There should be a constructive dialogue implemented between the assessors and the HTD to ensure a methodological exchange before prespecifying and conducting the corresponding analyses for a specific dossier.	limitation, rationale to give, nor binding to the analyses already performed in clinical studies.
Mihai Rotaru - EFPIA	General		Sub populations need to be interpreted in conjunction with the overall population Subpopulations of interest may be specified during the assessment scope as separate research questions, and a strong clinical, biologic or regulatory rationale needs to be given by the Member States to justify these. A priori, the effect estimate from the overall study population is the best estimate within each subpopulation. If a subpopulation needs to be analyzed for the scope of the JCA, this should be reflected as a "Sub-PICO" of the overall PICO on the overall study population, rather than as a separate research question. This would ensure that effects in subsets vs full population are analysed, evaluated and interpreted in conjunction together, taking into account the credibility of the results for the subsets according to good scientific best practice for subgroup analyses.	There is no request for rationale from MS in the scoping process guideline. Moreover, the scoping process is not data driven, and PICO therefore could be independent of the study population
Mihai Rotaru - EFPIA	General		Guideline review and methodological advances The guideline should be open for newly developed methods established after this guideline comes into effect. A corresponding review process for an update of the	These discussions are out of the scope of this

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			guidelines should be implemented to ensure that the guideline reflects the current state-of-the art. There should be ongoing scientific discourse between assessors, academia, and industry to discuss and consider latest developments in the methodology. Ideally, platforms are established (independent of specific guideline review processes) to promote exchange.	guideline.
Mihai Rotaru - EFPIA	General		Post-hoc analyses "Post-hoc" refers to any analysis that is specified after the data of a clinical trial have been observed (in contrast to the confirmatory setting in which "post-hoc" is often used for any analysis that has not been specified in the Statistical Analysis Plan (SAP) of the clinical trial). Pre-specification, therefore, may not only refer to analyses specified in the protocol and/or SAP of a trial, but also to any complementary analysis that is requested by the HTA with a strong scientific rationale, as well as any statistical analysis that a sponsor may specify in a separate SAP, i.e., HTA SAP, before conducting the analysis in order to meet a HTA body's request. Additional complementary analyses may also serve to strengthen the robustness and consistency of the data, even if not pre-specified. We therefore recommend to change the wording throughout the document. A better distinction could be: "prespecified in the context of the CSR", "requested for the JCA with a rationale", "well defined to show robustness of results", "after observation of the data".	We think the factual distinction between what was planned according to the SAP of a submitted study and what was performed post-hoc is sufficient for allowing MS to perform their appraisal at the national level. "Well-defined to show robustness of results" is something to be appraised at the MS level.
Tanja Podkonjak, Takeda	General		This guidance document, D4.5, is named Applicability of Evidence, however guidance document D4.6 Validity of Evidence also includes discussions and recommendations on the applicability of evidence. In the two documents 'applicability' have different meanings and cover different	We agree this guideline is not about applicability of evidence and is

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			concepts: D4.5 refers to 'methodological issues related to inferential statistical analyses' whereas D4.6 discussed 'external validity and generalisability'. To avoid confusion, we suggest using different terms to distinguish between these two guidance documents and the elements to which they refer.	about the specific issues that are covered within the guideline. Nonetheless, we are bound to the title because of contract with the EU.
Tanja Podkonjak, Takeda	General		Overall, we are concerned that there is no harmonisation of methods nor a pan-EU perspective reflected in the current document. Instead, the guidance document suggests a fragmented approach with the analyses and sub-analyses of all MS requests presented. Takeda support the JCA being a truly pan-European assessment which includes a harmonised approach to methods and the analyses required (i.e. populations of interest). Any deviations or outlying needs of MS are best addressed in local, complimentary evidence submissions, as accounted for in the HTA Regulation's stipulation that countries may perform complementary assessments (Recital 15).	These concerns are out of scope of this guideline and are addressed in the scoping process guideline.
Tanja Podkonjak, Takeda	General		Overall, unplanned, additional analyses should be kept to a minimum despite potential variety of MS requests, particularly for non-stratified subgroups. As subgroup analyses have several recognized limitations there should generally be a parsimony usage of these exploratory analyses, i.e., by focusing on key endpoints and on key subgroups of biological interest. When interpreting the results of subgroup analyses the following aspects and limitations should be considered • Lack of power • Consistency of effects in the multiple testing framework • Likelihood of false positive results • Qualitative interaction vs quantitative interaction (i.e., are there substantial differences in subgroup categories?) • Comparability between treatment groups within the subgroups	There are no limitations in the HTAR to limit MS requests (see 4.2 'Scoping process'). Appraisal of effects are out of scope of the JCA and will be left at national

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Conclusions about differential effects in subgroups should only be drawn based on adequate statistical interaction tests and only with sufficient credibility through biological plausibility (with clinical, pharmacological, or mechanistic rationale) and replication (in multiple data sources) (Kisser 2021). Finally, new guidelines on this topic should be open for innovative approaches beyond interaction tests.	level only.
Tanja Podkonjak, Takeda	General	132-133	The guidance states that 'while recommendations in this guideline may be better suited for RCTs, they can apply to various study designs.' (line 132-133). In the discussion, individual clinical studies are separated from evidence synthesis and mainly focus on RCTs. However, guidance document D4.6 on Validity of Evidence, includes both RCTs and observational studies as being used in a JCA. Several considerations of multiplicity, subgroup analyses, sensitivity analyses and others could be different between RCTs and observational studies. For example, some multiplicity adjustments suitable for observational studies (i.e., multiple time points) might be missed in the individual clinical studies section in the current guidance on Applicability of Evidence We request an additional section and discussion be added to the current guidance which addresses single-arm trials specifically, and another which addresses observational studies specifically.	While the requirements of reporting we proposed are more likely to be associated with inferential statistical hypothesis testing performed in RCTs, the factual elements for reporting we address in this guideline can be used in other settings.
Tanja Podkonjak, Takeda	General	Section 3.2	The scope of the HTA process is not to repeat the EMA decision, but it is to estimate the comparative effectiveness of new drugs, as well as the certainty of the results and the strengths and weakness of the analyses/methods and evidence submitted by the HTD (see lines 96-98). Defining a hierarchy of endpoints by the prespecified hypotheses and multiplicity adjustment procedures is implicitly already a value judgement, as significance levels and uncertainty of analyses that are acceptable are a national decision of the member states.	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Therefore, for endpoints which have not been part of the prespecified testing hierarchy no adjustment for multiplicity should be applied. As different countries may apply different significance levels for their decision making on the primary endpoints of the trial, conclusions on the significance of a given result should be avoided. Not meeting a prespecified significance level should not be flagged as a "factor that affects certainty of the evidence". Takeda supports EFPIA's proposal on the requirements for JCA reporting.	
Tanja Podkonjak, Takeda	General		Prespecified test hierarchies and formal alpha-adjustments Takeda is concerned about the high emphasis placed on controlling for the familywise error rate (FWER) and the requirement for test hierarchies on the impact this may have on the endpoints which are able to meet these stringent criteria. Requiring test hierarchies would limit the number of endpoints usable for the assessment.1 In specific, we are concerned about the impact of this approach on PROs and HRQoL which are often exploratory endpoints in clinical trial designs but are relevant endpoint for patients, caregivers, society and many healthcare systems and should not be excluded from a JCA due to a rigid requirement on prespecified test hierarchies. These are currently required by select HTA agencies, and we recommend these specific preferences requested by single MS remain at the level of national assessments. Reference: (1.) Kisser, A (2021). https://link.springer.com/article/10.1007/s10198-021-01400-2	Already addressed issue.
Hervé Tchala Vignon, Zomahoun/ INESSS			No comments	Thank you!
EFSPI	general		Hypothesis testing We recommend recalling, for example in the introduction section, the different purposes of HTA vs Marketing Authorization (MA). Whilst MA has a strong focus on confirmatory conclusions (testing framework), HTA aims at providing causal effect estimates for domains needed for reimbursement decision making. Beyond providing estimates along with their uncertainty, this also includes a qualification of the strength of the underlying evidence. The current draft Guideline D4.5 approaches many topics	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			with a testing framework, which does not consider other approaches usefulness for achieving JCA reports fulfilling the objectives laid out in the EU HTA Regulation.	
EFSPI	general		Prespecification "Post-hoc" refers to any analysis that is specified after the data of a clinical trial have been observed (in contrast to the confirmatory setting in which "post-hoc" is often used for any analysis that has not been specified in the Statistical Analysis Plan (SAP) of the clinical trial). Pre-specification, therefore, may not only refer to analyses specified in the protocol and/or SAP for the CSR of a trial, but also to any analysis that is requested for the JCA as part of the standard complementary analyses and laid out in the guidelines as well as any statistical analysis that a sponsor may specify in a separate SAP, i.e., HTA SAP, before conducting the analysis in order to meet a HTA body's request. The clinical trial SAP that is developed for the regulatory process is considered to be the foundation for the analyses used in the JCA, especially with regards to the operationalization of endpoints. However, due to differing research questions, complementary statistical analyses may be required for the JCA. To minimize data driven analyses, a strong scientific rationale should be given by the MS during the scoping phase, when requesting complementary analyses for the JCA. In this context, such analyses are not to be considered "post-hoc". Additional complementary analyses may also serve to strengthen the robustness and consistency of the data, even if not pre-specified. Early involvement of the HTD in the scoping process and an agreement on the analyses to be presented is absolutely critical to ensure that analyses needed for the JCA can be properly prespecified.	Already addressed issue.
EFSPI	general		Dealing with multiplicity The scope of the HTA process is not to repeat the EMA decision, but it is to estimate the effectiveness of new drugs, as well as the certainty of the results and the strengths and weakness of the analyses/methods and evidence submitted by the HTD (see lines 96-98). Due to this guideline, effectiveness also includes safety (lines 134-	The guideline only addresses how elements pertaining to certain types of statistical

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			135), whereas in clinical studies, safety is tackled very differently than effectiveness (limited pre-planning, limited hypothesis testing etc). The pre-specification in the trial is done to protect against multiplicity and data driven decisions. But when the primary analysis is well-defined, sensitivity analyses should not give rise to multiplicity concerns. Adjustment for multiplicity should therefore not be mandatory, and definitely not for secondary or sensitivity analyses, nor should there be the need to control the alpha level across all defined PICOs. To avoid chance findings due to multiplicity issues, the scope of the JCA should be based on a very limited set of PICOs and well-defined analyses that are necessary to support the HTA processes at member state level. Whenever feasible, the research question should be answered on the basis of results with a high certainty, like the preplanned analyses of the trial. Early involvement of the HTD in the scoping process and an agreement on the analyses to be presented is absolutely critical to ensure that analyses needed for the JCA are properly prespecified.	analyses should be adequately reported when assessors will encounter these kinds of analyses. How PICO should be requested and how HTDs should be involved in this process are out of the scope of this guideline.
EFSPI	general		Subgroup analyses As subgroup analyses have several recognized limitations there should generally be a parsimony usage of these exploratory analyses, i.e., by focussing on key endpoints and on key subgroups of biological interest. When interpreting the results of subgroup analyses the following aspects and limitations should be considered: • Lack of power • Consistency of effects in the multiple testing framework • Likelihood of false positive results • Qualitative interaction vs quantitative interaction (i.e., are there substantial differences in subgroup categories?) • Comparability between treatment groups within the subgroups Conclusions about differential effects in subgroups should only be drawn based on adequate assessment of the credibility, including statistical interaction tests and biological plausibility (with clinical, pharmacological, or mechanistic rationale), prespecification and replication (in multiple data sources). Finally, new guidelines on this topic should be open for innovative approaches beyond interaction tests.	Duplicated comments. See answer above.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
EFSPI	general		Estimands Paragraph 7 claims to be about sensitivity analyses, but deals with estimands. Estimands should have their own section. It should be clarified how the different types of estimands that may occur in a clinical study relate to the operationalization discussed in Section 4.2.4. Also, one of the requirements for JCA reporting is "Accurate and unambiguous endpoint definitions". This is a good opportunity to align with regulatory language and introduce estimands into HTA reporting.	While we agree the estimand framework has some value, we do not think it is necessary to discuss it more within this guideline. A discussion about its value in the context of HTA is better suited in the D4.6 guideline.
MTE	General		The document would benefit to further define on "terms" used, and/or build a glossary of terms/definition eg. with some examples (incl. terms as in 5.1)	Terms are defined the first time they appear in the guideline.
MTE	General		An involvement of HTD in PICO's discussion, contributing to put forward the consideration of the statistical analysis plan to have a clear understanding of the statistical power to support additional analysis, should be ensured. Also further consideration on a priori discussions and consideration within the JSC to manage expectation on subpopulation, post-hoc analysis, etc. to be included in the JCA (and what should be left for national analysis) need to be established.	These concerns are out of the scope of this guideline.
MTE	General		The scope of the JCA should be based as much as possible on the prespecified analyses in trials and should be limited to analyses that are absolutely necessary for the JCA. The focus should be on an agreed, limited set of appropriate endpoints. Complementary, sensitivity and subgroup analyses should not be a	Duplicated comments. See answer above.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			regular occurrence in the JCA. The base case approach should be to conduct meaningful sensitivity analyses for situations where the investigation of the robustness of the results is crucial. The decision to conduct additional analyses should be explained on a case-by-case basis, and a strong scientific, clinical, biologic or regulatory rationale be given by the requestor. In some cases, it could be worth to include the expert views (medical experts and statisticians). Early involvement of the HTD in the scoping process and alignment on analyses needed for the JCA is therefore of utmost importance. There should be a constructive dialogue implemented between the assessors and the HTDs to ensure a methodological exchange before prespecifying and conducting the corresponding analyses for each JCA dossier.	
MTE	General		The document would benefit in the evidence synthesis analysis from consideration of special trial designs (including those referred to in D4.6 validity of clinical studiesand also adaptive trial design (possible in context of interim analysis) and especially for medical technologies more general of consecutive- adaptive evidence generation over time and associated analysis with an increased strength of evidence over time.	This guideline is about specific issues that can apply to various study designs. We do not think there is a need to expand the guideline.
MTE	General		The document would benefit if also considerations were given on how to handle JCA with a limited availability of evidence at time of assessment.	These concerns are out of the scope of the guideline.
MTE	General		JCA methodology should be consistently applied. Subpopulations of interest may be specified during the assessment scope as separate research questions, and a strong clinical, biologic or regulatory rationale needs to be given by the member states to	These concerns are covered in the scoping process guideline and

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			request subpopulations to be analysed. If analyses of patient subsets from a trial population need to be used, results in these subpopulations should be analysed and interpreted consistently with principles of best practice for subgroup analyses defined in the guidance (preplanning, interaction test; biologic plausibility). A priori, the effect estimate from the overall study population is the best estimate within each subpopulation, and the main population in the PICOs should be the overall study population. We recommend that subpopulations should be only requested if there is a strong rationale. MS can always ask for complementary analyses, if a subpopulation analysis is very specific to their national assessment. If a subpopulation needs to be analysed for the scope of the JCA, this should be reflected as a "sub-PICO" of the PICO on the overall study population, rather than separate research question. This should ensure that effects in subsets vs full population are analysed, evaluated and interpreted in conjunctionwith each other. The HTD should have the possibility to cross reference the information in the submission dossier.	are out of scope of this guideline.
MTE	General		The JCA reporting would benefit from a segmentation on the strength of evidence and a more clear guidance on the optimal use of eg. real world evidence and different trial designs in evidence synthesis. Within the reporting consideration on eg. 3 classes of strength in evidence could be considered (in line with professional societies).	These concerns are tackled within the D4.6 guideline validity of clinical studies.
MTE	general	-	The flow of this guidance is somewhat confusing, particularly the repetitiveness in the titles (i.e. On JCA reporting) and the information contained in each of the boxes. Consideration should be given to reduce repetition and improving the format and flow to the overall	Thank you. We will adopt this proposition for the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			document.	
MTE	general	-	The Requirements for JCA reporting should be numbered, have a title, and be presented as a checklist. This may mean some of the dot points need expansion or should be combined.	Thank you. We will consider the relevance of this proposition for the next version of the draft.
Natacha Bolaños, Lymphoma Coalition Marjorie Morrison, Lymphoma Coalition	General		**<u>Endpoints</u>** Indolent lymphomas are characterized by their incurability. There is an estimated median survival that extends beyond 15 years for lymphomas such as follicular lymphoma, Waldenstrom macroglobulinemia, and others. Due to their slow histology growth, "measuring overall survival after first-line therapy is an impractical study endpoint." (1) Treatment that addresses immediate disease control must also consider the longer-term consequences of therapy and risk of toxicities patients may experience over time. By design, surrogate endpoints in oncology clinical trials are used in regulatory and clinical decisions. As far back as 2014, regulatory bodies, namely the US Food and Drug Administration or FDA, "approved drugs for 83 oncology indications with 66% approved on the basis of surrogate outcomes." (2) As surrogate endpoints are reliant on retrospective analysis of completed clinical trials, the systemic challenges associated with data needs to be universally addressed with standardised and mandatory data collection and reporting processes implemented across all registries in Europe. Overall survival and progression-free survival are clinical endpoints however, patients with lymphomas that are slow to progress (low tumour burden) often experience management options that require prolonged surveillance. Therefore, it is impractical to focus on these endpoints exclusively as many patients with lymphoma are diagnosed at an age where there are competing causes of death and/or existing comorbidities to consider. Thus, it would be advantageous when addressing	These concerns are out of the scope of this guideline and are covered in the scoping process guideline and the upcoming endpoint guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			"endpoints of interest" to take into consideration factors that contribute to variability in response to initial treatments and outcomes. Additionally, factors such as the time required to complete clinical trials or studies where overall survival endpoints and progression-free endpoints are applied (estimated at five to eight years in trials for follicular lymphoma), the risk of potential bias with the introduction of subsequent therapies, and delays in the development of newer therapeutic interventions should also be considered. (3) **Given the scarcity of registries and/or registry data, in addition to the lack of differentiating between lymphoma sub-types, it is essential to consider endpoints that establish early or late progression of disease, complete response (and how long this can be maintained), and quality-of-life.**	
EHA	General		Document D 4.5 and the PICO question: It is unclear whether the text refers to methods or interpretation. There should not be differences in the methods used and that multiple testing for example should be handled in the same way everywhere. Some countries would look at HR and others at the differences in medical survival but the methods are similar. Also the consistency between a research question and the PICO should not be assessed differently. Same issue lines 333 and 336 with the different metrics that can be used by different countries. There should be a minimum set of core effect measures (eg in hematology and for other medical specialties as well) that every country would use.	The purpose of the guideline is to allow a factual reporting of the elements MS need to perform their appraisal at the national level. Effect measure are tackled within the D4.4 guideline "Endpoints". How PICO are requested is the scope of the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			For appropriate subgroup definition, health care professional expert and patient opinion/consultation should be sought early, ideally to pre-specify most relevant subgroups. This may be a gateway to more patient-centric research in the next, subsequent iteration (e.g. as a subpopulation for PICO). Subgroups should ideally be defined using IVDR or other methods that allows clinical implementation and, potentially, subpopulation-focused trial using identical criteria in the future. Subgroups should ideally not only be seen as tools to reduce uncertainty, but also as opportunities for improvements in outcome focused on special unmet need groups. For this purpose, EU in vitro diagnostic device regulation (IVDR) legislation should ideally be dynamically reviewed to serve the purpose of implementation of more patient-centric diagnostics in clinical evidence generation for HTA purposes, and EUnetHTA should seek active engagement with IVDR legislators for this purpose.	D4.2 guideline scoping process.
Dr. Thomas Ecker, Ecker + Ecker GmbH	general		Ecker + Ecker GmbH, a healthcare consultancy based in Germany with strong expertise in the early benefit assessment, welcomes the establishment of a European Health Technology Assessment (HTA) fostering closer cooperation between member states on health technology assessment by introducing a permanent framework for this joint work. The legal requirements for a European HTA have been determined as a legislative act by the end of 2021 with the EU regulation 2021/2282. From 2025, before placing innovative medicinal products on the market, oncology products and ATMP are subject to a European joint clinical assessment. In the next step, Orphan Medicinal Products	These concerns are out of the scope of this guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			(OMPs) will follow beginning in 2028 and from 2030, all medicinal products will have to go through the European assessment. While the regulation does not come into force until 2025, the process of implementation is already ongoing to ensure effective application from January 2025 onwards. At present, the development of a methodology for joint HTA work is facilitated by the European Network for Health Technology Assessment (EUnetHTA) 21 consortium.	
Dr. Thomas Ecker, Ecker + Ecker GmbH	general		Some points, especially in the JCA requirement boxes, are repeated numerous times. A more compact structure with one summarizing box containing all general requirements at the end of the introductory section of a chapter could be helpful for the overall readability.	Already addressed issue.
Dr. Thomas Ecker, Ecker + Ecker GmbH	general		The requirements need to be specified unambiguously, so that the HTD can submit all necessary data and evidence. The draft guideline does not meet this need, as it is vague at multiple points, especially: • It is unclear, which subgroup analyses have to be submitted (only for a priori planned endpoints or for every endpoint). • It is unclear, which measures in particular are meant in the JCA requirement boxes e.g., `appropriate measures for statistical precision´. Subjective und ambiguous clauses such as `appropriate´ should not be used. It is unclear, if estimands should be used outside of the concept of sensitivity analyses. This should be stated explicitly.	These concerns are out scope and are tackled within the D4.2, D4.4 and D4.6 guidelines.
Sebastian Werner vfa	General		The guidance "Applicability of evidence" discusses multiplicity, subgroup, sensitivity, and post hoc analyses for individual clinical and evidence synthesis studies. The guidance does not endorse a particular methodological approach for the appraisal of these aspects, as different member states could consider these methodological aspects differently. The guidance gives no recommendation on how these aspects might influence the joint clinical assessment regarding the degree of certainty of the relative effects. It gives only directions on the reporting of "all the necessary elements" that member states need to carry out the national appraisal of the clinical added value of the health technology.	There is no statement in the HTAR for requiring convergence in HTAb methodology. On the contrary, the HTAR allows

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			The guideline is clear that several methodological approaches will be used by the member states in dealing with the multiple hypothesis testing, subgroup, sensitivity, post hoc analyses, and other aspects (effect measures and operationalisations for an endpoint). The guideline is also clear that the submission dossier must comprise all the necessary elements for these assessments of the member states. Thus, the guidance indicates that multiple methodological approaches for data analysis must be addressed in the submission dossiers and joint clinical assessments. The obligation to address multiple methodological approaches in submission dossiers and joint clinical assessment clearly constitutes duplication. Further, the objective of the European HTA Regulation to harmonize the clinical assessment through convergence of HTA methodology is not met. The notion that multiple methodological approaches must be addressed to accommodate differences in national appraisal, sends a disappointing signal against a harmonized European methodological framework for assessing applicability and degree of certainty of the clinical evidence. Keeping multiple methodological approaches pose risks to a workable and efficient European HTA System. The vfa recommends forming clear recommendations on how to deal with the multiplicity, subgroup, sensitivity, and post hoc analyses in the joint clinical assessment with details on how these aspects might influence the degree of certainty of the relative effects. Further, the vfa recommends establishing a harmonized European methodological framework for joint clinical assessment that provides a uniform methodological approach for data analysis and synthesis that member states commonly accept. The European methodological framework should not be built as a collection of multiple methodological approaches of different member states.	differences between MS, by stating that the data request (scoping process) 'should be inclusive and reflect Member States' needs' (article 8) and that MS can 'consider the parts of those reports relevant in that context' (article 13).
Sebastian Werner vfa	General		It is recognized that Market Authorization (MA) and Health technology assessment (HTA) follow different goals, resulting in different views on multiplicity or "post-hoc" definition when using the evidence of pivotal studies, which were primarily designed for MA purposes and which are "re-used" for HTA-purposes (Leverkus 2018a, Leverkus 2018b). The central aspect of market authorization for new drugs is the question whether a drug is efficacious (with positive benefit – risk) or not. For this binary decision the Neyman-Pearson approach is a necessary concept for regulatory decision making	As stated in the introduction (line 101-113), MS appraises certain aspects differently. Therefore, the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			to allow formal, confirmatory conclusions for the primary endpoint(s), while secondary and exploratory endpoints only provide supportive or exploratory information: A null hypothesis must be accepted or rejected and a control for type I errors must be implemented to prevent false positive decisions. Additionally, the benefit must outweigh possible risks regarding safety. For HTA the principles of evidence-based medicine apply to answer clinical questions for a retrospective systematic review of all available and relevant evidence discussing the aspects of internal and external validity and sources of bias. The statistical methods are based on the Fisher approach: each endpoint is analysed separately, and the corresponding p-value is an estimate for the strength of the evidence without controlling overall type I and II errors (Lehmann, 1993). In contrast to the binary decision problem in market authorization, HTA handles an estimation problem to evaluate the extend and certainty of an added benefit compared to other therapies based on all patient-relevant outcomes. So, the statistical methods should be based on the Fisher approach: each endpoint is analysed separately, and the corresponding p-value is an estimate for the strength of the evidence without controlling overall type I and II errors. Therefore, given the specific HTA context prespecified clinical trial test hierarchies and formal alpha-adjustments are not recommended in the HTA environment (Kisser 2021). The vfa recommends considering the differences of the statistical approaches of Regulatory and Health technology assessment and a re-assessment of the importance of the multiplicity problem in the context of HTA. Prespecified clinical trial test hierarchies and formal alpha-adjustments are not recommended in the HTA environment. • Leverkus 2018a. Leverkus F, Gillhaus J, Knoerzer D, Kupas K, Nicolay C, Hennig M: AMNOG meets EMA - Methodological Areas of Debate from an Industry Point of View (Part 1). Pharmazeutische Medizin 2018, Jahrgang 20, Heft 1, März • Leverkus 2018b. Leverkus F, Gillhaus J, Knoerzer D, Kupas K, Nicolay C, Hennig M: AMNOG meets EMA - Methodological Areas of Debate from an Industry Point of View (Part 2). Pharmazeutische Medizin 2018, Jahrgang 20, Heft 2, Juni • Lehmann, E. L. 1993. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? Journal of the American Statistical Association, 88(424), 1242-9. • Kisser 2021. Kisser, A., Knieriemen, J., Fasan, A. et al. Towards compatibility of EUnetHTA JCA	guideline is intended to help assessors in reporting all the necessary elements to cover every MS needs.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			methodology and German HTA: a systematic comparison and recommendations from an industry perspective. Eur J Health Econ 23, 863–878 (2021). https://doi.org/10.1007/s10198-021-01400-2	
Sebastian Werner vfa	General		It is recognized that Market Authorization (MA) and Health technology assessment (HTA) follow different goals, resulting in different views on "post-hoc" definition. "Post-hoc" refers to any analysis that is specified after the data of a clinical trial have been observed. Specifically, in the Marketing authorization confirmatory setting "post-hoc" is often referred to as any analysis that has not been specified in the Statistical Analysis Plan (SAP) of the clinical trial. However, "pre-specification", may not only refer to analyses specified in the protocol and/or SAP of a trial, but also to any analysis that is requested by the HTA bodies as part of the standard complementary analyses as well as any statistical analysis that a sponsor may specify in a separate SAP, i.e., a HTA SAP, before conducting the analysis to meet a HTA body's request. The clinical trial SAP that is developed for the regulatory process is the foundation, especially with regards to the operationalization of endpoints. However, due to possibly deviating research questions (PICO), complementary statistical analyses on specific subpopulations may be required in the HTA context. In this context, such analyses might not be considered "post-hoc" as they are defined by the scope of the HTA as "a-priori" analyses. Additional complementary analyses may also serve to strengthen the robustness and consistency of the data, even if not pre-specified. The vfa recommends elaborating on the definitions of "prespecification" and "post hoc" to address differences between Regulatory and Health technology assessment. The specific context of HTA should be considered. Analyses requested by the HTA authorities regarding specific PICO questions, should not be designated as "post hoc" but rather be considered as "prespecified" hypothesis to be tested in the systematic review (i.e., joint clinical assessment).	Already addressed issue.
Sebastian Werner vfa	General		Health technology assessment (HTA) in contrast to regulatory assessment for Marketing Authorization, aims at providing causal effect estimates for additional domains needed for reimbursement decision making, as well as a qualification of the strength of evidence supporting these estimates. For HTA, exploratory analyses may play a bigger role than in regulatory approval. The "importance" of endpoints for HTA may differ from the endpoint hierarchy in the trial. Exploratory endpoints may be key for HTA. Therefore,	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			regulatory assessment has a stronger focus on formal testing, while HTA focuses on estimation. Therefore, the differences resulting from different goals should also be reflected in a different handling of multiplicity, the estimand concept, predefinition, and acceptance of post hoc analyses. The vfa recommends considering the differences of Regulatory and Health technology assessment by differential handling of multiplicity, the estimand concept, predefinition, and acceptance of post hoc analyses. The Joint clinical assessments should not be a repetition of regulatory assessment. The joint clinical assessment should consider the results of the regulatory assessment, without repeating it or calling it into question.	
Sebastian Werner vfa	General		Subgroup analyses have many recognized methodological limitations that decrease their credibility. Therefore, they should be used with parsimony, by focussing on key endpoints and on key subgroups of specific biological plausibility. When interpreting the results of subgroup analyses the following aspects should be considered: biological plausibility, lack of statistical power, consistency of effects in the multiple testing framework, likelihood of false positive results, qualitative interaction vs quantitative interaction (size of the interaction) and comparability of characteristics between treatment groups within the subgroups. Conclusions about differential effects in subgroups should only be drawn based on adequate statistical interaction tests and only with established biological plausibility through clinical, pharmacological, or mechanistic rationale and possibly replication in multiple data sources (Kisser et al. 2021). Decisions about biological plausibility should be informed by the discussions of the regulatory approval. The vfa recommends limiting the requests for subgroup analyses to a minimum. Requests should be focused on key subgroups with established or possible biological plausibility for differential effects. The request should also focus on key endpoints, which, according to biological rationale, are likely to show the effect. Conclusions about differential effects in subgroups should only be drawn based on adequate statistical interaction tests and only with established biological plausibility. Decisions about biological plausibility should be informed by the discussions of the regulatory approval.	There is no request for an explicit rationale from MS in the scoping process guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			• Kisser 2021. Kisser, A., Knieriemen, J., Fasan, A. et al. Towards compatibility of EUnetHTA JCA methodology and German HTA: a systematic comparison and recommendations from an industry perspective. Eur J Health Econ 23, 863–878 (2021). https://doi.org/10.1007/s10198-021-01400-2.	
Sebastian Werner vfa	General		Sensitivity analyses may be useful to estimate the robustness of the results. However, sensitivity analyses for all analyses, as a regular requirement, is not reasonable. The general approach should be to conduct meaningful sensitivity analyses for situations where the investigation of the robustness of the results is crucial. The vfa recommends conducting sensitivity analyses in special data situations where the investigation of the robustness is crucial. The decision to conduct sensitivity analyses should be justified on an individual basis.	We do not require sensitivity analyses for all outcomes that will be reported according to the scoping process. It will be clarified in the next version of the draft.
Sebastian Werner vfa	General		There should be a constructive dialogue implemented between the assessors and the applicants to ensure a methodological exchange before prespecifying and conducting the corresponding analyses for a specific dossier.	These concerns are out of the scope of the guideline.
Sebastian Werner vfa	General		There should be ongoing scientific discourse between assessors, academia, and industry to discuss and consider latest developments in the methodology. Ideally, platforms to promote exchange should be established (independent of specific guideline review processes and individual clinical assessments).	These concerns are out of the scope of the guideline.
Sebastian Werner vfa	General		The guideline should be open for innovative methods established after this guideline comes into effect. A corresponding review process for an update of the guidelines should be implemented to ensure that the guideline reflects the current state-of-the art.	These concerns are out of the scope of the guideline.
Storz-Pfennig/ Ermisch – GKV-SV	General	n/a	The draft addresses a several issues regarding the applicability of the evidence base to the PICO questions of the assessment. Trials forming the evidence base should be selected according to their relevance and need to be applicable to the assessment questions. However, it is not the primary objective of the assessment to evaluate these individual trials. Therefore, we support specification that the assessment should	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			answer the PICO question resulting from the scoping process, including analyses of subpopulations and relevant endpoints and evidence synthesis/meta-analysis as necessary. It is to be expected that HTDs support claims of effectiveness by sufficiently reliable evidence according to the PICO questions. Member states may require different information on different PICO-elements. Not all aspects (multiplicity etc.) discussed in the draft are of the same relevance in view of applicability. Multiplicity issues and, in particular, sensitivity analysis primarily refer to the validity and robustness of effect estimates as such and only to a lesser degree to applicability in terms of particular subpopulations or comparators. While methods regarding multiplicity and other issues discussed in the draft guidance in evidence synthesis/meta-analysis are comparatively undeveloped, no grave concerns are identified and carefully defined PICO-questions may avoid or alleviate issues raised regarding pre-specification in the context of evidence synthesis based on already acquired data. Evidence synthesis produced by the HTD for the purpose of HTA and reported in the dossier should provide the necessary data for answering the PICO questions as well as address methodological issues.	
Advanced Medical Services GmbH – AMS	general		With regard to the **Submission Dossier Template** we would like to propose a "design" with **text boxes** that contain specific descriptions and definitions of what is required and how it has to be presented (for an example see the submission dossier template for the German Benefit Assessment that has text boxes shadowed in grey). Thus, the health technology developer (HTD) will have a comprehensive instruction of what is expected in the Submission Dossier corresponding to the listings in the boxes of D4.5: "**Requirements for JCA reporting**".	These concerns are covered by other deliverables of the EUnetHTA 21 initiative.
Bayer	general		It is mentioned several times that member states might have their own requirements why it must be made sure that as many requirements as possible are met in the JCA or that member states should complement JCA by national assessment. This refers to effect measures (p.12, line 333-335), operationalisation of endpoints (p.12, line 336-340), subgroup analyses (p.13, line 368-369), and acceptability of missing date (p.17, line 493-495). Beyond additional national assessment due to missing PICOs at JCA level, this level of granularity will lead to a very high number of data not being assessed in JCA and consequently to extensive additional national HTA assessments with high data demand. It is therefore of high importance to give leeway to national assessment with a sense of proportion and rather focus on a common understanding of this methodology and a common sense of added value.	Already addressed issue.
Bayer	general		The conclusions regarding single arm trials and basket trials are challenging for	The current

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			oncological products. This guidance would continue the disparate access in EU for transformative cancer therapies that we see today. There is no target-oriented consideration of alternatives such as external controls or indirect comparisons (even if relegated to other consultation papers on meta-analyses and indirect comparisons), which implies that these alternatives will not play a considerable role in the assessment	guideline covers specific considerations that can apply to various study designs. The assessment of the certainty of results of individual studies is covered in the validity of clinical study guidelines. Regarding indirect comparison, two guidelines are produces (a methodological and a practical one).
Bayer	general		Many topics in this document adopt a hypothesis testing framework (e.g., the reporting requirements in Section 4.2). While hypothesis testing is important for regulators in the "pre-marketing authorization" stage, HTA should place greater focus on adequate effect size estimation (without neglecting uncertainty quantification). Regulators often make decisions based on a randomized controlled study comparing the active treatment against placebo or standard of care. The study is designed to attain appropriate power for statistical testing in regulatory decision-making. Conversely, HTA requires comparative effectiveness estimates versus all active treatments available on the market. Direct head-to-head comparisons in RCTs are often unavailable. Evidence synthesis may rely on observational data sources or on	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			the synthesis of disparate sources (and often external published data), e.g., network meta-analyses, indirect treatment comparisons, generalizability or transportability analyses. In general terms, relative effect estimates of interest are less precise and more uncertain. The methodologies that are applied to adjust for confounding or to account for different sources of data are not powered for hypothesis testing. Moreover, it is debateable whether hypothesis testing is of relevance in systems where the relative effect estimates are used to inform the effectiveness inputs to a health economic model, and the decision is based on cost-effectiveness in this model. Our suggestion is to move away from a hypothesis testing framework and place greater focus on appropriate (causal) effect size estimation, which is more relevant for decisions made by HTA bodies.	
Bayer	general		The document places heavy focus on multiplicity adjustment (e.g., reporting requirements in Section 4.2). Unplanned analyses are common in HTA. Member states often make post-hoc requests about different subpopulations, outcomes/endpoints, etc. As a result, we suggest that multiplicity adjustments should be mandatory, particularly for secondary and sensitivity analyses. "Almost all proposals for controlling the type I error rate assume that this is a responsibility that the trialist/sponsor/ designer should shoulder. Within the context of making a formal bid for registrations of a pharmaceutical this is a sociological constraint one can accept (just) as being practical, since, currently, regulators prefer not to be burdened with the responsibility. However, it is hardly logical. Why should end-users accept some complicated scheme (...) which uses an approach they may not favour? When it comes to making their own inference, they are better off with all the unadjusted p-values rather than some mangled construct." (S Senn. Section 4: "Who should adjust for multiplicity? The author or the reader?" doi: 10.1002/bimj.201700032) The quote is based on the regulatory approval context but also applies in the HTA context. Different HTA agencies might have different research questions, value judgements and requirements, some unplanned, and controlling the alpha level across all PICOs seems to be unnecessary.	The main purpose of this guideline is to allow proper reporting of all the necessary elements MS can use for their appraisal at national level. The guideline does not require adjustments must be made for all PICOs. It only requires to report how it was performed in the corresponding submitted evidence. We will consider if a

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				clarification is necessary for the next version of the draft.
James Ryan AstraZeneca	General		Thank you for the opportunity to respond to the consultation. We have inputted into the EFPIA response to this consultation and want to confirm that we are aligned with their response.	Thank you.
Richard Birnie Lumanity HEOR	General		In the guidance on evidence synthesis the requirement to report null and alternative hypotheses along with associated p-values seems overstated. In the vast majority of cases evidence syntheses are not explicitly formulated in a hypothesis testing framework. Instead, the focus is on estimation of the magnitude and precision of relative treatment effects. The usual aim in evidence synthesis is to answer a research question in the general form: "What is the relative effect of *new treatment* compared to *alternative treatment options* for people with *health condition*?" Reframing this in terms of null and alternative hypotheses according to whether a p-value crosses what is ultimately an arbitrary threshold is likely to lead to oversimplification and misuse of p-values from poorly specified tests. In addition, substantial further work would be required to properly specify the hypotheses and identify properly designed and appropriately powered tests. As the guidance rightly notes, although methods from analysis of individual trials can be adapted to evidence synthesis, in practice this is almost never done and robust ways of doing this have not yet been worked out. We would recommend removing the requirements for reporting hypotheses and p-values for evidence syntheses. Instead we suggest focus on estimation of the relative effects, clear definition of the research question and proper assessment of the evidence base. We recognise that much of this is covered by other deliverables	The main purpose of this guideline is to allow proper reporting of all the necessary elements MS can use for their appraisal at national level. We agree that because evidence synthesis are based on data already analyzed, their analysis frequently differ from a strong confirmatory framework. We think this has been emphasized within the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				guideline. Therefore, the guideline does not imply evidence synthesis without strong confirmatory framework will be systematically dismissed.
Roche	general	general	We recommend recalling, for example in the introduction section, the different purposes of HTA vs Marketing Authorization (MA). Whilst MA has a strong focus on confirmatory conclusions (testing framework), HTA aims at providing causal effect estimates for domains needed for reimbursement decision making. Beyond providing estimates along with their uncertainty, this also includes a qualification of the strength of the underlying evidence. Therefore, the Practical Guideline D4.5 should provide recommendations and establish best practices for the estimation of causal treatment effects. Unfortunately, the current draft Guideline D.5 approaches many topics with a testing framework, which we consider of limited usefulness for achieving JCA reports fulfilling the objectives laid out in the EU HTA Regulation.	Already addressed issue.
Silke Walleser Autiero Medtronic	general	-	The flow of this guidance is somewhat confusing, particularly the repetitiveness in the titles (i.e. On JCA reporting) and the information contained in each of the boxes. Consideration should be given to reduce repetition and improving the format and flow to the overall document.	Duplicated comment.
Silke Walleser Autiero Medtronic	general	-	The Requirements for JCA reporting should be numbered, have a title, and be presented as a checklist. This may mean some of the dot points need expansion or should be combined.	Duplicated comment.
Silke Walleser Autiero	general	-	Consider the addition of a glossary of terms/definitions	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Medtronic				
GSK	General	General	There should be a constructive dialogue implemented between the assessors and the applicants to ensure a methodological exchange before prespecifying and conducting the corresponding analyses for a specific dossier.	Duplicated comment.
GSK	General	General	There should be ongoing scientific discourse between assessors, academia, and industry to discuss and consider latest developments in the methodology. Ideally, platforms are established (independent of specific guideline review processes) to promote exchange.	Duplicated comment.
GSK	General	General	The guideline should be open for innovative methods established after this guideline comes into effect. A corresponding review process for an update of the guidelines should be implemented to ensure that the guideline reflects the current state-of-the art.	Duplicated comment.
GSK	General	General	Multiplicity It is recognized that Market Authorization (MA) and HTA follow different goals, resulting in different views on multiplicity or "post-hoc" definition when using the evidence of pivotal studies, which were primarily designed for MA purposes and which are "re-used" for HTA-purposes (Leverkus 2018a, Leverkus 2018b): Within MA, the focus is on primary endpoints used for formal confirmatory conclusions (testing problem), whereas in HTA the focus is to assess the added benefit of the new drug (estimation problem). Given the specific HTA context prespecified clinical trial test hierarchies and formal alpha-adjustments are not recommended in the HTA environment (Kisser 2021).	Duplicated comment.
GSK	General	General	As subgroup analyses have several recognized limitations these exploratory analyses should generally be used parsimoniously, i.e., by focussing on key endpoints and on key subgroups of biological interest. When interpreting the results of subgroup analyses the following aspects and limitations should be considered: •Lack of power •Consistency of effects in the multiple testing framework •Likelihood of false positive results •Qualitative interaction vs quantitative interaction (i.e., are there substantial differences in subgroup categories?) •Comparability between treatment groups within the subgroups Conclusions about differential effects in subgroups should only be drawn based on adequate statistical interaction tests and only with sufficient credibility through	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			biological plausibility (with clinical, pharmacological, or mechanistic rationale) and replication (in multiple data sources) (Kisser 2021). Finally, new guidelines on this topic should be open for innovative approaches beyond interaction tests.	
GSK	General	General	Sensitivity analyses should not be a regular requirement in the HTA. The general approach should be to conduct meaningful sensitivity analyses for situations where the investigation of the robustness of the results is crucial. The decision to conduct sensitivity analyses should be explained on an individual basis. In some cases, it could be worth including expert views (medical experts and statisticians).	Duplicated comment.
GSK	General	General	The "Requirements for JCA reporting" listed in chapters 3.2, 4.2, 5.2., 6.2, 7.2, 8.2., 9.2. and 10.2 are mostly standard requirements for Good Scientific Practice and are therefore typically reported in the Clinical Study Reports (CSR) of the corresponding studies. For the purpose of JCA reporting there should be references to the corresponding sections of the CSR.	Duplicated comment.
Jasmine Toomey PHMR	NA	General	The guideline references many different documents which could be integrated to an appendix.	We do not think a comprehensive integration of all the documents referenced in the guideline would be advisable.
ISPOR	General		This is a good, very useful document, with clear and specific definitions and very relevant technical information for critical reading and the analysis of scientific literature in the generation of the HTA reports. The contents show an integral and descriptive approach. The document is helpful for the assessment and reporting processes at the national level.	Thank you!
ISPOR	General		Overall, this document summarizes these important statistical topics in a high level, but some aspects of the discussions may made clearer with more details. In particular, references to good case examples for how multiplicity adjustments are made in protocols (with pre-specified JCA country requirements). or use of sensitivity analyses, would be helpful.	This document is a practical guideline for helping assessors in reporting

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				necessary elements and is not intended to be a methodological or statistical textbook.
ISPOR	General		We do have one significant concern. From reading this document one would get the impression that the only source of evidence in scope is RCTs (despite a brief comment about "various study designs" on l. 133). HTA Regulation (EU) 2021/2282 (35) notes that observational studies can be helpful. And D4.6.1 (validity of clinical studies) includes some discussion of study designs based on RWE, which might range from an external control arm or supplemental survival data to a full retrospective cohort study. However, we found the discussion in D4.6.1 quite limited relative to the attention some other major agencies are paying to RWE. This particular consultation is similarly limited in that respect. Use of non-randomized study designs, primarily based on RWD, creates nuances for each of the 4 main statistical areas covered, introduces other analytical considerations, and suggests special attention to evidence synthesis involving both RCT and RWE results (especially if they are considered different levels of evidence). This would be important additional information, particularly given the proliferation of new treatment innovations with small or precisely defined patient populations, where regulatory approval may have been obtained based on a limited evidence base; this is where JCA could be particularly helpful for member states. Will a future consultation be paying more explicit attention to considerations related to RWE?	This document addresses specific issues that can be generally applies for various study designs. Considerations about RWE in the context of JCA are covered in the D4.6 validity of clinical studies guideline.
ISPOR		General	Your sections on real-world evidence and non-randomized studies are well-written but in general reflect traditional thinking in these areas. However, recent developments in natural experiments (note the 2021 Nobel Prize in economics) and target trial emulation (note the RCT-DUPLICATE work) have highlighted the potential for valid causal inference with observational data. We also think the value of external validity that real world evidence can contribute to decision-making is understated. While we certainly support the investigation of potential biases via tools like ROBINS-I, we thought your discussion of this area could have been more forward-looking, particularly given a growing need for post-approval evaluations. A fuller explication of	This comment refers to the D4.6 validity of clinical studies guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			this viewpoint can be found in a recent Value in Health Commentary: Berger ML, Crown WC. How Can We Make More Rapid Progress in the Leveraging of Real-World Evidence by Regulatory Decision Makers? *Value in Health*, Volume 25, Issue 2, 167 – 170.	
GSK	8-13, 15-16	Requirements for JCA reporting box	What's the meaning of 'unambiguous endpoint definitions'?	Adequate reporting of endpoints are more detailed in the D4.4 guideline on endpoints. We will refer to this guideline to complement this requirement.
ISPOR	6-10	3 Multiple Statistical Hypothesis Testing in Individual Clinical Studies (lines 150-258)	The discussion of multiplicity adjustment should be under the context of pre-defined primary, secondary and sensitivity analyses. For instance, one may conduct a sensitivity analysis to assess the treatment effect in a subpopulation. Since the purpose of this sensitivity analysis is to assess the robustness of the primary analysis result and it does not contribute to the main interpretation of the study finding, multiplicity adjustment accounting for this sensitivity analysis is not needed. The examples provided for multiple testing between Line 187-194 may or may not lead to multiplicity issues, depending on the pre-defined study plan on primary, secondary and sensitivity analyses. Therefore, the language of this paragraph here should not be deterministic. And it would benefit from more in-depth discussions to avoid ambiguity.	The cited paragraph proposes examples of situations leading to potential multiplicity issues. It does not imply de facto how these situations should be handled. We agree that depending on the context, the analysis of a

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				subpopulation could be considered either as a sensitivity analysis, but it could also be considered as a confirmatory analysis to confirm a benefit in another population than the primary one.
ISPOR	6-10	3 Multiple Statistical Hypothesis Testing in Individual Clinical Studies (lines 150-258)	It's worth noting that if the study is for regulatory submission, the multiplicity adjustment can also depend on your regulatory purpose, eg, efficacy (where multiplicity is usually corrected for) vs. safety (where it often is not).	Our guideline is intended to help assessors when reporting elements pertaining to analyses that has been performed in the framework of statistical hypothesis testing in their corresponding evidence submitted, irrespective of the type of outcome. The

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				proper way to perform safety analyses is not the concern of the guideline.
ISPOR	6-10	3 Multiple Statistical Hypothesis Testing in Individual Clinical Studies (lines 150-258)	In addition to statistical significance and corresponding approaches handling multiplicity problems in different scenarios (e.g., multiple outcomes, multiple time points, multiple treatments, multiple groups, or multiple effect measures), it might be useful to discuss suggestions on how one can utilize clinical significance to guide decision-making in different scenarios.	The guideline is not intended to be a methodological textbook.
Roche	10 - 13	Section 4	Evidence synthesis methods are important tools to provide relative effect estimates in the absence of direct, head-to-head studies. They achieve this, for example, through the analysis of published RCT data. As such, evidence synthesis is typically exploratory in nature and the formal confirmatory testing framework seems not applicable to evidence synthesis studies (the draft Guideline D4.5 acknowledges this in lines 269/70). However, the reporting requirements in Section 4.2 are still heavily centered around testing. We recommend to remove these requirements and to focus, instead, on estimation and established best practices, such as pre-specification in the scoping process (leading to the EU HTA PICO) [1-3]. [1] Jeroen P. Jansen et al., "Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1," Value in Health 14, no. 4 (June 2011): 417–28, https://doi.org/10.1016/j.jval.2011.04.002. [2] David C. Hoaglin et al., "Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2," Value in Health 14, no. 4 (June 2011): 429–37, https://doi.org/10.1016/j.jval.2011.01.011.	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			[3] Jeroen P. Jansen et al., "Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report," Value in Health 17, no. 2 (March 2014): 157–73, https://doi.org/10.1016/j.jval.2014.01.004.	
DVSV	16-18	450-486	The chapter on estimands and ICEs is hard to understand on its own, mainly because no examples for ICEs are given.	We will consider if an example is necessary for the next version of the draft.
ISPOR	13-15	5 Subgroup Analyses in Individual Clinical Studies (lines 341-405)	The considerations of subgroup analyses may be different for superiority test and non-inferiority test. For instance, could the document discuss in the context of non-inferiority test, should the margins be the same within each subgroup, or should different margins be applied depending on subgroup characteristics?	The guideline is not intended to be a methodological textbook. Appraisal of the analyses that performed will be left at the discretion of MS.
ISPOR	13-15	5 Subgroup Analyses in Individual Clinical Studies (lines 341-405)	Given the fact that most subgroup analyses do not have sufficient power, in addition to the detailed reports on methods and results from subgroup analyses, it may be useful to present post hoc (or observed) power analyses. Also see and reference the following: Wang et al. Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. *NEJM* 2007; 357:2189-2194.	We do not think post-hoc power analyses are useful since true difference in the population is not known and therefore when H0 is not rejected, an absence of difference still cannot be ruled out.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
ISPOR	16-18	7 Sensitivity Analyses in Individual Studies (lines 445-506)	In addition to sensitivity analyses related to the five attributes of the estimand (i.e., population, treatment, variable (endpoint), intercurrent events and the summary measure), it might be useful to include competing approaches as part of sensitivity analyses. In practice, it is common that multiple models can give undistinguishable goodness-of-fit to the same data set, but different interpretations or conclusions. For example, in evidence synthesis studies, it is common that fixed-effects model can give different conclusions from random-effects model. Reporting results from both approaches. Along with their strengths and limitations, can enhance the summary of evidence generated from systematic reviews and meta-analyses.	This is a technical discussion of statistical modelling that is beyond the scope of the guideline.
Tanja Podkonjak, Takeda	13-14	362-371 375-377	Current text: Pre-specification 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). The first paragraph (line 362-371) dismisses unplanned subgroup analyses as 'lack[ing] statistical robustness' yet the following sentence (line 375-377) states that these may be requested by MS and should be provided. It is unclear what the EUnetHTA recommended methodology and approach is for subgroup analyses which were not prespecified – regardless if they are conducted based on HTD or MS request. The data credibility, and the strengths and limitations of these analyses, regardless whom initiated the request, should be treated in the same way. Credibility describes the extent to which subgroup findings can be concluded as being well substantiated and hence relied on for decision making. Credibility depends on the degree of well-founded, a priori definition, the biological plausibility for a particular finding and replication and not on a policy need for the request.	The distinction between subgroups and subpopulations will be clarified in the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			We request the guidance clarifying the existing conflicting guidance and apply the same principles to subgroups and subpopulations, unless a biologic or clinical rationale for their difference can be provided. In general, Takeda supports different PICO requests to the population being based on biological or clinical rationale only.	
Dr. Thomas Ecker, Ecker + Ecker GmbH	12-13	3.2	In chapter 3, the topics of multiple operations and multiple effect measures are missing. As this is problematic in individual studies it should be mentioned. A possible solution would be a separate subsection analogue to chapter 4.	We agree with this suggestion. A new subsection will be added in the next version of the draft.
GSK	16-17	450-486	Due to the importance of the estimand concept, there should be an extra chapter dedicated to this.	While the estimand concept has some value, it is an extension of the PICO approach (which already addresses the 3 amin attributes of the estimand concept). We do not think it is necessary to detail more the concept in this guideline. The value of the estimand concept in the context of HTA will be clarified in the D4.6 guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
GSK	21-22	585-643	Please add the following references (related to earlier comments): Leverkus 2018a. Leverkus F, Gillhaus J, Knoerzer D, Kupas K, Nicolay C, Hennig M: AMNOG meets EMA - Methodological Areas of Debate from an Industry Point of View (Part 1). Pharmazeutische Medizin 2018, Volume 20, Issue 1, March Leverkus 2018b. Leverkus F, Gillhaus J, Knoerzer D, Kupas K, Nicolay C, Hennig M: AMNOG meets EMA - Methodological Areas of Debate from an Industry Point of View (Part 2). Pharmazeutische Medizin 2018, Volume 20, Issue 2, June Kisser 2021. Kisser, A., Knieriemen, J., Fasan, A. et al. Towards compatibility of EUnetHTA JCA methodology and German HTA: a systematic comparison and recommendations from an industry perspective. Eur J Health Econ 23, 863–878 (2021). https://doi.org/10.1007/s10198-021-01400-2, accessed 2022-06-23	We do not have to add references on demand without justification of their usefulness.
ISPOR	8-9	243-254	It would be helpful to provide further guidance on whether the nominal alpha / alpha spent at interim analysis should be used to analyse all subsequent endpoints / PICOs requested by JCA for the HTA dossier.	This is a technical discussion that is out of the scope of the guideline.
ISPOR	8-9	241, 254, 258	The requirements for reporting are largely related to prespecified statistical planning and transparency - these requirements seem to address statistical reporting rather than multiplicity specifically. There is also very little practical guidance on methods; references in this section would be helpful.	The primary purpose of the guideline is to help assessors to report adequately. The purpose of the guideline is not to be a statistical textbook.
ISPOR	15-16	6 Subgroup Analyses in Evidence Synthesis	In some cases, meta-regression is another way of exploring heterogeneity; some discussion of this approach seems merited. This is particularly important in evidence synthesis using direct or indirect	We do not reject meta-regression but only describe

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
		Studies (lines 406-444)	treatment comparisons. Meta-regression model focuses more attention on the studies with a lower sampling error and it is able to achieve this by assuming a mixed-effects model. MR model accounts for the deviation from the true overall effect due to sampling error and between-study variance or heterogeneity. Also, you can use one or more variables to predict differences in the true effect sizes.	their shortcomings in certain situations.
Mihai Rotaru - EFPIA	p.16	429-440	Paragraph should be deleted, as it already includes a value judgement on the acceptability of evidence, which is subject to the MS decisions on national level.	Already addressed issue.
GSK	8, 10-13, 15-16	Requirements for JCA reporting box	Is the hypothesis test required for analysis using real world data?	The document is about factual reporting of methodological elements pertaining to statistical hypothesis testing based on the evidence of data submitted by HTD irrespective of the type of study.
Mihai Rotaru - EFPIA	5	86-88	Please add also the ICH-E9 addendum (E9(R1)) as reference here – as the estimand concept is part of this document.	We will consider if this addition is necessary.
EFSPI	5	102-5 / 1.1	Current text: "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." Per the EU HTA Regulation, decision making should indeed remain with the Member	Already addresses issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			States. However, we would welcome an attempt to establish a consensus on scientific best practices at the EU level. The Practical Guideline D4.5 should seek to endorse recommendations from the scientific literature, for example on subgroup analysis [1]. Also, we welcome the attempt to adopt the recommendations from the ICH E9 (R1) Addendum and encourage the use of it across all Guidelines where applicable (for example Guideline D4.6).	
Natacha Bolaños, Lymphoma Coalition	5	101	If we consider that health technology assessment focuses specifically on the added value of a health technology in comparison with other new or existing health technologies, and that national health authorities can take evidence-based decisions on the pricing or reimbursement of health technologies, we realise that assessing the clinical added value of a health technology at a national level is not only limited by the methodological aspects, but is mainly limited because of the differences in access. Comparisons cannot be harmonised is access to medicines is not harmonised.	These concerned are out of the scope of the guideline.
François Houÿez (Eurordis)	5	89-94	Different Member States might define different PICO questions due to the national context, and the different questions should be based on scientific and/or medical evidence. However, the HTA cooperation should be aiming at synthesising different PICO questions into one that could satisfy the largest number of Member States (defining the Majority PICO). Other PICO questions (Minority PICOs) should not bee too many or it could delay significantly the moment when the HTD can submit a dossier. The document does not explain whether the applicant could submit first the evidence in relation with one first PICO question (e.g the majority one), and the rest of the dossier at a later stage.	These concerns are related to the D4.2 scoping process guideline.
François Houÿez (Eurordis)	5	95-100	Member States draw conclusions regarding the clinical added value of the health technology assessed. The JCA report concludes on the relative effectiveness and the certainty of results. It would be excellent if the JCA report could include conclusions that could be understood by others than HTA bodies and decision makers. Healthcare professionals, patient organisations and other representatives of the civil society are entitled to understand the purpose of European joint clinical assessments and the view of experts about the utility of the technology in question. Factual assessment of the effectiveness and the certainty of results might not be enough for a larger public to understand the importance of such joint reports.	These concerns are out of the scope of this guideline

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Sebastian Werner vfa	5	88	Please also add the ICH-E9 addendum (E9(R1)) as reference here – as the estimand concept is part of this document.	We will consider if adding this reference as suggested makes sense for the next version of the draft.
BAH	5	102	"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." One of the main goals of EU-HTA is harmonization. Therefore, methodology and its consideration have to be the same for all member states.	Already addressed issue.
Roche	5	102-5 / 1.1	Original text: *"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."* Per the EU HTA Regulation, decision making should indeed remain with the Member States. However, we support the establishment of a consensus on scientific best practices at the EU level. When International consensus by scientific organizations e.g. HTAi, ISPOR, CIRS, DIA are available, they could be considered at the EU level. Consensus should include patient and caregivers' voice when applicable, and those consensus should be preferred as representing all stakeholders. The Practical Guideline D4.5 should seek to endorse recommendations from the scientific literature, for example on subgroup analysis [1]. Also, we welcome the attempt to adopt the recommendations from the ICH E9 (R1) Addendum and encourage the use of it across all Guidelines where applicable (for example Guideline D4.6). [1] Sun et al., "How to Use a Subgroup Analysis: Users' Guide to the Medical Literature," JAMA 311, no. 4 (January 22, 2014): 405–11, https://doi.org/10.1001/jama.2013.285063.	Duplicated comment.
GSK	5	86-88	Please add also the ICH-E9 addendum (E9(R1)) as reference here – as the estimand	Duplicated

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			concept is part of this document.	comment.
ISPOR	5	102	Member states are required to give "due consideration" to JCA reports, but the paragraph goes on to explain that differences may relate to different member state interpretation of consistency / mismatch between HTD research questions and JCA assessment scope PICO questions. This is helpful context, but it would be helpful to more explicitly spell out the implications of such variation, and in what scenarios this might or might not be appropriate. Distinction and overlap between EUnetHTA guidelines is addressed in the final paragraph of the Introduction.	Duplicated comment.
Matias Olsen, EUCOPE	6	158	"Sampling hazard" is confusing as hazard is a technical term used in survival studies. It would be better to use "sampling error"	It is already specified that sampling hazard is a form a random error.
Matias Olsen, EUCOPE	6	166	Add: "which is the probability of the occurrence of a difference under repeated sampling (experiment repeated several times)…"	We think our definition of the p-value is accurate.
Matias Olsen, EUCOPE	6	229-234	The statement is not technically correct. Consider: In Bayesian inference the sampling variability is not taken into account as the current data are considered as fixed, while population parameters and hypotheses concerning them are considered to be random. Thus, the question of type-1 error is not relevant. However, it remains relevant in clinical research to ask what might happen if the same study was conducted again. Type-I error rates in Bayesian designs (studies based on Bayesian inference) can also be evaluated via simulations to account for sampling variability.	The comment is technically not correct. In Bayesian inference the sampling variability is taken into account as the likelihood is the distribution of the study data (considered as random) conditional on the parameters (considered as

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				fixed). The type 1 error is also relevant if you use Bayesian inference to test a frequentist hypothesis, e.g., by assessing whether the zero effect is included in the credible interval. Therefore, no change is required.
Mihai Rotaru - EFPIA	6	142-149	Please add the definition for "evidence synthesis studies" in this section.	It will be added in the next version of the draft.
Mihai Rotaru - EFPIA	6	169	Current wording: if the p value is less than the α level, the alternative hypothesis is accepted Proposed wording: if the p value is less than the α level, the null hypothesis is rejected Rationale: A clinical trial is designed to reject or not reject a null hypothesis. Rejecting a null hypothesis does not mean we should accept the alternative hypothesis.	It will be corrected in the next version of the draft.
Marko Ocokoljic (SIOPE)	6	132, 133	"Therefore, while recommendations in this guideline may be better suited for RCTs, they can apply to various study designs like single arm studies (which are highly	This precision is vague,

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			beneficial in rare and ultra rare disease settings and paediatrics)."	speculative and unnecessary in the context of this guideline.
EFSPI	6	158-159	Actual wording: "Thus, the risk of wrongly claiming the existence of treatment effectiveness needs to be controlled at an acceptable level which is achived via statistical hypothesis testing." Suggest: "The risk of wrongly claiming the existence of treatment effectiveness can be controlled at an acceptable level via statistical hypothesis testing".	We do not think the suggested change adds value to the statement.
EFSPI	6	165	Current wording: "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." Proposed wording: "Statistical hypothesis testing in the spirit of Neyman-Pearson relies on comparing a test statistic computed from the data to a critical value derived from the distribution of the test statistic under the null hypothesis. The critical value is determined such that the probability to reject the null hypothesis although it is actually true in the population is bounded by a pre-specified number, the significance level alpha. This is often set to 0.05. If the test statistic exceeds the critical value the null hypothesis is rejected. This decision process of comparing test statistic to critical value can be translated in comparing the p-value of a test (defined as the probability of the occurrence of a difference at least as large as the one observed if the null hypothesis is true) to the significance level. But a priori p-values have nothing to do with hypothesis testing within the Neyman-Pearson framework. In Fisher's sense, p-values are used for a quantification of evidence against a null hypothesis."	While we technically agree with this comment, in practice, in clinical research, statistical hypothesis testing in most cases is interpreted and reported by comparing the p-value to the alpha level, and a binary decision is taken. Estimates of test statistics are rarely reported. We think this technical

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				distinction could lead to confusion for assessors, because it does not match the usual practices in terms of reporting.
Sebastian Werner vfa	6 14	102-104 493 - 494	*"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."* *"The acceptability of missing data is subject to member state differences in interpretation of their relevance within their respective decision-making process."* The guidance does not include an explanation for why <u>different methodological approaches</u> in the member states for handling multiplicity, subgroup, sensitivity, post hoc analyses, or missing data are justified. As the European HTA Regulation requires the development of methodologies following international standards of evidence-based medicine, it is unclear why available standards are not considered to achieve a harmonized method for handling multiplicity, subgroup, sensitivity, post hoc analyses, or missing data for joint clinical assessment. One of the main goals of the European HTA Regulations was to harmonize the clinical assessment by convergence of HTA methodology.	Duplicated comment.
Sebastian Werner vfa	6	142	Please add the definition for "evidence synthesis studies" in this section.	Duplicated comment.
Sebastian Werner vfa	6	143-145 146-148 532-538	*"In the context of this document, the terms "planned" and "<u>prespecified</u>" 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 "<u>post hoc analysis</u>" 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*	Already addressed issue.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			a HTD" "*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."* Pre-specification or planning of analyses increases the credibility of results. These results are not influenced by the investigators knowledge of the data. Post hoc analyses conducted by the investigators are usually referred to as "Data driven analysis" and are usually less credible. However, in the context of HTA these relationships are not straight forward, as post hoc analyses are usually requested by the HTA authorities. With the definition of the PICO that might not fully match the characteristics of the clinical study, investigators must re-analyse the study results according to these requirements. As HTA authorities do not know the data, it cannot be easily assumed that these post hoc analyses are "Data driven" and less credible. Indeed, for a HTA assessment, these PICO requirements might be seen as a "prespecified" hypothesis to be tested in that assessment. These "prespecified" analyses might reach comparable credibility as prespecified analyses in the statistical analysis plan. The vfa recommends elaborating on the definitions of "prespecification" and "post hoc" to address differences between Regulatory and Health technology assessment. The specific context of HTA should be considered. Analyses requested by the HTA authorities regarding specific PICO questions, should not be designated as "post hoc" but rather be considered as "prespecified" hypothesis to be tested in the systematic review (i.e., joint clinical assessment).	
Roche	6	165-9 / 3.1	Original text: *"Statistical hypothesis testing relies on estimating the p value, which is the probability*	Already addressed

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			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." A priori, hypothesis testing has nothing to do with p-values. p-values can be used to make the decision in a hypothesis test, but their initial (quantification of evidence against a null hypothesis in Fisher's sense) was completely unrelated to hypothesis testing. In the interest of educating also non-statistical stakeholders we recommend to reformulate this paragraph as follows: Proposed new text: *"Statistical hypothesis testing in the spirit of Neyman-Pearson relies on comparing a test statistic computed from the data to a critical value derived from the distribution of the test statistic under the null hypothesis. The critical value is determined such that the probability to reject the null hypothesis although it is actually true in the population is bounded by a pre-specified number, the significance level alpha. This is often set to 0.05. If the test statistic exceeds the critical value the null hypothesis is rejected.* *This decision process of comparing test statistic to critical value can be translated in comparing the p-value of a test (defined as the probability of the occurrence of a difference at least as large as the one observed if the null hypothesis is true) to the significance level, but a priori p-values have nothing to do with hypothesis testing within the Neyman-Pearson framework."*	issue.
GSK	6 19	146-148 Section 9.1	"Post-hoc" refers to any analysis that is specified after the data of a clinical trial have been observed (in contrast to the confirmatory setting in which "post-hoc" is often used for any analysis that has not been specified in the Statistical Analysis Plan (SAP) of the clinical trial). Pre-specification, therefore, may not only refer to analyses specified in the protocol	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			and/or SAP of a trial, but also to any analysis that is requested by the HTA bodies as part of the standard complementary analyses as well as any statistical analysis that a sponsor may specify in a separate SAP, i.e., HTA SAP, before conducting the analysis in order to meet a HTA body's request. The clinical trial SAP that is developed for the regulatory process is considered to be the foundation, especially with regards to the operationalization of endpoints. However, due to deviating research questions complementary statistical analyses may often be required in the HTA context. In this context, such analyses are not to be considered "post-hoc" as they are defined by the scope of the HTA. Additional complementary analyses may also serve to strengthen the robustness and consistency of the data, even if not pre-specified.	
GSK	6	142-149	Please add the definition for "evidence synthesis studies" in this section.	Duplicated comment.
ISPOR	6	130	On line 130 it is stated that this guideline predominantly deals with methodological issues related to inferential statistical analyses. This is a helpful clarification, and could be reflected earlier (or even in the title of the guidance document).	We will consider if it needs to be specified elsewhere for the next version of the draft.
ISPOR	6	133-135	This document uses "effectiveness" as a common term to describe efficacy, effectiveness and safety (Line 133-135). Given that dossiers may include evidence from both trial and non-interventional, real-world observational study designs, where "effectiveness" is usually considered to mean treatment outcomes in real world, usual care conditions. In addition, sometimes "effectiveness" has been considered a more comprehensive term meaning something more along the lines of "net benefit". That said, we don't have a better general term to suggest than "effectiveness" here – terms like "outcome" or "endpoint" don't work any better in this document - so these considerations may simply bear more explanation here. We can provide references if desired.	While we acknowledge there may be debated about the best terms to use regarding "effectiveness" or the potential difference between outcomes and endpoints, we do not think these

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				conceptual discussions will hamper the how the guideline will be interpreted.
ISPOR	6	133-135	In addition to the definitional point on "effectiveness" above, some of the issues and considerations should be discussed separately with respect to efficacy/effectiveness and safety endpoints. For instance, when a study targets efficacy endpoint, the risk of false positives should be limited, so multiplicity adjustment is necessary when multiple testing to control for type I error rate. However, if a study targets safety endpoint, the economic cost of false negatives could be more significant compared to false positives. Therefore, we want to limit the risk of false negatives to ensure adequate study power to detect safety risks, and we generally are not worried about multiplicity controls for safety endpoint.	Already addressed issue.
ISPOR	6	169	The statement "the alternative hypothesis is accepted" is technically incorrect. Generally in hypothesis testing, if the p-value is less than the alpha level, we have strong evidence to believe that the null hypothesis is not true, but it doesn't necessarily mean the evidence is strong enough to believe that the alternative hypothesis is true. Therefore, the statement here should be "the null hypothesis is rejected".	Already addressed issue.
DVSV	7	175-183	As far as I know, the FWER is always defined as the probability of at least one false positive, regardless of how many of the tested null hypothesis are in fact true. However, the FWER can be controlled in two ways (in the strong or the weak sense). Not sure if it is necessary or helpful to introduce - at least for me - untypical defintions of FWER here ("global FWER" and "multiple level FWER").	Various members of the HOG think this distinction has value.
DVSV	7	185-196	"Thus, for the rest of the document, we use the term FWER to mean "FWER in a strong sense"." The FWER can be controlled in the strong sense, but I am not familiar with the term "FWER in a strong sense". Consider rewording, e.g.: "Thus, for the rest of the document, we use the term FWER control to mean "FWER control in a strong sense"."	We do not think adding "control" leads to a better comprehension of the concept at hand here.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Matias Olsen, EUCOPE	7	187	While these tests are appropriate for individual clinical studies, e.g. RCTs, it is not clear how they could (or should) all apply to evidence synthesis, e.g. NMAs.	This subsection is precisely about original clinical studies. Evidence synthesis is dealt with in separate sections.
Matias Olsen, EUCOPE	7	195	Would help to include clarification of how this would be applied to evidence synthesis studies.	This subsection is precisely about original clinical studies. Evidence synthesis is dealt with in separate sections.
Mihai Rotaru - EFPIA	7	215	Current wording: rejecting or accepting the null hypotheses Proposed wording: Rejecting or not rejecting the null hypothesis Rationale: A clinical trial is designed to reject or not reject a null hypothesis. Not rejecting a null hypothesis does not mean we should accept the null hypothesis.	Already addressed issue.
EFSPI	7	202-205	Treatment effects based on early stopping for efficacy are in theory biased, but this bias is in all reasonable scenarios negligible (see e.g. https://journals.sagepub.com/doi/10.1177/1740774509102310 or https://pubmed.ncbi.nlm.nih.gov/27271682).	We only point potential shortcomings while not judging if these shortcomings are relevant in

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				a particular context.
MTE	7	173	"The probability a of a type 1 error for one significant test is the comparison wise error rate (CER) [3]." Suggest changing "significant test" to "statistical test" or "hypothesis test".	We think this statement is correct as the type I error is the error one can make when declaring a test is significant.
Roche	7	196 / 3.1	*"formal completion of trial"*. We recommend to phrase this more precisely, as a trial typically has a "preplanned final analysis" (where the prespecified number of events, e.g., is reached), but the trial will continue to be run for years to come until its "completion" (where "completion is understood as stopping any data collection).	It will be clarified in the next version of the draft.
Roche	7	203 / 3.1	Treatment effects based on early stopping for efficacy are in theory biased, but this bias is in all reasonable scenarios negligible (see e.g. https://journals.sagepub.com/doi/10.1177/1740774509102310 or https://pubmed.ncbi.nlm.nih.gov/27271682). We therefore invite EU HTA to reformulate this paragraph accordingly.	Duplicated comment.
GSK	7	175-186	FWER is preferred in clinical trials as it keeps falsely rejected hypotheses low with few significant outcomes. As genomic and exploratory data is also increasingly considered, we could also look at the false discovery rate (FDR) which is the expected proportion of wrongly rejected hypotheses, and as such allow for a few incorrectly rejected hypotheses whilst also having more true significant hypotheses.	We agree with this comment, but the evidence that assessors will have to look at will almost certainly only be related to control of the FWER. Therefore, we do not think introducing the FDR is currently relevant in the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				context of this guideline.
GSK	7	190-191	Should 'same consequence is assessed at different time points' be included in the interim analysis section?	This sentence deals with the idea of testing the same consequence at different time points as part of the final and main analysis and not as part of a strategy of interim analyses.
ISPOR	7	3 Multiple Statistical Hypothesis Testing in Individual Clinical Studies (lines 150-258)	Footnote: For more complex situations, it's worth noting that the use of simulation studies to explore the type I error control is an overall good strategy to help better understand the multiplicity adjustment issue.	We do not think this technical comment is needed for this practical guideline.
Matias Olsen, EUCOPE	8	241-242	Are the null and alternative hypotheses, as well as the type 1 error rate, required for Bayesian analysis?	It depends if Bayesian Inference was used to test a frequentist hypothesis or not.
Matias Olsen, EUCOPE	8	242-254	Clarification on applicability/examples for evidence synthesis studies would be helpful for section 3.2.2	This chapter is about original clinical studies only. Evidence

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				synthesis are dealt within separate sections.
EFSPI	8	229-230	Actual wording: "RCTs are designed to answer a specific research question in a binary manner (i.e., concluding if there is a true effect or not) […]". Proposed wording: "RCTs are designed to obtain the evidentiary support for a research hypothesis in a binary manner (i.e. concluding if there is evidence in favor of a true effect or not)"	We do not think this conceptual distinction, while technically more correct, will add value in the context of this practical guideline.
MTE	8	241-242 BOX	The JCA reporting requirements presented in the BOX are related to pre-specified statistical planning and transparency reporting rather than multiplicity specifically. Consider adding specific multiplicity considerations that speak to how to handle multiple testing issues.	This guideline is aimed at assessors for allowing adequate reporting. It is not intended to be a statistical textbook.
MTE	8	242	Propose to separate our interim analysis done by DSMB (data safety monitoring board) for study safety to stop prematurely or to extend trial period to respond to the safety question and trial designed interim analysis on effectiveness (seeking to prematurely conclude the study outcomes) and have a separate reporting the JCA.	Methodological concerns apply for both situations. There is no need to separate them.
MTE	8	3.2.1	Requirements for JCA reporting box: sample size/statistical power should also be evaluated	Estimation of the required sample size for ensuring a minimal amount

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				of power is useful for planning a study but its assessment does not add value after the study was performed. If the null hypothesis is rejected, the possible error is the type-1-error. If not, it's impossible to know if it's because of lack of power or true absence of difference in the population.
Roche	8	236-41 / 3.2.1	Original text: *"Prospective specification of all data analyses …"* This implies that all analyses required for HTA purposes (additional domains) would need to be defined a-priori. However, due to complexity and variety of member state requests, there may be a need for further analyses to answer the intended HTA requests. In the case that additional analyses are required to answer the HTA question (i.e., outside of the pre-specified analyses) all non-pre-specified analyses should be restricted to a minimum set of clearly defined research questions. Proposal:	This will be clarified for the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Please reformulate the paragraph to reflect the extension of analyses to non-pre-specified variables within reason of clearly defined research questions defined by the pan-EU PICOs.	
Silke Walleser Autiero Medtronic	8	241-242 BOX	The JCA reporting requirements presented in the BOX are related to pre-specified statistical planning and transparency reporting rather than multiplicity specifically. Consider adding specific multiplicity considerations that speak to how to handle multiple testing issues.	Requirements are meant to be elements that needs to be factually reported by assessors for allowing MS to draw their conclusions at national level. These requirements are not meant to be instructions about how analyses should be done.
Elaine Stamp PHMR	8	Summary box	Reporting should be done according to PICOT. Would expect a full description of population (including baseline characteristics and inclusion /exclusion criteria), and a description of the intervention. Also, should include a sample size calculation and a description of how missing data is to be dealt with.	These concerns are out of the scope of the guideline.
ISPOR	8	235	One of the requirements for JCA reporting is "Accurate and unambiguous endpoint definitions". Is this a good opportunity to align with regulatory language and introduce the term "estimands" into HTA reporting?	Already addressed issue.
ISPOR	8	235	Requirements for appropriate reporting of methods and results in a JCA are concise and straightforward, which can facilitate assessor/co-assessor's review. However, it is important to separate primary/secondary endpoints rather than combine into multiple endpoints. Moreover, statistical methods for dealing with multiplicity for primary	JCA are expected to answer to PICO questions from

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			endpoints, and secondary endpoints if applicable should be reported in a JCA	the HTA perspective. MS can request the outcomes they see fit without ranking them. Therefore, we do think the distinctions between primary and secondary are relevant in the context of JCA.
ISPOR	8	3.2.1 / 3.2.2.	The multiplicity concept seems to be restricted to the context of a single RCT. The guidance should consider other situations where non RCT are needed for the analyses needed for the PICOs for JCA. More broadly, the document should provide some guidance on the PICOs and CER level needed to test across all analyses requested by the JCA. The list of PICOs will determine the testing strategy, which may also vary across the different requests/PICOs from the MS. Should we rather apply a more targeted approach for each single MS (but across PICOs), or have a general approach for each single PICO? This may also require clear directions from JCA on the ordering and importance of each of the PICOs.	PICO won't be ranked or ordered, and have to be considered with the same importance. Only definitive PICO will be communicated to HTD, but without any details on MS' individual request.
DVSV	9	255-257	Chapter on more than two treatment groups seems to be missing.	Requirements are proposed. We do not think this subsection needs more content.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Matias Olsen, EUCOPE	9	254-255	Add: "…spending designs and their boundaries, and the desired FEWR level. For adaptive designs, which method is chosen for multiplicity control, for example, promising zone design, use of the conditional error-rate function, inverse normal combination of stage-wise p-values or other weighted methods.".	We do not think adding these methodological details are necessary in the context of this practical guideline.
Paolo Morgese – ARM	9	253-255	The interpretation of the durability of effect and it is associated uncertainty is the main consideration for HTAs of ATMPs. It is important that across JCAs that there is a consistent position on the timing of the data cut used for the analysis. Experience with the Beneluxa and Finose joint assessments and individual HTA country assessments for ATMPs is that the timing of the data cuts for the clinical analysis has been highly variable. Some consortia/agencies have used the data-cut which has underpinned the EPAR whereas other consortia/agencies have requested data right up to the point of the clinical-effectiveness assessment for the HTA. With durability of effect being the main driver and the requirement for extrapolation to the lifetime horizon for HTA, the reporting of the data cut-off should be consistent across JCAs. The data-cut for the study report may not be the same as for the EMA submission nor the follow interactions e.g. D120 when a more recent data-cut might be required	These concerns are out of the scope of this guideline.
MTE	9	255	Could also add to the title "comparative studies"	We do not think this addition is useful.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
MTE	9	3.2.3	When there are more than two treatment groups, it is important to specify what the comparisons are and this is not mentioned. For example, for groups A/B/C there are 3 potential comparisons, will all 3 comparisons be performed? Also, it is possible when there are multiple comparisons, only one comparison is confirmatory and others are exploratory.	We think the requirements for reporting we are proposed will allow MS to draw the conclusions they need in response to their PICO questions.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	9	253	Original wording: "Results for interim analysis and the decisions regarding clinical study continuation should be reported." Comment: Interims analyses may be performed for a number of different reasons. Thus, an interim analysis does not necessarily add new information compared to other data cutoffs. This may be the case, for instance, if follow-up data collection for a certain endpoint has already been almost completed at the time of the previous data cutoff or if the time period between two data cutoffs is short. Therefore, interims analyses should only be reported if they add a significant amount of information relevant for the benefit assessment of a new treatment [1]. Suggestion for rewording: "*Results for relevant interim analyses […] should be reported.*"	'Significant amount of information relevant for the benefit assessment' interpretation is left at MS assessment. MS could have different requirements for 'significant'. Every interim analysis should be reported.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ina S. L. Buchholz SKC Beratungsgesellschaft mbH	9	Section 3.2.2, Box "Requirements for JCA reporting"	Original wording: ○ […] ○ "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)."	We disagree with this comment, and we will keep the original wording.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Comment: Although it is worthwhile to briefly outline DMC's involvement in the study, a detailed description of the specific recommendations is superfluous in the dossier. Recommendations of DMC are sufficiently documented in the study documents. The same applies to modifications of the study protocol which are highlighted in the study protocol amendments. Further, these levels of detail are not necessary to assess the efficacy and safety of a new treatment and should therefore not be obligatory in the dossier.	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ina S. L. Buchholz SKC Beratungsgesellschaft mbH	9	Section 3.2.2, Box "Requirements for JCA reporting"	Original wording: o "How the endpoints were tested (statistical methods), including the method chosen for controlling for multiplicity, [...]" Comment: As stated by the EUnetHTA authors, correction for multiplicity often causes problems. Therefore, corresponding information on the statistical procedures should only be expected in the dossier provided that multiplicity adjustment has been performed. The formulation "if performed" was also chosen in D4.5 by the EUnetHTA authors in several other sections in a very similar context. Similarly, the German IQWiG state in their method paper: "*When appropriate and possible, the Institute applies methods of adjustment for multiple testing* [2]*."* Suggestion for rewording: "*How the endpoints were tested (statistical methods), including, if performed, the method chosen for controlling for multiplicity, [...]*"	We will modify the draft accordingly.
Sebastian Werner vfa	9	Requirements for JCA reporting	What is meant by 'respective cutoff date, with corresponding follow-up'? What kind of follow up to a cutoff should be reported?	We will clarify that it is data cutoff date.
Sebastian Werner vfa	9	Requirements for JCA reporting	Please change 'sponsor or regulatory body.' to 'e.g. sponsor, regulatory body or HTA body'.	We will modify the draft accordingly
BAH	9	254 (box)	"When unplanned interim analyses were conducted, why they were deemed necessary and by whom (sponsor or regulatory body)."	We will modify the draft accordingly

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			In Consequence interim analyses could not be called from HTA bodies?	
Silke Walleser Autiero Medtronic	9	255	Could also add to the title "comparative studies"	Duplicated comment.
Elaine Stamp PHMR	9	Line 254	Should mention that results for interim analysis are usually considered by an independent data monitoring committee so that key trial personnel remain blinded, and bias is not introduced. It is mentioned in the summary box	This is out of scope of this guideline.
Elaine Stamp PHMR	9	Line 255	I think this section could benefit from further detail such as: • Rationale for using a multi-arm design • Specification of the research question referring to all of the treatment groups. • Hypotheses to be tested with clear statement of primary comparisons • Description of trial design – many options available for multi-arm (parallel, factorial, cross-over) Sample size calculation	This is out of scope of this guideline.
EFSPI	10f	Section 4	Evidence synthesis methods are important tools to provide relative effect estimates in the absence of direct, head-to-head studies. They achieve this, for example, through the analysis of published RCT data. As such, evidence synthesis is typically exploratory in nature and the formal confirmatory testing framework seems not applicable to evidence synthesis studies (the draft Guideline D4.5 acknowledges this in lines 269/70). However, the reporting requirements in Section 4.2 are still heavily centred around testing. We recommend to remove these requirements and to focus, instead, on estimation and established best practices, such as pre-specification in the scoping process (leading to the EU HTA PICO) [1-3]. [1] Jeroen P. Jansen et al., "Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1," Value in Health 14, no. 4 (June 2011): 417–28, https://doi.org/10.1016/j.jval.2011.04.002.	This guideline is about how to report elements pertaining to methodological issues such as multiple hypothesis testing. These concerns are out of the scope of this guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			[2] David C. Hoaglin et al., "Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2," Value in Health 14, no. 4 (June 2011): 429–37, https://doi.org/10.1016/j.jval.2011.01.011. [3] Jeroen P. Jansen et al., "Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report," Value in Health 17, no. 2 (March 2014): 157–73, https://doi.org/10.1016/j.jval.2014.01.004.	
MTE	10	4.1	"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." – suggest adding that because of this, the emphasis of meta-analysis should generally be on estimating intervention effects rather than testing for them. (ref: https://handbook-5-1.cochrane.org/chapter_16/16_7_2_multiplicity_in_systematic_reviews.htm)	We think our statement is enough.
Storz-Pfennig/ Ermisch – GKV-SV	10	Sect. 4.2.1	Evidence synthesis/meta-analysis will be provided by the HTD in the dossier according to specified PICO-questions (e.g. D5.1 Submission Dossier Guidance). They will not be taken from the published literature; thus, consideration of "heterogeneous interests" is not relevant. Likewise, although it might "almost never [be] the case" that individual patient data is available for published meta-analyse/evidence synthesis based on clinical trial publications, this is different in den current context. In particular, when trials conducted by/on behalf of the HTD form the evidence base for the evidence synthesis in the dossier, the HTD will be able to and can be expected to provide such analyses, if it is relevant for the PICO questions in terms of outcomes, patient groups etc.	These elements are the scope of the D4.3.1 and D4.3.2 guideline. We will add references to these guidelines in the text.
Matias Olsen, EUCOPE	11	283	Replace: "The CER level for each statistical test"	This section is not multi-arm trials.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			With: "The CER level for each comparison to be tested in a multi -arm trial". Add: "For trials using Bayesian inference, the operating characteristics (type-I error rate and power) must be established via extensive simulations. The success criterion based on the posterior probability is typically chosen via simulations to ensure that the type-I error is preserved at the nominal rate. The simulation report should be provided as an appendix to the SAP.".	We do not think that currently the guideline should precise more technical elements regarding Bayesian inference.
Tanja Podkonjak, Takeda	11	294-298	Current text: Methods of multiple hypothesis testing in NMA is still under debate however, in the context of a JCA scope the relevant comparator(s) are defined. Therefore, the multiplicity due to multiple groups is not the main problem if the relevant comparison is new intervention vs one control. The current guidance does not address the likely scenario where there is more than one comparator identified through the scoping process resulting in multiple PICOS. Many technologies will have direct evidence vs one control (i.e., one comparator) but are unlikely to have direct evidence vs multiple comparators. Further guidance is needed on multiple hypothesis testis in NMA in a situation where there are multiple comparators but not direct evidence vs the intervention for all.	Each comparator will represent a PICO on its own. Thus, multiple comparators should not be an expected issue.
MTE	11	283-284 BOX	The first dot point is unclear and needs rewording.	It will be clarified for the next version of the draft.
MTE	11	289-293	This sentence makes the recommendation unclear. More explanation of what is a matter of debate vs what is required in the JCA would be helpful.	The requirements proposed in the box below are explicit.
Storz-Pfennig/ Ermisch – GKV-SV	11	Sect. 4.2.2	We support the notion that multiplicity issues will not be of grave concern if NMA are used only to assess one intervention to a control. However (s. comment on sect.	We agree with this comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			4.2.1), if the evidence synthesis is created by the HTD for the purpose of including it in the dossier to answer the PICO questions, comparators/treatment arms not relevant to the PICO questions might not be needed.	
Silke Walleser Autiero Medtronic	11	283-284 BOX	The first dot point is unclear and needs rewording.	Duplicated comment.
Silke Walleser Autiero Medtronic	11	289-293	This sentence makes the recommendation unclear. More explanation of what is a matter of debate vs what is required in the JCA would be helpful.	Duplicated comment.
ISPOR	11	299	This section (4.2.3) and the next (4.2.4) provide examples with limited discussion of the concepts generally. Additional guidance/references as to how to handle these issues from a conceptual perspective would be helpful.	The guideline is primarily intended for helping assessors to perform an adequate reporting. It is not meant to be a methodological textbook.
Matias Olsen, EUCOPE	12	315	It would be helpful if the specific statistical considerations were included in the D.4.3 guidelines to clarify which apply to indirect comparisons, as some of these methods are more appropriate for clinical studies, e.g. RCTs	D4.3.2 guideline is already published.
Matias Olsen, EUCOPE	12	338-340	If Member States require different operationalizations of an endpoint, how will that information be communicated to sponsors preparing the JCA report? Will it be part of the PICO report?	These concerns are tackled within the D4.2 and D4.4 guidelines.
Mihai Rotaru - EFPIA	12	309-313/4.2.3	Multiple time points Current wording: If individual patient data are available, methods for dealing with multiple time points	References to the D4.3.1 and D4.3.2 will be added.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			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. Proposed wording: If individual patient data is available, or pseudo IPD can be generated from published Kaplan-Meier survival functions (Guyot 2016), evidence synthesis methods for dealing with survival data generated at multiple time could be considered. For example, fractional polynomials can be used to produce difference in restricted mean survival time for a selected time point. Rationale: IPD for all RCTs in a NMA of survival data is seldom available. It is accepted practice to generate pseudo IPD from published KM plots to overcome this data limitation. Reference: Guyot P, Ades A, Ouwens M, Enhancing secondary analysis of survival data: reconstructing the data from published Kaplan Meier survival curves. BMS Med Res Meth 2012; 12:9	
EFSPI	12	334-335	Current wording: [...] 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." To make this operational, the effect measure of interest needs to be part of the scope requested by member states. It is not merely a technical issue for the HTD to choose an operationalization - different operationalizations generally correspond to different policy questions. To overcome multiplicity issues, the scope of the JCA should be based on a very limited set of PICOs and well-defined analyses that are necessary to support the HTA processes at member state level.	These concerns are tackled within the D4.2 and D4.4 guidelines.
EFSPI	12	309-313/4.2.3	Current wording: 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	Duplicated comments.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			survival time for a selected time point. Proposed wording: If individual patient data is available, or pseudo IPD can be generated from published Kaplan-Meier survival function (Guyot 2016), evidence synthesis methods for dealing with survival data generated at multiple time could be considered. For example, fractional polynomials can be used to produce difference in restricted mean survival time for a selected time point. Rationale: IPD for all RCTs in a NMA of survival data is seldom available. It is accepted practice to generate pseudo IPD from published KM plots to overcome this data limitation. Reference: Guyot P, Ades A, Ouwens M, Enhancing secondary analysis of survival data: reconstructing the data from published Kaplan Meier survival curves. BMS Med Res Meth 2012; 12:9	
MTE	12	307	"A different solution to the problem of different time points is to use a summary effect measure over time" – suggest adding this is only possible when the individual patient data are available. Should be noted that it is often not possible to have access to the individual patient data.	References to the D4.3.1 and D4.3.2 guidelines will be added.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	12	333	Original wording: "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." Comment: According to HTA Regulation (EU) 2021/2282 Article 25 "The assessment scope for joint clinical assessments should be inclusive and should reflect all Member States' needs in terms of data and analyses to be submitted by the health technology developer." Hence, the final assessment scope provided to the HTD shall enable the submission of a dossier fully meeting the needs of every member state. Consequently,	These concerns are tackled within the D4.2 and D4.4 guidelines.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			there should only be the need for one effect measure per outcome commonly accepted by all member states. A situation in which different member states may require different effect measures must be avoided.	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	12	338	Original wording: "[...]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." Comment: As mentioned above, according to HTA Regulation (EU) 2021/2282 Article 25 "The assessment scope for joint clinical assessments should be inclusive and should reflect all Member States' needs in terms of data and analyses to be submitted by the health technology developer." Hence, the final assessment scope provided to the HTD shall enable the submission of a dossier fully meeting the needs of every member state. Consequently, equal requirements for the operationalisation of certain endpoints should be established among all member states. A situation in which different member states may require different operationalisations for a certain endpoint must be avoided.	These concerns are tackled within the D4.2 and D4.4 guidelines.
Sebastian Werner vfa	12	304	*"A solution to limit the issue of multiple time points can be to choose a single time point for the analysis."* This might be contradicting to different PICO/requirements of member states.	There is no contradiction. MS picks the endpoints for their appraisal.
Sebastian Werner vfa	12	333-335 337-339 119-120	*"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."* *"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."* *"Thus, all the requirements for assessment and reporting mentioned in this guideline assume that HTDs present the necessary elements in the submission dossier."* The guideline states that several methodological approaches will be used by the	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			member states in dealing effect measures and operationalisations for an endpoint. The guideline is also clear that the submission dossier must comprise all the necessary elements for these assessments of the member states. Thus, multiple methodological approaches for data analysis must be addressed in the submission dossiers and joint clinical assessments. Multiple methodological approaches for data analysis constitute duplication. Further, convergence of HTA methodology for a harmonized European clinical assessment framework is not achieved. Multiple methodological approaches pose risks to a workable and efficient European HTA System. The vfa recommends forming clear recommendations on effect measures and operationalisations of endpoints. The vfa recommends establishing a harmonized European methodological framework for joint clinical assessment that provides a uniform methodological approach for data analysis and synthesis that member states should commonly accept. The European methodological framework should not be built as a collection of multiple methodological approaches of different member states.	
BAH	12	304	"A solution to limit the issue of multiple time points can be to choose a single time point for the analysis." Under certain circumstances, different PICOs lead to different analysis time points. (As mentioned in your guideline in line 334, related to "(…) multiple effect measures for an outcome need to be reported in the JCA.")	Duplicated comment.
Matias Olsen, EUCOPE	13	343	Add: "Patients may respond differently to treatments because of demographic factors and baseline factors such as predictive or prognostic biomarkers, disease characteristics,"	This addition does not add value to this subsection.
Mihai Rotaru - EFPIA	13	Line 342f, Section 5.1	Current wording: 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	Subpopulation are not analysed as separate sub-PICOs, but as separate PICOs (see 4.2

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			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. Proposed wording: 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. Subpopulations, defined as a subset of the patient population targeted as described in the therapeutic indication, are subgroups which may be specified during the assessment scope (see EUnetHTA Practical Guideline D4.2.1 Scoping process) and are analysed as separate Sub-PICOs. All concepts for analysing and interpreting subgroup analyses are also applicable to subpopulations. Rationale: JCA methodology should be consistently applied. A priori, the effect estimate from the overall study population is the best estimate also within each subgroup/subpopulation. Accordingly, priority should be given to higher certainty results from overall study population unless major heterogeneity can plausibly be demonstrated based on established best practice. If a subpopulation needs to be analysed for the scope of the JCA, this should be reflected as a "Sub-PICO" of the PICO on the overall study population, rather than a separate research question. This should ensure that effects in subsets vs full population are analysed, evaluated and interpreted in conjunction.	'Scoping process'). The assessment scope, which is the starting point and the basis for the JCA, is not data driven (see 4.2 'Scoping process') and is based on MS needs. Accordingly, there is no rational to define a priori some results to be 'the best estimate'.
Tanja Podkonjak, Takeda	13	354	Current wording: 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.	We define here what is an interaction as a statistical concept, whether this interaction is considered

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Proposed wording: Variables that represent methodological characteristics of a study are not regarded as potential effect modifiers and therefore their potential impact on estimating treatment effectiveness "may" 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, a "statistically significant" effect modification is referred to as an interaction between a treatment and the relevant variable. Rationale: The term "evident effect modification" leaves room for wide interpretation. To assess interaction, it is standard to either conduct a statistical test of the terms in question or plot the data for examination of the pattern characteristics of interaction when sample size is in question. (1).Applied Regression Analysis and other Multivariable Methods. Kleinbaum, Kupper and Morgenstern (c) 1987 pg 164.	significant or not after statistical hypothesis testing.
Tanja Podkonjak, Takeda	13	347-353	Current text: 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. The current definition and distinction between a subgroup and a subpopulation is unclear. Could the guidance document and authors further elaborate what is the biological or clinical distinction between a subgroup and a subpopulation? We note that throughout the guidance document different recommendations regarding analyses (pre-specified, post hoc) and how the validity of the results are made for subgroups and subpopulations.	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Unless there is a clear biological or clinical rationale for a subgroup vs a subpopulation, it is unclear what the justification is to apply different principles to the two. Based on the current definitions, subgroup and subpopulation should be treated the same and the same rules, handling of data, analyses and their interpretation applied to both.	
EFSPI	13	Line 342, Section 5.1	Current wording: "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." Proposed wording: 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. Subpopulations, defined as a subset of the patient population targeted as described in the therapeutic indication, are subgroups which may be specified during the assessment scope (see EUnetHTA Practical Guideline D4.2.1 Scoping process) and are analysed as separate PICOs. All concepts for analyzing and interpreting subgroup analyses are also applicable to subpopulations. Rationale: JCA methodology should be consistently applied. Subpopulations of interest may be specified during the assessment scope as separate research questions. A priori, the effect estimate from the overall study population is the best estimate within each subgroup/subpopulation. The draft guidance on validity of clinical trials states that "there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			the PICO question can be sufficiently answered on the basis of higher-certainty results." Accordingly, priority should be given to higher certainty results from overall study population unless major heterogeneity can plausibly be demonstrated based on established best practice. If the research questions cannot be answered with the results on the ITT population, the subpopulations should be analysed and interpreted in consistence with principles and best practice for subgroup analyses (see comment on page 14, lines 378f).	
EFSPI	13	395-396	Current wording: "[…] it cannot be ruled out that any differences detected between subgroups are caused by systematic errors such as confounding". In a randomized study, such errors will by definition not be systematic, but at most be an example of so-called chance confounding, not to be confused with structural confounding (doi: 10.1097/EDE.0000000000000564). Therefore we propose to reword accordingly, and better talk about imbalances instead of confounding, as this wording is strongly linked to structural confounding.	We will modify the text accordingly.
MTE	13	364-367	This statement is too strong to be considered accurate. While they may be less robust than pre-specified analyses, they do not lack all robustness. Indeed, multiple testing corrections could be applied to the set analyses that were not pre-specified.	It will be clarified for the next version of the draft.
Dr. Thomas Ecker, Ecker + Ecker GmbH	13	362-367	**Statement in the guideline:** "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." **Comment:** We agree that evidence of unplanned subgroup analyses is limited. However, the text	MS may choose their subgroup analyses, therefore no criteria possible to define.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			should be more specific about the consequences, especially whether unplanned subgroup analyses will be considered at all. If yes, clear criteria as to when unplanned subgroup analyses will be considered should be given.	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	13	364	Original wording: "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." Comment: Subgroup analyses required for HTA may depend on the separate PICOs and can therefore not be prespecified in all cases. It must be ensured that a priori planned subgroup analyses as well as (unplanned) post hoc subgroup analyses that are not controlled for multiple hypothesis testing are equally accepted by all member states.	They do not have to be equally accepted by all MS. This is why MS need to know if a given subgroup analysis was planned or not (and if a MS has requested a particular subgroup analysis, it can be expected it will give due consideration to this analysis).
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	13	368	Original wording: "Nevertheless, member states may require further subgroup analyses than those planned at the single study level for assessment at a national level" Comment: All member states should agree on common requirements with respect to subgroup analyses in order to ensure reliability and transparency of the HTA process. A situation in which different member states may require different subgroup analyses must be avoided. Moreover, a feasible number (e.g. 4) of effect modifiers to be investigated should be determined. The investigation of the following four potential effect modifiers has proven to be feasible in the HTA process: gender, age, disease severity or stage, geographic region.	These concerns are within the scope of the D4.2 guideline.
Sebastian Werner	13	341ff	Please include comprehensive discussion on possible limitations of subgroup results.	This guideline is

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
vfa				not intended to be a methodological textbook.
Sebastian Werner vfa	13	362	*"Prespecification of subgroups is being encouraged in the planning of individual clinical studies as it can lend credibility to positive or negative subgroup findings."* Please add, that due to the limitations when interpreting the results of subgroups, the number of subgroups and especially not prespecified subgroups should be strictly limited. Generally, requests for non-pre-specified analyses should be kept to a minimum. The specific context of HTA should be considered. Analyses requested by the HTA authorities regarding specific PICO questions, should not be designated as "post hoc" but rather be considered as "prespecified" hypothesis to be tested in joint clinical assessment.	Duplicated comment.
Roche	13	348 / 5.1	The principal stratification strategy put forward in the ICH E9(R1) addendum allows to define postbaseline "subgroups" in a strict sense within a causal inference framework. We suggest to add this option here, to also allow to draw causal conclusions based on post-baseline variables, at the expense of having to make unverifiable assumptions. But at least the inferential frame would be transparent. And the questions (e.g. effectiveness by dose, response, AE occurrence, etc.) remain scientifically relevant, even if difficult to answer.	We will consider if a sentence is needed for the next version of the draft.
Roche	13	341- / 5.1	Subgroup analyses are associated with many challenges and limitations, which is acknowledged in the draft Guideline D4.5. However, the current proposal of D4.5 is in our view not sufficiently recommending to follow established good practices such as those by Sun et al [1-2]. This will not ensure that subgroup analyses are limited to a meaningful set. For example, the current draft Guideline D4.5 only "encourages" prespecification of subgroups (line 362), while "inviting" member states to ask for further subgroup analyses (lines 368-371). We recommend to include much stronger wording to ensure best practices - for analysis and interpretation - are being followed. For example, only a small number of scientifically plausible effects should be tested and these effects should be prespecified. The interpretation should consider whether chance can explain the	This is not a methodological guideline, nor a best-practice guideline intended for HTD. The interpretation is out of scope of the JCA, and is to be left at MS

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			observed effect. [1-4] Candidates for the investigation of subgroup effects may typically include stratification factors and the subgroups defined in the trial protocol (as for those it seems more likely that the conditions outlined above are being fulfilled). For all other factors, the effect estimate from the overall study population is the best estimate within each subgroup/subpopulation, unless major heterogeneity can be demonstrated based on established best practice. Finally, the Guideline D4.5 should acknowledge that subpopulations lead, per se, to similar issues. Therefore, subpopulations should be treated similarly to subgroups. This means subpopulations should be only requested based on a strong biological foundation, and such subpopulations should be specified as "sub-PICOs" of the overall PICO rather than as separate research questions. Only this approach will ensure that the full population estimate and the subgroup effects are analyzed and interpreted jointly. [1] Sun et al., "How to Use a Subgroup Analysis: Users' Guide to the Medical Literature," JAMA 311, no. 4 (January 22, 2014): 405–11, https://doi.org/10.1001/jama.2013.285063.Sun et al (2014) [2] Sun et al., "Is a Subgroup Effect Believable? Updating Criteria to Evaluate the Credibility of Subgroup Analyses," BMJ 340 (March 30, 2010): c117, https://doi.org/10.1136/bmj.c117. [3] EMA Guideline on the investigation of subgroups in confirmatory clinical trials (https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf) [4] Clarke M, Halsey J. DICE 2: a further investigation of the effects of chance in life, death and subgroup analyses. International Journal of Clinical Practice 2001; 55: 240-242	level. Subpopulation and subgroups are not similar and there is no rational to be provided by MS during the assessment scope (see 4.2 'Scoping process').

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Silke Walleser Autiero Medtronic	13	364-367	This statement is too strong to be considered accurate. While they may be less robust than pre-specified analyses, they do not lack all robustness. Indeed, multiple testing corrections could be applied to the set analyses that were not pre-specified.	Duplicated comment.
GSK	13	366	For unplanned analysis that is not controlled for multiple hypothesis testing, is there any guidance on how to interpret the results, especially P values?	The appraisal of such a result is the concern of MS at national level.
ISPOR	13	5.1	A few examples and references in section 5.1 would be helpful, particularly for "variables that represent methodological characteristics of a study" (l. 356).	We will consider is such an example will be helpful for the next version of the draft.
DVSV	14	387-396	"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." This statement is not clear. The situation described in this paragraph refers to random sampling error: Allocation to treatments was randomised but by chance, prognostic factors might be unbalanced between treatment arms in general or within subgroups, in particular for small sample sizes. This is clear. However it is currently unclear why these random imbalances should lead to more type I errors, confounding or "systematic errors", that is, when the randomisation/allocation process was working as intended and an adequate interaction test is used.	It will be clarified for the next version of the draft.
Advanced Medical Services GmbH	14	387-400	Provide specific criteria for the definitions of subgroups for post hoc analyses, especially a minimum subgroup sample size. Subgroups should have a minimum sample size in order to yield meaningful results. Otherwise, numerous subgroup analyses might be conducted but yield uncertain results. Those results will not allow	MS can request the subgroups they see fit. Therefore, no

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			to draw meaningful conclusions or to make evidenced-based decisions. In addition, we would recommend to limit the number of subgroups for post hoc analysis.	criteria can be defined.
Mihai Rotaru - EFPIA	14	Line 378	Current wording: 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. Proposed wording: A discussion of the credibility of the subgroup 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. Therefore, best practices for analysis and interpretation of subgroups, as well as subgroups defined during the scoping phase and defined as subpopulations, should be applied. That is: • Only a small number of anticipated effects are prespecified and tested. • Interpretation of differences in effect should include: 1. assessing the likelihood that the differences in effects can be explained by chance: this should include the use of a homogeneity or interaction test [17–20] but also take into consideration the resulting sample sizes of the subgroups 2. assessing if a significant subgroup effect is independent, and 3. putting subgroup effects in context, e.g., by considering if an interaction is consistent across studies and across closely related outcomes within the study, and if a biologic rationale supports the hypothesised interaction Rationale: A significant interaction test is only the first step in identifying effect modifiers. Even	Subgroup analyses requirement are made during the assessment scope, and the specific request is therefore out of scope (see 4.2 'Scoping process').

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			with a test for heterogeneity, results can still be chance findings, and interpretation needs to include a medical/biological rationale as well. The EMA guideline on subgroup analysis also highlights the importance of credibility, which refers to "the extent to which subgroup findings can be concluded as being well substantiated and hence relied on for decision making. Credibility depends on the degree of well-founded, a priori definition, the biological plausibility for a particular finding and replication". Cochrane conclude in their handbook: "Importantly, and irrespective of the analytical method, where multiple subgroups have been investigated and/or subgroups effects lack biological plausibility, results should be viewed with caution (Clarke and Halsey 2001). Where there is no particular evidence that trial or participant characteristics impact on the results, emphasis should be placed on the overall effects. " New guidelines on this topic should be open for innovative approaches beyond interaction tests. References: 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 EMA Guideline on the investigation of subgroups in confirmatory clinical trials (https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf) Clarke M, Halsey J. DICE 2: a further investigation of the effects of chance in life, death and subgroup analyses. International Journal of Clinical Practice 2001; 55: 240-242	
Mihai Rotaru - EFPIA	14	Line 391-396	Current wording: 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	We agree and it will be corrected for the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			small sample sizes, it cannot be ruled out that any differences detected between subgroups are caused by systematic errors such as confounding. Proposed wording: In such cases, due to the unbalanced prognostic variable the effect estimates within the subgroups may be biased, and this bias can lead to different results in the different subgroups. Therefore, subgroups with very small sample sizes should be interpreted with even more caution. Rationale: The unbalanced prognostic variable does not affect the treatment received in the trial, the word "confounder" is wrong here.	
Mihai Rotaru - EFPIA	14	397-400	Should be deleted. Interactions of higher order are typically of very limited meaningfulness and should not be in the scope of the JCA.	What is meaningful and is to be included in the scope of the JCA is what MS require during the assessment scope.
Tanja Podkonjak, Takeda	14	388-396	Current text: 'Furthermore, in very small sample sizes, prognostic variables 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.' Proposed wording: 'Furthermore, in very small sample sizes, prognostic variables 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 effect estimates within the subgroups may be biased and this bias can lead to different results in the different subgroups.'	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Rationale: The current statement indicates that a prognostic variable cannot be a confounder because a confounder needs to be associated with both the treatment received and the outcome. However, later in the section the conclusion drawn is that 'a prognostic variable is therefore a confounder' which is confusing.	
EFSPI	14	378f	Current wording: "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." Proposed wording: A discussion of the credibility of the subgroup 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. Therefore, best practices for analysis and interpretation of subgroups, as well as subgroups defined during the scoping phase and defined as subpopulations, should be applied. That is: · Only a small number of anticipated effects are prespecified and tested. · Interpretation of differences in effect should include: 1. assessing the likelihood that the differences in effects can be explained by chance: this should include the use of a homogeneity or interaction test [17–20] but also take into consideration the resulting sample sizes of the subgroups 2. assessing if a significant subgroup effect is independent, and 3. putting subgroup effects in context, e.g., by considering if an interaction is consistent across studies and across closely related outcomes within the study, and if a biologic rationale supports the hypothesised interaction	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Rationale: The statement "An interaction test is a requirement" should be followed by some discussion on the limitations of the subgroup analyses and the use of such an interaction test. A single p-value for interaction should not be the only tool for identifying or excluding subgroups findings. Even with a test for heterogeneity, results can still be chance findings, and interpretation needs to include a medical/biological rationale as well. The EMA guideline on subgroup analysis also highlights the importance of credibility, which refers to "the extent to which subgroup findings can be concluded as being well substantiated and hence relied on for decision making. Credibility depends on the degree of well-founded, a priori definition, the biological plausibility for a particular finding and replication". Cochrane conclude in their handbook: "Importantly, and irrespective of the analytical method, where multiple subgroups have been investigated and/or subgroups effects lack biological plausibility, results should be viewed with caution (Clarke and Halsey 2001). Where there is no particular evidence that trial or participant characteristics impact on the results, emphasis should be placed on the overall effects. " References: 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 EMA Guideline on the investigation of subgroups in confirmatory clinical trials (https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf)	
EFSPI	14	397-399	Current wording: "[…] 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". The operational implications or intent of this paragraph are not clear at all. It is recommended to remove this paragraph to avoid confusion: higher-order interactions should not routinely be within the scope of the JCA.	Duplicated comment.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia	14	375	Original wording: "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	No rationale has to be provided by MS

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Frick SKC Beratungsgesellschaft mbH			require specific subgroup analyses." Comment: Cutoff value(s) for continuous variables defining subgroups that are justified with an adequate rationale should be commonly accepted by all member states. A situation in which different member states may accept different cutoff value(s) for the same type of subgroup analysis must be avoided.	regarding the subgroups they require during the scoping process.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	14	394	Original wording: "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." Comment: Since the results of subgroup analyses are hardly conclusive in case of very small sample sizes, subgroup analyses should only be required for characteristics for which the the resulting subgroups include at least 10 patients. Moreover, subgroup analyses for binary events should only be required if at least 10 events occurred in one of the resulting subgroups.	MS are free to appraise the results of the subgroup analysis the way they see fit, so such a rule of thumb cannot be defined.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	14	397	Original wording: "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." Comment: As mentioned in document D4.5, the interpretation of the results of subgroup analyses may be difficult, especially in case of multiple subgroups or multiple significant interaction terms. Further investigation of these interaction effects is rarely feasible due to the resulting small sample sizes. Hence, the results of subgroup analyses remain inconclusive in many cases and rarely inform the benefit assessment of a new treatment. The necessity and feasibility of subgroup analyses for HTA should therefore be reconsidered as a whole.	MS can require and then appraisal the results of the subgroup analyses they see fit.
Prof. Matthias P. Schönermark, M.D.,	14	402	Original wording: "In the JCA report, information regarding a priori planning of subgroup analyses,	This guideline is about factual

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH			consideration of multiplicity and definitions of subgroups in the protocol and SAP of the clinical studies assessed must be provided." Comment: As mentioned above, subgroup analyses may need to be conducted post hoc in order to comply with the respective PICOs. However, information to be provided for post hoc subgroup analyses are not specified in document D4.5. Since multiplicity adjustments can usually not be performed for post hoc analyses, information regarding consideration of multiplicity cannot be expected for this type of subgroup analysis. Similarly, the protocol and SAP of the clinical studies rarely contain information about post hoc subgroup analyses so that information regarding the definitions of subgroups cannot be expected to be part of the protocol or SAP for post hoc analyses but should be provided in the JCA only.	reporting of the elements MS needs to perform their appraisal at the national level.
Sebastian Werner vfa	14	375-376	*"In the case of subgroup analyses performed because of the assessment scope, justification for the choice of cut-off value(s) pertains to the member state(s) that require specific subgroup analyses."* The definition of subgroups as well as the definition of cut-offs should be prespecified and should clearly follow a biological and clinical rational. There should be an agreement on the definitions with the regulatory bodies.	Already addressed issue.
Sebastian Werner vfa	14	378 – 396	It should be added that a significant interaction test is only the first step in identifying effect modifiers. Further steps include the investigation of qualitative and quantitative interaction, i.e. are there different effect directions in the subgroup categories (qualitative interaction) or are there only different effect sizes in the subgroups, but all effect sizes are unidirectional. This directionality aspect has been addressed in the chapters on sensitivity analyses and should therefore be added here as well. Furthermore, the biological rationale is a key criterium to decide on the effect modification.	The guideline does not imply that interaction test is the only step possible but ask it as a requirement. Regarding the rationale; MS can require the subgroup analyses they see fit according to the

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				scoping process without specifying a rationale.
Sebastian Werner vfa	14	394-395	*"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."* Subgroups should be interpreted with caution. This especially holds in case of small sample sizes. Using the overall estimation of the study might have more certainty.	Already addressed issue.
Sebastian Werner vfa	14	397 - 400	Please add that interactions of higher order are typically of very limited meaningfulness – if they are based on rather small sample sizes.	This technical issue does not seem very relevant tro add.
Advanced Medical Services GmbH - AMS	14	387-400	Provide specific criteria for the definitions of **subgroups for post hoc analyses, especially a minimum subgroup sample size**. Subgroups should have a minimum sample size in order to yield meaningful results. Otherwise, numerous subgroup analyses might be conducted but yield uncertain results. Those results will not allow to draw meaningful conclusions or to make evidenced-based decisions. In addition, we would recommend to **limit the number of subgroups** for **post hoc analysis**.	Already addressed issue.
Bayer	14	387-396	"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. 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." We suggest rewording: Confounding is, by definition, systematic or structural.	It will corrected for the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Randomization only ensures balance on expectation. In a randomized study, imbalances arise due to chance, particularly with small sample sizes, not due to structural confounding. Also, chance imbalances do not induce bias because bias is a large-sample property (See Myth 2 in S Senn. Seven myths of randomisation in clinical trials. https://doi.org/10.1002/sim.5713).	
Bayer	14	378	Given the limited power of individual randomized studies and even evidence syntheses for interaction/heterogeneity testing, the statement "an interaction test is a requirement" can be questioned.	We only state that if a subgroup analysis is performed, we expect such a test.
Roche	14	372-7 / 5.1	*"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."* Cutoff values and rational for subgroup analysis on secondary endpoints are often difficult to pre-specify and should, therefore, be restricted to primary endpoint subgroup analysis in most cases. We recommend emphasizing this point in the guideline.	Already addressed issue.
GSK	14	378-396	It should be added that a significant interaction test is only the first step in identifying effect modifiers. Further steps include the investigation of qualitative and quantitative interaction, i.e. are there different effect directions in the subgroup categories (qualitative interaction) or are there only different effect sizes in the subgroups, but all effect sizes are unidirectional. This directionality aspect has been addressed in the chapters on sensitivity analyses and should therefore be added here as well. Furthermore, the biological rationale is a key criterion to decide on the effect modification. Any recommendation on threshold of 'small sample size' and 'very small sample size' to do a subgroup analysis?	Duplicated comment.
GSK	14	397-400	Please add that interactions of higher order are typically of very limited	Duplicated

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			meaningfulness if they are based on rather small sample sizes.	comment.
ISPOR	14	372-377	It may be worth noting that choosing different cut-offs for a subgroup variable requested by different countries may provide contradicting results, which could in turn merit further analysis.	It is true but it's the decision of a MS to choose one cut-off that is relevant for it. Cut-off has to be predefined during PICO process but could not be leading to decision for HTA.
ISPOR	14	387-396	Discuss some challenges of subgroup analyses in sample sizes. Could the document discuss the Bayesian approach in subgroup analyses to estimate the subgroup treatment effect? It's known that Bayesian approach can better handle small and imbalance subgroup sample size to obtain more reliable results (Henderson et al. 2016). Reference: Henderson, N. C., Louis, T. A., Wang, C., & Varadhan, R. (2016). Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology*, 16(4), 213-233.	The guideline is not intended to be a methodological textbook and we think this technical discussion will not add value for assessors.
MTE	15	5.2	Requirements for JCA reporting box - as mentioned in line 387 on page 14, the power of interaction tests is low, subgroup analyses (interaction tests) are usually considered exploratory rather than confirmatory. This should be the most important consideration for the JCA requirement, and should be noted that many requirements (e.g., hypotheses, multiplicity, …) would not be applicable if analyses are exploratory.	The guideline asks for factual reporting of methodological elements. How these elements will be appraised will be left at the national level.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
MTE	15	419	"An interaction is a requirement" – as indicated in line 425 when individual patient data are not available, "interaction" is then tested based on Cochran's Q for heterogeneity or F test for subgroup as predictor in meta-regression. In both cases there is no "interaction term" in the regression sense. So "An interaction is a requirement" might be confusing in line 419.	It will be clarified for the next version of the draft.
MTE	15	431	One major limitation for meta-regression is the sample size (# of studies) often is not large enough and should be pointed out. The Cochrane handbook suggests a minimum of 10 studies for each study-level variable.	Such limitations are out of the scope of this guideline.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	15	417	Original wording: "In addition, within the assessment scope, subgroups may be defined together with the PICO framework." Comment: As mentioned above, all member states should agree on common requirements with respect to subgroup analyses. A situation in which different member states may require different subgroup analyses, i.e., the required subgroup analyses differ from PICO to PICO, must be avoided.	These concerns are tackled within the D4.2 and D4.4 guidelines.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	15	426	Original wording: "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." Comment: As mentioned above, the results of subgroup analyses remain inconclusive in many cases and rarely inform the benefit assessment of a new treatment. Therefore, the necessity and feasibility of subgroup analyses for HTA should be reconsidered as a whole.	Duplicated comment.
Elaine Stamp PHMR	15	Summary box	It is important to state limitations of sub-group analyses, likely to be underpowered (small sample size) and patients may have other characteristics which vary simultaneously. High chance of false	The guideline asks for factual reporting of methodological

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			positives due to multiple testing and false negatives when not adequately powered.	elements. How these elements will be appraised will be left at the national level.
Marko Ocokoljic (SIOPE)	16	435, 436, 437, 438	"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 and Real World 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, whilst generated Real World Data could be provided by academic capability to satisfy regulatory HTA requirements [24,25]."	This is out of the scope of this guideline.
Tanja Podkonjak, Takeda	16	435-438	Current text: 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]. Current text states that bias cannot be balanced by adjustment when using aggregated data and that IPD data is needed for all studies included in the meta-analyses in order to balance the trials. In practice, it is unrealistic that IPD will be available for all trials as generally individual patient data is the proprietary information of the competitor manufactures. The HTD of the intervention under assessment would not have access to competitors' IPD and it may be the case that the HTD will only have access to the IPD of clinical trial conducted for the intervention under assessment. Using IPD for comparators and from observational studies, if available to the HTD, could allow to conduct more extensive adjusted analyses vs using aggregated data, which may potentially minimise bias. However, it should be considered that availability of IPD does not automatically guarantee more precise results or reduction of potential bias, as it depends on the quality of the IPD available for a comparator, and comparability between the IPD of the intervention and the IPD available for the comparator (e.g., assessment schedule, criteria, or definition of clinical endpoints may introduce bias in comparative effectiveness analyses even if IPD data are available).	We will add references to the D4.3.1 and D4.3.2 guidelines.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			Given the high likelihood that IPD will not be available for all (or even any) of the comparators that may be included in one or more PICOs, and the scoping guidance anticipates multiple comparators, it would be helpful for the guidance document to address and propose methods of how analyses could be conducted when IPD data are not available for one or more comparators and the use of aggregated data would be the only way to compare the intervention under assessment with a comparator included in a PICO during a JCA; for example, methods such as matching-adjusted indirect comparisons (MAIC) and simulated treatment comparisons STCs) are accepted by various health authorities and have been successfully employed in HTA appraisals around the world. Of course, potential limitations of meta-analyses and comparative effectiveness analyses using aggregated data for comparators should be mentioned (e.g., population, variables, and outcomes definitions may differ between studies; aggregated data may be historical data; exhaustive details on outcome definitions and data variables are based only on what is published in the aggregated data, adjusted analyses are only possible based on the variables and outcomes reported for the aggregated data), ; but also that should be the case for analyses using, if available, IPD for a comparator that also have limitations some of which are the same as when aggregated data may need to be used (e.g., population, variables, and outcomes definitions may differ between studies; exhaustive details on outcome definitions and data variables are based only on what is published for the IPD and would be subject to what the data owners report, adjusted analyses are only possible based on the variables and outcomes included in the IPD available, quality of the comparator IPD may lead to bias, comparator IPD may be available for geographies outside the EU [e.g., if only IPD for the comparator is available for a study conducted in Asian patients]).	
EFSPI	16	445	Sensitivity analyses should pertain not just to pre-specified estimands, but be described more generally: as per ICH E9, they are "A series of analyses conducted with the intent to explore the robustness of inferences from the main estimator to deviations from its underlying modelling assumptions and limitations in the data."	They are already described as such.
MTE	16	6	Similar to comment above for page 15, section 5.2, the power of subgroup analysis is often low. It should therefore be mentioned that the subgroup analyses are usually considered exploratory rather than	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			confirmatory. The "Requirements for JCA reporting" box should be revised accordingly (e.g. multiplicity, hypotheses are not applicable when analyses are considered exploratory).	
MTE	16	458-463	This section introduces ICEs and missing data as a potential type of ICE. However, this isn't always the case. It would be helpful if additional examples of ICEs and the role of sensitivity testing was provided in this section.	We will consider if clarifications are needed for the next version of the draft.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	16	435	Original wording: "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]." Comment: Individual patient data is only available for meta-analyses in rare cases in which the same manufacturer conducted multiple clinical studies that are sufficiently similar with respect to the studied patient population and other aspects of the study design. Hence, meta-analyses that include individual patient data cannot be considered an alternative approach in general. Suggestion for rewording: "*If available and accessable, the use of patient individual data for meta-analyses may lead to greater certainty of results, that is, more prec**ise r**esults [...]*"	Already addressed issue.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	16	444	Original wording: "All analyses should be reported." Comment: As mentioned above, interactions that do not reach statistical significance are hardly informative for the benefit assessment of a new treatment. Therefore, only the results of subgroup analyses with significant interaction term should be reported in the JCA report whereas the results of subgroup analyses for which the interaction term does not reach statistical significance should be presented in a separate appendix in the form of the original output tables generated by the statistics software [1].	Already addressed issue.
Sebastian Werner	16	450 - 486	It is appreciated that the draft document contains dedicated parts referring to the	Already

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
vfa			estimand framework per ICH-E9(R1). However, detailed recommendations for an appropriate and accepted use are lacking. The vfa recommends to give detailed recommendations on how to deal with estimands as part of a European method framework that member states commonly accept and apply. The evaluation of estimands as the primary, supplementary and sensitivity analyses should be consistent over member states. Due to the importance of the estimand concept, vfa recommends elaborating on the topic with an extra chapter.	addressed issue.
Richard Birnie Lumanity HEOR	16	Section 6.2	In relation to subgroup analysis in evidence synthesis we suggest the following statement should be removed from reporting requirements "The results (p values) of an appropriate interaction test for all subgroup analyses conducted" Section 6.1 rightly notes the numerous limitations of performing interaction tests in evidence synthesis. Mandating the reporting of p-values is likely to result in oversimplification of the problem and misuse of underpowered tests to label subgroup effects as 'significantly different' or 'not significantly different'	The guideline asks for factual reporting of methodological elements. How these elements will be appraised will be left at the national level.
Roche	16	458 / 7.1	The definition of estimands described within the guideline should be linked to the ICH E9 Addendum. Please include a reference to this addendum in this section of the report.	It will be added.
Silke Walleser Autiero Medtronic	16	458-463	This section introduces ICEs and missing data as a potential type of ICE. However, this isn't always the case. It would be helpful if additional examples of ICEs and the role of sensitivity testing was provided in this section.	Duplicated comment.
ISPOR	16	459	This section introduces ICEs and missing data as potential type of ICE then proceeds to focus on that. It would be helpful if they provided other examples of ICEs and the role of sensitivity testing in those cases.	Already addressed issue.
DVSV	17	478	"...with the aim of explore the impact..." should be "...with the aim to explore the impact ..."	It will be corrected.
Mihai Rotaru -	17	502	Current wording:	No change

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
EFPIA			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. Proposed wording: This sentence should be deleted. Rationale: The definition of statistical rigour depends on the statistical framework used for the decision making. This testing strategy is used for regulatory decision-making, which is different to that of HTA, which is determined by PICOs defined after the trial has readout. Therefore, discussion of this statistical rigour should not be part of the JCA report, and could infer a ranking of health outcomes that the Regulation excludes. This information is provided in the CSR and can be used by Member States who use it to appraise and value the evidence in their own local decision-making.	needed, it says 'if' so this is the same as for an endpoint with 'less rigor', we report the fact but say nowhere it should or should not be used.
EFSPI	17	499-501	"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.". First, if estimands are addressed in PICO scoping, it needs to be explicitly described in the scoping guidance document. Second, it is speculation to say that it may be rare that different estimands are called for. Different policy questions may very easily lead to different estimands. Suggest to remove the speculation.	It will be clarified for the next version of the draft.
EFSPI	17	502	Current wording: 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. This sentence should be deleted. Rationale: The definition of statistical rigour depends on the statistical framework used for the decision making in the national HTA process and should therefore be subject to the member states decision process.	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	17	493	Original wording: "The acceptability of missing data is subject to member state differences in interpretation of their relevance within their respective decision-making process." Comment: As mentioned above, the final assessment scope provided to the HTD shall enable the submission of a dossier fully meeting the needs of every member state. Consequently, there should be defined criteria for the acceptability of missing data agreed upon by all member states. A situation in which different member states may accept different amounts of missing data or interpret the relevance of missing data differently must be avoided.	The HTAR does not ask for a harmonization of the appraisal of methodological elements across all MS.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	17	500	Original wording: "[...] 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." Comment: As mentioned above, the final assessment scope provided to the HTD shall enable the submission of a dossier fully meeting the needs of every member state. Consequently, there should be a defined set of estimands for any given PICO accepted by all member states. A situation in which different member states may require different estimands for a certain PICO must be avoided.	Duplicated comment.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	17	504	Original wording: "The acceptability of sensitivity analyses is subject to member state differences in interpretation of their relevance within their respective decision-making process." Comment: As mentioned above, the final assessment scope provided to the HTD shall enable the submission of a dossier fully meeting the needs of every member state. Consequently, there should be defined criteria for the acceptability of sensitivity analyses agreed upon by all member states. A situation in which different member states may require and/or accept different sensitivity analyses or interpret the same sensitivity analyses differently must be avoided.	The HTAR does not ask for a harmonization of the appraisal of methodological elements across all MS.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
BAH	17	489	"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." What about unplanned interim analysis? (Also see comment no. 2)	It should be reported if an unplanned analysis was conducted and it should be flagged.
BAH	17	493	"The acceptability of missing data is subject to member state differences in interpretation of their relevance within their respective decision-making process." One of the main goals of EU-HTA is harmonization. Therefore, the acceptability of missing data has to follow this principle by consensus between member states.	Already addressed issue.
Roche	17	496-503 / 7.2	The text in lines 496-503 seems to imply that all estimands relevant for HTA were captured in the protocol and SAP - which is the ideal situation. However, there may be cases where estimands of interest for HTA are only identified later on. We recommend to extend this paragraph by including the following elements: 1) JSC should be used to align the needs for marketing authorisation and HTA and to ensure that HTA needs can be appropriately reflected in the protocol/SAP (and using the estimand framework). The scoping process should make use of the ICH E9 (R1) Addendum and clarify whether the estimands in the protocol and SAP are deemed appropriate for HTA purposes. If deemed necessary, additional estimands needed for HTA should be specified.	This document is about reporting not requesting analyses. If a study was conducted without the use of the Estimand framework (which can be completely unproblematic) doesn't mean it hasn't captured all relevant information also for HTAB's. Estimands identified by HTABs will be defined during

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				the PICO process.
Tanja Podkonjak, Takeda	18	517	Current text: 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. It is not clear whether 'prespecified in the study protocol and/or SAP' is in reference to the individual studies included in the evidence synthesis or for the evidence synthesis protocol/SAP. If latter, please also clarify if it is required this be pre-specified at the systematic literature review phase or at the indirect treatment comparison phase (i.e. when conducting the feasibility assessment).	It will be clarified for the next version of the draft.
EFSPI	18	512	"Sensitivity analyses are a set of analyses estimating the same effect but with different methodology […]". More clarity and alignment with the definition of sensitivity analyses in the estimand context (Section 7) is called for. Analyses that change the question of interest should be considered as supplementary analyses. For example, changing the population would amount to changing the question of interest and thus be considered a supplementary analysis. Whereas the random effects versus fixed effects is more reasonable to consider a sensitivity analysis. Furthermore, it is suggested to provide more clarity around the rationale expected for sensitivity analyses, for example along the following snippet from ICH E9 estimand addendum (which applies beyond the setting) about sensitivity analyses as a "[...] structured approach, specifying the changes in assumptions that underlie the alternative analyses, rather than simply comparing the results of different analyses based on different sets of assumptions. The need for analyses varying multiple assumptions simultaneously should then be considered on a case by case basis. A distinction between testable and untestable assumptions may be useful when assessing the interpretation and relevance of different analyses."	No change suggested, again this is reporting not requesting and the reader has hopefully understood that we are aware of the ICH E9 addendum and have endorsed it. We do not need to repeat what is already clearly explained in the addendum.
MTE	18	512-516	This is not entirely accurate. It could also include testing related to the effect of interest, such as a broader inclusion criteria to see if the results are consistent (similar to a falsification test).	Same as above, no change suggested. We endorse the ICH E9 and follow

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				the recommendations given there.
Marjorie Morrison, Lymphoma Coalition	18	597-603	**<u>Real-world data and real-world evidence</u>** While it is understood that randomized controlled trials (RCTs) are broadly considered the "gold standard" for measurement of treatment effectiveness and patient safety, and as RCTs aim for internal validity, the application and value of **external validity or generalizability** cannot be overlooked when considering real-world adaptation or application. In cases of rare lymphomas, there is a risk of low population validity that will require study designs to take into careful consideration the diverse challenges associated with emergent extraneous factors. Further, while the facilitation of clinical trials with a large sample of patients and less stringent inclusion is noted in the guidelines in relation to real-world data, it is not clear what measures will be implemented to universally and consistently **address the recruitment and diversity of patients with rare diseases in clinical trials** – more specifically, where patient size and patient characteristics present as challenges (given that "most clinical studies using real-world data are currently not RCTs") and where there are clear indications that confounding bias is a key consideration in relation to treatment effectiveness.	This is a comment for the D4.6 guideline.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ina S. L. Buchholz SKC Beratungsgesellschaft mbH	8, 10	Section 3.2.1, Box "Requirements for JCA reporting" (exemplary); Box "Requirements for JCA reporting"	Original wording: ○ "How the endpoints were tested (statistical methods), including, if relevant, the multiplicity procedure that was used, […]" Comment: As stated by the EUnetHTA authors, correction for multiplicity often causes problems. Therefore, corresponding information on the statistical procedures should only be expected in the dossier provided that multiplicity adjustment has been performed. The formulation "if performed" was also chosen in D4.5 by the EUnetHTA authors in several other sections in a very similar context. Similarly, the German IQWiG state in their method paper: "*When appropriate and possible, the Institute applies methods of*	It will be corrected for the next version of the draft.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
		(respective parts) in Section 3.2.3	*adjustment for multiple testing."* Suggestion for rewording: "*How the endpoints were tested (statistical methods), including, if ~~relevant~~ performed, the multiplicity procedure that was used, […]*"	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	18	Section 7.2, Box "Requirements for JCA reporting"	Original wording: "There should be a detailed description of the chosen estimand(s), with a focus on the five attributes as well as the ICE strategy." Comment: According to document D4.5, "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." Hence, reporting information about the methodology of the prespecified analyses per study protocol and SAP should be sufficient.	No change suggested. The first bullet point states that the estimand is either defined in the trial protocol or should be provided in response to the PICO requirements. Hence, we expect that those that have not used the estimand framework in the trial protocol will use it in relation to the PICOs.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC	18	Section 7.2/8.2, Box "Requirements for JCA reporting"	Original wording: "All sensitivity analyses should be presented in the report, [...]" Comment: It is not exactly defined which sensitivity analyses are included in "all" sensitivity analyses.	We do not think this requirement should be reformulated.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
Beratungsgesellschaft mbH			Suggestion for rewording: "*All of the performed sensitivity analyses should be presented in the report, [...]*"	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	18	Section 8.2, Box "Requirements for JCA reporting"	Original wording: "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." Comment: It is not specified which kind of "analysis" this sentence refers to. Suggestion for rewording: *"[...] whether the reported sensitivity analysis [...]"*	It is in the section 'Sensitivity analyses in evidence Synthesis', so it refers to those.
Silke Walleser Autiero Medtronic	18	512-516	This is not entirely accurate. It could also include testing related to the effect of interest, such as a broader inclusion criteria to see if the results are consistent (similar to a falsification test).	Duplicated comment.
ISPOR	18	512	Sensitivity analysis could also include tests for something related to the effect of interest like broader inclusion criteria to see if results are consistent or completely opposite like a falsification test.	The guideline does not aim to be a comprehensive methodological textbook.
DVSV	19	529-530	"Both the power of a study and the certainty for correctly rejecting the null hypothesis are..." Statment currently not entirely clear. The "power of a study" and the "certainty for correctly rejecting the null hypothesis" are different wordings for the same concept/probability?	No change, the power is not the same as the certainty to reject the null hypothesis.
Tanja Podkonjak, Takeda	19	521-522	Current text: Thus, in the strictest sense, post hoc analyses are all analyses that are performed because of the results of a previous analysis. This is not necessarily the case. Post-hoc analyses are not only conducted due to the results of a previous analysis, they may also be conducted due to changes in the external environment, shifts in treatment paradigms, patient presentation or	This refers to improper post-baseline sub-group/population analyses within a trial (which would be

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			regulatory changes. Development of medicines takes time and many years may pass from initial protocol design to the trial read-out. As the external environment is not static, post hoc analyses may be required due to a response to external changes. We recommend the current wording be modified to fully reflect reasons why post hoc analyses may be conducted, beyond only the trial results.	the result of a PICO request) or post-launch data generation (years after approval). The first is covered by the document, the second is irrelevant given the JCA timelines.
Tanja Podkonjak, Takeda	19	532-535	Current text: 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. Unplanned post hoc analyses, such as those requested by a HTA as a consequence of the PICO process, should be avoided unless there is a strong justification that the research question cannot be sufficiently answered on the basis of higher certainty results, i.e., the lower certainty of unplanned subgroup analyses is justified by the gains in external validity in comparison to results at study population level. With this perspective, subgroup analyses, beyond pre-specified and stratified groups in the trial, should only be conducted if there is strong clinical, biological or regulatory rationale. Multiplicity should be considered in these situations as multiple analyses can lead to deviating conclusions, and when analyses are data driven. To avoid wrong conclusions, the scope of the JCA should be based on prespecified analyses in the trial and limited to analyses that are justified based on regulatory, clinical or strong biologic rationale to estimate the magnitude of the benefit of the new treatment. Therefore, complementary analyses and subgroup analyses should be avoided as	Already addressed issue

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			much as feasible, and the focus should be on an agreed, limited set of important endpoints. Multiplicity adjustment is then no longer needed for the estimation.	
EFSPI	19	528ff	Actual wording: "[...] unplanned post hoc analyses violate the principles of inferential hypothesis testing. [...] Post hoc analyses should be clearly identified as such to distinguish them from the primary analyses in the JCA". Suggest to remove the first statement. The purpose of JCA is not statistical hypothesis testing, it is to generate evidence that can support evidence-based policy making at a member state level. For this purpose, unplanned post-hoc analyses can sometimes play an important role.	No change, this is a simple fact in statistics and has no consequence with respect to the reporting or use during assessment for the individual HTABs.
EFSPI	19	537-538	Actual woding: "Post hoc analyses should be clearly identified as such to distinguish them from the primary analyses in the JCA.". It is not clear what is a primary analysis. PICOs requested by member state give rise to "primary" analyses in the JCA dossier. But it may often be the case that these analyses are unplanned, in the sense of involving subpopulations, different outcomes etc than those considered and pre-specified in clinical protocols/SAPs. While the principles of defining prospectively all analyses prior to DB lock should be the ultimate goal, this cannot be applied in this context. Suggest to reword to be more precise.	No change needed, both types should be identified as post-hoc. Regardless of whether they are conducted motivated by the developer or the PICO process.
EFSPI	19	549-553	Current wording: "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.". This seems to cover the only scenario where the HTD presents subgroup claims	We agree with the comment, but this is the concern of the scoping process guideline.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			proactively. In the case where subgroup analyses are provided on member state request, member states should provide a clear clinical/scientific rationale. It should not be the task of the HTD to rationalize the member state subgroup request.	
ISPOR	19	9 Post Hoc Analyses in Individual Clinical Studies (lines 518-553)	If "supplementary analyses" are provided, it should be specified whether they were pre-specified or post-hoc analyses.	No change needed, for reporting purposes the same rules apply for all analyses.
ISPOR	19	9 Post Hoc Analyses in Individual Clinical Studies (lines 518-553)	While post-hoc analyses are sometimes merited due to unforeseen considerations, it should be noted that there is also potential bias in the selection of which post-hoc analyses are reported and which are not reported.	Selective reporting is indeed a general concern pertaining to all clinical research.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Sebastian Vinzens, M.Sc. SKC Beratungsgesellschaft mbH	20	Section 10.2, Box "Requirements for JCA reporting"	Original wording: "The report should clearly distinguish between planned analyses and unplanned post hoc analyses." Comment: Unplanned post hoc analyses may be necessary in order to comply with the respective PICO(s). Therefore, the report should not distinguish between planned analyses and unplanned post hoc analyses in terms of acceptability and interpretation of the results. Suggestion for rewording: "*Planned analyses and unplanned post hoc analyses should be clearly flagged as such in the report, respectively*."	We do not think this proposition for rewording is sound.
Sebastian Werner vfa	20	555-560	Indeed, full pre-specification on evidence generation is not possible. Nevertheless, it is mentioned as requirement in various parts of the complete document. Please revise, as this is not possible.	We have already highlighted within the guideline

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
				situations where full pre-specification can be difficult to achieve.
Storz-Pfennig/ Ermisch – GKV-SV	20	557-558	The statement that "Full pre-specification is difficult and often not possible for systematic reviews because knowledge is already available for the underlying studies." might be misunderstood. Thus, we suggest changing the statement to: "Pre-specification is limited by data available from underlying studies". We also refer to comments on Sect. 4.2.1 (the HTD should use available data to adapt evidence synthesis to the assessment PICO).	We do not agree, this is an invitation to deliver poor data and claim that pre-specification wasn't possible. It can be considered as a reversed argumentation!
BAH	20	558	"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. " This is a contradiction regarding to the prespecified analyses that are requested before.	We will consider in the next version of the guideline if a clarification needs to be made.
Roche	13,15	364, 369, Box / 5.1	The 'importance' of endpoints for HTA may differ from the endpoint hierarchy in the trial. Please add text to clarify whether subgroup analyses for patient relevant endpoints (secondary, exploratory and post hoc endpoints) except the primary endpoint are considered to be unplanned and therefore no multiplicity adjustment would be required. For example, patient reported outcomes (e.g. from signs/symptoms, function, to quality of life) often have a lower sensitivity compared to the primary endpoint. Therefore, there will be less power to detect treatment effects for such patient-centric measures and it is unlikely to see anything but nominal significance (without Type 1	The guideline is about factual reporting of methodological and results elements of the evidence submitted. How these elements will be appraised is left

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
			error adjustment) for them. However, given the importance of patient-centric endpoints in HTA, we would encourage language be included for accepting patient-centric endpoints even if they are exploratory (meaning the endpoints were prespecified in the trial protocol but without Type 1 error control).	at the national level.
ISPOR	14,15	378 and 419	The statement "An interaction test is a requirement" should be followed by some discussion on the limitations of the subgroup analyses using such an interaction test. Rightfully, there is mention of the limited power of the test. There should also be mention of the risk of type I error and a wording suggesting that a single p-value for interaction should not be the only tool for identifying or excluding subgroups findings.	Already addressed issue.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Lydia Frick SKC Beratungsgesellschaft mbH	15, 16	Section 5.2/6.2, Box "Requirements for JCA reporting"	Original wording: "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)." Comment: Interactions between a treatment and a subgroup characteristic that do not reach statistical significance show that there is no effect modification and thus do not have relevant impact on the benefit assessment of a new treatment. Therefore, reporting of all results for each subgroup (estimates, effect measures, measure of statistical precision and p values) does not seem reasonable. Instead, only the results of subgroup analyses with significant interaction term should be reported in the JCA report whereas the results of subgroup analyses for which the interaction term does not reach statistical significance should be presented in a separate appendix in the form of the original output tables generated by the statistics software[1].	Duplicated comment.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ina S. L. Buchholz SKC Beratungsgesellschaft mbH	8, 9, 10, 11, 15, 16	Section 3.2.1, Box "Requirements for JCA reporting" (exemplary); Box "Requirements for JCA	Original wording: ○ […] ○ "The α level used to determine if the study was a success. ○ […] ○ For the results for a given test, whether the test was appropriately controlled for multiplicity […]." [For section 4.2.1] ○ "If control for multiplicity was performed, if it was appropriately conducted or not." [For section 5.2 and 6.2]	Duplicated comment.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
		reporting" (respective parts) in Section 3.2.2; Section 3.2.3; Section 4.2.1; Section 5.2; Section 6.2	○ "Whether each statistical test for subgroup analysis was appropriately controlled for multiplicity or not, […]" Comment: As stated by the EUnetHTA authors, correction for multiplicity often causes problems. The control for multiplicity should not be obligatory. Instead, multiplicity adjustment should only be expected if appropriate and reasonable from the methodological perspective. A p-value < 0.05 has been established as adequate for the evaluation of clinical study results in terms of statistical significance. In accordance with this view, the German IQWiG state in their method paper: "*The convention of speaking of a statistically significant result if the p-value falls below the significance level 0.05 (p < 0.05) is quite reasonable in many cases. When appropriate and possible, the Institute applies methods of adjustment for multiple testing* [2]*."*	
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ina S. L. Buchholz SKC Beratungsgesellschaft mbH	8, 9, 10, 11 (twice), 12, 13, 15, 16	Section 3.2.1, Box "Requirements for JCA reporting" (exemplary); Box "Requirements for JCA reporting" (respective parts) in Section 3.2.2; Section 3.2.3; Section 4.2.1; Section 4.2.2; Section	Original wording: ○ […] ○ "Null and alternative hypotheses that are tested. ○ […], the desired FWER level and which FWER was controlled (global level or multiple level). ○ […] ○ The CER level for each statistical test (i.e., the significance level required for each test)." Comment: This level of detail is not necessary to assess the efficacy and safety of a new treatment. Therefore, this information is deemed superfluous in the dossier in terms of making the HTA process an appropriate, focused, and value-added assessment. In any case, this information can usually be taken from the study documents such as the study protocol and SAP and do not need to be included in the dossier.	This guideline is about factual reporting of the methodological and results elements MS need to perform their appraisal at the national level.

Please add extra rows as needed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	HOG response
		4.2.3; Section 4.2.4; Section 5.2; Section 6.2		

Please add extra rows as needed.