**EUnetHTA 21 Public Consultation Comments and Responses**
**of D4.3.1 on Practical Guideline Direct and Indirect Comparisons**

| Name organisation | Country |
|---|---|
| AstraZeneca | Europe Global |
| Bayer AG & Bayer Vital GmbH | Germany |
| EFPIA | |
| EUCOPE | Belgium |
| European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) HTA SIG | Europe |
| European Organisation for Rare Diseases (Eurordis) | France |
| European Union of General Practitioners/Family Physicians - UEMO | Belgium |
| F. Hoffmann-La Roche Ltd (Roche) | Switzerland |
| German Medicines Manufacturer´s Association (BAH) | Germany |
| GSK | Headquartered in the UK, but local operating companies across Europe, therefore GSK is directly impacted by the HTA Regulation |
| IGES Institut GmbH and HealthEcon AG: "IGES LifeScience" | Germany |
| Institut national d'excellence en santé et en services sociaux (INESSS)* | Canada - Québec |
| ISPOR – The Professional Society for Health Economics and Outcomes Research * | Headquarters is based in the USA, but nearly 20% (1 in 5) of our membership lies within the European Union. |
| Lumanity | Lumanity is a global company with several European entities, including in Ireland and the Netherlands. |
| Lymphoma Coalition, Lymphoma Coalition Europe (LCE) | France |
| Medtronic | Switzerland |
| SKC Beratungsgesellschaft mbH (SKC) | Germany |
| Takeda Pharmaceuticals International AG | Switzerland, Belgium with pan-EU local operating companies |
| Verband Forschender Arzneimittelhersteller (vfa) e.V. | Germany |

**This is an organisation outside EU/EEA countries*

**Comments recieved after the deadline**

| **Name organisation** | **Country** |
|---|---|
| Edwards Lifescience | Europe |
| GKV-SV | Germany |

**EUnetHTA 21 Public Consultation Comments and Responses**
**of D4.3.1 on Practical Guideline Direct and Indirect Comparisons**

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | General | | In general, Takeda believes that the tone of the current guidance is mainly objective, factual, and overall reflects an up-to-date picture of the currently available evidence synthesis methods and approaches for healthcare decision making.<br><br>Takeda supports the Hands-on Group's (HOG) recommendation to merge the EUnetHTA 21 Methodological Guideline "Direct and Indirect Comparisons" (D.4.3.2) with "Practical Guideline Direct and Indirect Comparisons" (D4.3.1) to ensure consistency between the two guidelines. However, in doing so, we request that the balanced and objective tone and language of D.4.3.2 be adapted to be more aligned with the current practical guidance D.4.3.1, without value judgements and inclusion of non-randomised studies that is achieved in this guidance document. | Thank you |
| Dr. Norbert Gerbsch<br><br>for IGES Institut GmbH and HealthEcon AG | general | | Comment:<br><br>In general the Practical Guideline Document provides concrete guidance where possible (RCT, meta analyses, and handling of cases with few available studies). In cases where no gold standard exists (comparisons without randomization, indirect comparisons) relevant methodological approaches are discussed. In those cases it is on the assessor/co-assessor to identify and justify the appropriate methodological approach. The document provides a compendium of methods well known and therefore well established in our work on national level in member states. | Thank you |
| EFSPI | general | | An exchange between HTAb and HTD is absolutely necessary to allow clarifying questions from both sides, to discuss data availability and the range of appropriate methodological analyses used for the JCA. The current guideline does not refer to the draft guideline D7.1., and also does not describe a continuous exchange between HTAb and HTD to discuss appropriate methodological analyses and requirements. We recommend to include that. | We added D7.1.1 in Section 7. |

**EUnetHTA 21 Public Consultation Comments and Responses**
**of D4.3.1 on Practical Guideline Direct and Indirect Comparisons**

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| EFSPI | general | | We welcome the flexibility offered to HTD in choosing the most appropriate methods to use for ITC. However, given current proposals that both do not foresee interactions between HTD and the Assessors/Co-Assessors on choice of methodology and data sources, as well as the mandate that PICO(s) selection is policy rather than data driven, it seems likely that the choices of data sources and methodologies could be highly complex. In light of that, more attention should be given to (a) the training required for Assessors/Co-Assessors, (b) the availability of independent statisticians with expertise in both data sources and indirect treatment comparisons methodology to provide technical input to the assessors. | We agree, but this is out of the scope of this guideline. |
| Dr Daniel Widmer<br><br>UEMO | General | | Very technical document for assessors and co-assessors emphasizing the importance of methodological competencies. | Thank you |
| S. Walleser Autiero,<br><br>Medtronic | General | | There is a need for a link or more detailed acknowledgment regarding how the previous Methods guide and this Practical guideline are to be used and the synergies between them. | See Section 2 |
| Matias Olsen, EUCOPE | General | | As this practical guideline states, “there might be exceptional circumstances in which methods of evidence synthesis will need to be applied despite uncertainty or doubt as to their validity.” The EMA definition of ‘exceptional circumstances’ describes this as “A type of marketing authorisation granted to medicines where the applicant is unable to provide comprehensive data on the efficacy and safety under normal conditions of use, because the condition to be treated is rare or because collection of full information is not possible or unethical.”. A more pragmatic recognition of the practical challenges with obtaining evidence under certain conditions needs to be included throughout this guideline.<br><br>As also noted in our comments on deliverable D4.6 “Validity of clinical studies”, increasing the expectation of certainty of results beyond EMA | Thank you for your comment. We refer to our general response to the public consultation of D4.3.2, which also has relevance for your comments. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | accepted thresholds would only delay patients access while high unmet needs exist.<br><br>It is important to recognise that interventions in areas of high unmet need, e.g. therapies for rare diseases or paediatric populations and, ATMPs rely on less conventional methodological approaches in order to improve their interpretability. There should be a stronger recognition that evidence generated outside of the randomised controlled trial (RCT) design can bring highly relevant information to support HTA decisions. This is essential to ensure the success of the EU HTA procedure and to improve patients access to innovative therapies across the EU. We have addressed this in individual comments throughout the document. | |
| Matias Olsen, EUCOPE | General | | According to the project plan version 1.0, this deliverable D4.3.1 was initially named “Comparators and comparisons” while deliverable D4.3.2 was named “Direct and indirect comparisons”. This was later changed into “D4.3.1 Practical Guideline Direct and Indirect Comparisons” and “D4.3.2 Methodological Guideline Direct and Indirect Comparisons”.<br><br>The name “Comparators and comparisons” in the initial project plan suggested that the topic of the comparators may also be discussed in more detail. However, this is not the case. Since the choice of the comparator is also not discussed in detail in other guidelines, a separate guideline for the choice of the correct comparator may be necessary.<br><br>EUnetHTA 21 offers a limited amount of JSC. The second open call for applications ends on August 31st and afterwards no further JSC are planned at this stage. This means that most of the companies will not | The original title for both guidelines was “Comparators and comparisons: Direct and indirect comparisons”.<br><br>However, as described in the project plan, the objectives of this project were to update of existing EUnetHTA guideline and to give practical advice for direct and indirect comparisons. Consequently, the name was later modified to reflect the objectives. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | have an opportunity to discuss and confirm the comparators during a consultation.<br><br>Considering how important the correct choice of the comparative therapy is for the outcome of HTA procedure, comprehensive guidelines for the choice of the comparator are needed. Otherwise, lack of clarity on the appropriate comparative therapy may risk the quality of the assessment and lead to increased risk of discontinued assessments. | |
| Matias Olsen, EUCOPE | General | | Deliverable D4.3.1 is more detailed for some methods of indirect treatment comparisons than deliverable D4.3.2, for example with regards to NMA with time-to-event data and specification of the estimand/inferential goal using propensity score methods with non-randomised evidence. Does any guidance apply for these methods for which there is detailed assessment guidance in D4.3.1 but less detailed methodological guidance in D4.3.2? | We do not repeat details given in D4.3.2 in D4.3.1. Everything that is described in D4.3.2 is also valid for D4.3.1. |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Dr. Lydia Frick<br><br>SKC Beratungsgesellschaft mbH | general | | Comment:<br><br>D4.3.1 mainly describes very specific methods to be applied when combining the results of multiple trials through evidence synthesis. However, this kind of evidence can rarely be generated for JCA as for new innovative drugs the number of clinical trials is usually limited and the evidence basis necessary to robustly apply the methods described in D4.3.1 is usually not available (e.g., "[… ] at least five studies are required for a reliable assessment", "meta-analyses with fewer than five studies are problematic in most cases", "[…] two studies with identical design, which can be found after drug approval", "Statistical detection of inconsistency requires more data than are required to establish a treatment effect"). Therefore, the methods described in D4.3.1 must not be seen as the gold standard of evidence to be presented for JCA but rather represents additional options of evidence generation in very few cases. | We disagree; the Guideline also covers the evidence synthesis if there are only 2 or 3 studies (see e.g., Section 4.3.1).<br><br>The Guideline describes how to deal in practice with evidence syntheses in JCA reports and provides corresponding guidance for the assessors and co-assessors. Within this goal, the Guideline reflects state-of-the-art |

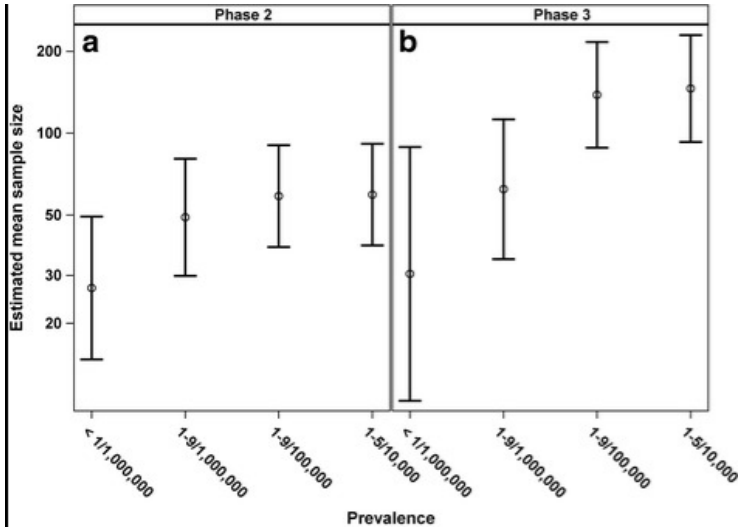
| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The authors of D4.3.1 themselves mention that certain methods might not be feasible in practice (e.g., "[…] the random-effects model is the appropriate choice, although its use might not be feasible in practice"). Generally, the practical feasibility of methods to be applied in a certain case, rather than the hypothetical highest possible methodological standard, should be the driving factor for the decision whether data presented in the dossier is considered for JCA. | evidence synthesis for healthcare decision-making. |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Sebastian Vinzens, M.Sc.<br><br>SKC Beratungsgesellschaft mbH | general | | Comment:<br><br>Multiple different PICOs may be included in the assessment scope for a JCA. However, it is highly unlikely that an HTD will be able to provide evidence of the highest possible quality for all given PICO constellations. Therefore, inherent limitations of the methods for evidence synthesis (e.g., "[… ] such assessments cannot explore the potential impact of unknown effect modifiers") should in no case lead to the dismission of data generated with those methods. Furthermore, it is criticized that consequences for JCA if a chosen method fails to meet the high methodological standards to a certain degree or if one method is chosen over another in a scenario where multiple methods are applicable are not clearly outlined in D4.3.1. Generally, if the data presented for JCA meet the high methodological standards described in D4.3.1, this data should always be considered for the JCA of the new health technology despite methodological limitations. | We disagree; the Guideline (together with the framework given in D4.3.2) clearly describes which method should be preferred in which data situation. We refer also to Section 2 of the Guideline. |
| BAH | general | | One of the main goals of EU-HTA is harmonization. The BAH therefore proposes to find a consensus between HTA bodies and EMA regarding the data quality of real-world data to be accepted. HTA guidelines should therefore require the same data quality standards for the assessment of evidence from non-randomised real-world data that are set for regulatory approval. | Thank you, but it is not within the scope of this guideline to align with regulatory guidelines. |
| Sebastian Werner | General | | The vfa welcomes the document as it attempts to give more practical advice for assessors/co-assessors within the framework described in the Methodological Guideline "Direct and Indirect Comparisons". | Thank you, but the goal of this guideline is not to determine the validity of |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| vfa | | | However, the guideline lacks clear recommendations about the JCA's validity assessment and acceptable methodology that assessors and HTD can use for direct and indirect comparisons. The guideline implies that validity assessments are not a subject matter of the JCA report itself but of Member States only. The vfa does not support this view. The JCA report should comprise scientific assessments of validity of submitted comparisons, including conclusions. This document therefore should provide guidance on acceptability of methodological approaches. Methodological assessments should follow international standards of evidence-based medicine and should be based on well-defined criteria allowing flexible, context specific approaches ensuring best fit for the data situation. Consensus on validity in the JCA can best be reached based on a harmonised European methodological framework that is accepted by the Member States. | the comparative assessment within the context of national decision making but to clearly describe the limitations, where needed, associated with the methods which is consistent with the HTA Regulation. |
| Sebastian Werner<br>vfa | General | | The guideline calls into question the fundamental principle of evidence-based medicine to use the best available evidence to inform health care decisions. The guidance sends out the signal that only data with "preferable methods" are useful for JCA and that data with less appropriate methods with high uncertainty are considered not valid, even when they constitute the best available evidence. This notion is especially harmful for health technologies with specificities in their evidence generation, in which the required data for preferable methods might not be readily available due to ethical (disease areas with high unmet need) or feasibility reasons (rare or orphan diseases). Data with higher uncertainty due to specificities of the health technology can be a valid source of evidence and should be considered. JCA must follow international standards of evidence-based medicine. The principle of evidence -based medicine to "use the best available evidence" must be ensured. | We disagree and refer to our general response to the public consultation of D4.3.2. |
| Sebastian Werner<br>vfa | General | | A constructive methodological exchange between the assessors and the HTD is needed to ensure highest level of quality of the JCA. The vfa recommends establishing a PICO meeting with the possibility of dialogue about methodological aspects of the dossier preparation. | We added a reference to D7.1.1 in Section 7. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Additionally, a scientific discourse between assessors, academia, and industry should be promoted, establishing platforms for exchange. | |
| Sebastian Werner<br><br>vfa | General | | The vfa appreciates that the guideline provides recommendation on frequent reviews of the guidelines on direct and indirect comparisons under the direction of the future Methodological Subgroup. These reviews are necessary to ensure an up-to-date and state-of-the art guidance on direct and indirect comparisons. A public consultation on updated guidelines must be ensured. | Thank you |
| Mihai Rotaru, EFPIA | General | | EFPIA would like to thank the Hands-on Group (G-BA, HAS, IQWIG, NCPE, NOMA) for producing this objective and factual practical guide (PG) that we feel, in general, reflects state-of-the-art evidence synthesis for healthcare decision making. We are also pleased to see that the PG contains no personal opinions or value judgements on the acceptability of evidence from non-randomised studies for the European Joint Clinical Assessment.<br><br>We support the HOG's recommendation to update the methods guide for Direct and Indirect Comparison every 3 years as we agree that evidence synthesis methods are constantly evolving and new methods emerging. It is therefore important that these guides continue to reflect state of the art practice and international standards of evidence-based medicine, as stipulated in the Regulation. That said, it is equally important that the same frequency of update is reflected for other EUnetHTA21 methods guides, for example, the forthcoming consultation D4.4 Endpoints. | Thank you |
| Mihai Rotaru, EFPIA | General | | There are multiple approaches for conducting indirect comparisons, all of which are underpinned by different assumptions and data requirements. We believe this draft practical guide would benefit by providing assessors/co-assessors with further guidance on the context specific factors (for example, regulatory and ethical considerations related to trial design, epidemiology and its impact on data availability) that should be considered when assessing methods(s) used for direct | There are a number of other deliverables within EUnetHTA21 (including validity of clinical studies, applicability of clinical evidence and scoping where the issues |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | and indirect comparison. Currently, it is unclear how the research questions, the context of evaluation and the associated available evidence should all be taken into consideration.<br><br>EFPIA therefore recommend introducing explicit wording that context specific considerations may be needed. The current draft guideline implicitly acknowledges this need by asking that both statisticians and healthcare professionals should be consulted (line 134) and that "there is an element of subjectivity in the assessment of many assumptions" (line 118)." | described in the comment are raised. We do not consider it within the scope of this particular guideline to provide the overview as described. |
| Roche | general | | An exchange between HTA and HTD is absolutely necessary to discuss data availability and appropriate methodology for a given context and decision problem. .<br><br>We believe that continuous dialogue between the HTD and HTA from JSC onwards is needed and particularly important around scoping where all PICO related questions should be clarified.<br><br>The current guideline does not refer to the draft guideline D7.1., where such interactions are specified.<br><br>We recommend referring explicitly to D7.1. and to ensure that D7.1 offers two-way communication between HTA and HTD to meaningfully discuss appropriate methodology and requirements. | Thank you - We added a reference to D7.1.1 in Section 7. |
| François Houÿez, Eurordis | General | | Eurordis agrees direct and indirect comparisons need to be conducted with the highest scientific rigour, both for orphan medicinal products and non-orphan, and for medical devices.<br><br>Recital 2 of Regulation (EC) No 141/2000 specifically states that "patients suffering from rare conditions should be entitled to the same quality of treatment as other patients". | Thank you – we agree with your observations. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The same quality can only be ensured if highest scientific standards are also applied to orphan medicines, when assessing their benefits and risks, and when assessing their relative effectiveness. Even If small populations and relative lack of knowledge on many rare diseases make the development of health technologies and the evaluation of the clinical research results more complex, no special case should exist when assessing the relative effectiveness of orphan medicinal products.<br><br>In particular, when no direct comparison exists, the evaluation of the relative effectiveness should follow highest statistical methods to reduce uncertainty and obtain clear conclusions. The decision by the HTD not to conduct a head-to-head comparison but to use indirect comparisons instead should not be driven by the impression that less rigour can be applied when analysing indirect results.<br><br>This said, head-to-head comparisons usually require a high number of trial subjects, and clinical superiority can hardly be demonstrated, except when the difference between new product and comparator is important.<br><br>The HTD might then be tempted to compare the new product with a placebo, which poses ethical dilemma when a medicine is already authorised for the condition and should serve a standard of care comparator product.<br><br>This can drive developers to explore the use of the new product in patients who failed all other treatment options, making the development even more complex, with usually even fewer patients, but where placebo can be ethically valid.<br><br>This highlights the imperative for regulatory/HTA scientific advice for the development of orphan medicinal products, to maximise chances that R&D can conclude.<br><br>In addition, the analysis of clinical trial data from the NIH clinical trials registry (www.clinicaltrials.gov) showed that among rare diseases with | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | higher prevalence, efforts could be made to increase the sample size for confirmatory /phase III trials.<br><br>  <br><br>(W, Willis A, Tudur Smith C, Day S, Miller F,<br>Madan J, Posch M, Zohar S, Stallard N. Does the low prevalence affect the sample size of interventional clinical trials of rare diseases? An analysis of data from the aggregate analysis of clinicaltrials.gov. Orphanet J Rare Dis. 2017 Mar 2;12(1):44) | |
| Natacha Bolanos, Lymphoma Coalition | General | | Patient-reported outcomes<br><br>The role of patient-reported outcomes measures - with respect to such areas as daily life activity, psychological health, social functioning, and overall life quality - must be properly integrated and considered given the increasingly common useage to support timely comparison of treatment outcomes however, we note that this area is generally overlooked in the documentation. We wish to reiterate that any delay in the integration of patient-reported outcomes in comparative | Thank you and we agree on the importance of patient reported outcomes. The outcomes to be examined will be determined within the scoping process and the development of the PICO. HRQOL, as |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | evidence for new treatments/therapies will greatly reduce the value of comparative research for improving the decisions of physicians, payers, and patients. | described is an important outcome that is often included in the comparative assessment. We believe that the methods described in the document can apply to patient reported outcomes. |
| James Ryan<br>AstraZeneca | General | General | Thank you for the opportunity to comments on this Practical Guideline. AstraZeneca supports the response from EFPIA. | Thank you. |
| GSK | General | General | Some HTA bodies do not currently accept indirect treatment comparisons (ITCs) as evidence. Is there any guidance on how Member States should incorporate the EU HTA assessment of ITCs into their decision making? | It is not within the scope of this guideline to advise on national decision-making. |
| Sebastian Werner<br>vfa | 6 | 105-110 | "Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which a direct or indirect treatment comparison was submitted. It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so."<br><br>The guideline gives reporting requirements to assessors regarding methodological aspects of direct and indirect comparisons to enable Member States' evaluation of validity. The guideline lacks | We refer to our general response to the public consultation of D4.3.2, which also applies here. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | recommendations about the JCA's validity assessment per se and an acceptable ("valid") methodology that HTD can use for direct and indirect comparisons. Thus, the guideline is of limited use for the HTD in choosing the appropriate methods for their dossier preparation. Most importantly, the guideline implies that validity assessments are not a subject matter of the JCA report itself but of Member States only.<br><br>JCAs must provide a scientific evidence-based assessment (HTAR, Recital 2). Scientific conclusions on validity are in the scope of the JCA, as validity is central for the degree of (un)certainty (HTAR, Recital 28). Validity conclusions are of scientific nature and should be based on well-defined methodological criteria to ensure that decisions about validity in the JCA are scientifically objective. Therefore, the JCA report should give scientific conclusions about the validity of submitted direct and indirect comparisons and the guideline document should provide guidance on acceptability of methodological approaches. Member States might deviate from the validity assessments of the JCA report, after considering national complementary clinical analyses.<br><br>The vfa supports the inclusion of validity assessments and scientific conclusions in the JCA report. EUnetHTA21 and the CG should use its best endeavours to reach a consensus. The assessment should follow international standards of evidence-based medicine and should be based on well-defined methodology with flexible, context specific approaches ensuring best fit for the data situation. Consensus on validity can best be reached based on a harmonised European methodological framework that is accepted by the Member States. | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Matias Olsen, EUCOPE | 6 | 106-110 | The EU HTA procedure should establish a common methodology accepted across Member States, e.g. for indirect treatment comparisons to ensure the aims of the EU HTA Regulation are met.<br><br>The EU HTA procedure should not be reduced to a mere evidence-gathering exercise; despite the fact that the degree of uncertainty which is acceptable by each Member State may differ, guiding thresholds should be established at the EU level to prevent multiple and duplicative assessments with additional evidence requests.<br><br>The Regulation on health technology assessment (EU) 2021/2282 Recital 6 underscores the aim to prevent the current parallel and multiple assessments insofar that they "…lead to both duplication and variation in outcomes, resulting from the specific national healthcare context.".<br><br>If there are no guiding thresholds included in the methodology to complement the guidance on which limitations might arise, when certain conditions apply, and the approaches to be applied when certain evidence is available, the EU HTA Procedure is unlikely to prevent multiple clinical assessments with further requests for evidence at Member State level.<br><br>This is a critical point for the success of the EU HTA procedure, as HTD would in practice have to undergo two clinical assessments, one at the EU level and another at the national level. This would have the unintended effect of increasing the burden for developers, especially | It is not within the scope of this deliverable to determine the acceptability of common methods across Europe or indeed to determine the common acceptability thresholds on uncertainty. This would also be outside the scope of the HTA Regulation. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | for small to mid-sized companies, prolonging the time to place products on the market and delay patients' access to innovative therapies. | |
| Sebastian Werner<br>vfa | 6 | 124 - 126 | *"The argument that the required data to apply preferable methods are not available, is insufficient to demonstrate the validity of the results coming from less appropriate methods with high uncertainty."*<br><br>This comment of the authors "to the argument" sends out the clear signal that only data with preferable methods are useful for JCA. Data with less appropriate methods with high uncertainty are considered not valid, even if they would constitute the best available evidence. Thus, this comment clearly calls into question the fundamental principle of evidence-based medicine to use the best available evidence to inform health care decisions [1]. This is especially harmful for health technologies with specificities in their evidence generation, in which the required data for preferable methods might not be readily available due to ethical (disease areas with high unmet need) or feasibility reasons (rare or orphan diseases). In these specific situations, adapted methods for validity assessment should be applied. The use of adapted methods for JCA of health technologies with specificities in their evidence generation is set by the HTAR (Recital 24). Data with higher uncertainty due to specificities of the health technology are still a valid source of evidence and should be considered. JCA must follow international standards of evidence-based medicine. The principle of evidence -based medicine to "use the best available evidence" must be ensured.<br><br>The vfa proposes to delete this sentence. | Thank you for your suggestion. We do not agree that this should be deleted, as it is an important point. However, we have reformulated this text. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | [1] Sackett et al. BMJ 1996; 312:71<br>https://www.bmj.com/content/312/7023/71 | |
| Matias Olsen, EUCOPE | 6 | 124-126 | The lack of data should be enough argument in itself to resort to alternative methods of interpretation of data when the preferred methods cannot be applied. Lack of data in itself is never used to demonstrate the validity of a method. Uncertainty is managed using sensitivity analysis. Some level of evidence even with uncertainty will always be better than no evidence at all.<br><br>Remove:<br><br>"~~The argument that the required data to apply preferable methods are not available, is insufficient to demonstrate the validity of the results coming from less appropriate methods with high uncertainty.~~" | Repeated comment theme from one above (Sebastian Werner) |
| Mihai Rotaru, EFPIA | 6 | 124-126 | **Application of preferable methods**<br><br><u>Current wording:</u><br>The argument that the required data to apply preferable methods are not available, is insufficient to demonstrate the validity of the results coming from less appropriate methods with high uncertainty. | We disagree and refer to our general response to the public consultation of D4.3.2, which also applies here. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br><br>~~The argument that the required data to apply preferable methods are not available, is insufficient to demonstrate the validity of the results coming from less appropriate methods with high uncertainty.~~ **When data to perform an indirect comparison is limited (for example, for orphan medicinal products or advanced therapy medicinal products), the HTD must provide sufficient evidence to justify their chosen method and undertake rigorous sensitivity analyses to characterise uncertainty in their estimates of relative treatment effect.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>Firstly, EFPIA believes that the current wording “preferable methods” and “less appropriate” are subjective and therefore likely to vary depending on the EU assessors’ and co-assessors’ views and opinions on different methods of evidence synthesis.<br><br>Secondly, insufficient data may be unavoidable in certain circumstances, for example when developing treatments for rare diseases. In this case, by definition, there will be fewer studies and fewer events to conduct robust comparisons. We believe that the particular context should be appreciated by the assessor, as requested in the EU Regulation 2021/2282.<br><br>As such, EFPIA recommends that this statement in the document is revised based on the proposed wording provided. | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Reference:<br><br>1. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance) | |
| EFSPI | 6 | 107-10 / 2 | Current wording<br>"It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so."<br><br>Suggested wording:<br>"It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). ~~Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so.~~<br><br>Rationale:<br><br>The Practical Guideline D4.3.1 should foster the adherence to (recognized) best practices in the conduct of direct and indirect treatment comparisons. The guideline therefore discusses the assumptions, strengths and limitations of given approaches based on the available literature, which will help assessors and co-assessors to evaluate the scientific validity of presented indirect comparisons. Therefore, JCA reports should include a detailed discussion on the | We refer to our general response to the public consultation of D4.3.2, which also applies here. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | scientific validity of presented indirect comparisons, their uncertainty, and the strengths and limitations of the underlying evidence base and/or methods."<br><br>We note that it is important to continue to allow individual member states their current flexibility in what methodologies and data sources they consider acceptable and how they define value. There are methodologies and data sources that maybe considered scientifically valid (i.e., would be considered acceptable for a peer-reviewed journal) but that does not mean those methods or data sources are considered equally acceptable by all agencies. The JCA should allow for that, and the assessment of the validity of a method should not be driven by the acceptability of a certain MS, and validity and acceptability should not be mixed up in the wording. | |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Sebastian Vinzens, M.Sc.<br><br>SKC Beratungsgesellschaft mbH | 6 | 107, 118 | Original wording:<br><br>"It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs)."<br><br>and<br><br>"[…] decisions might vary between MSs."<br><br>Comment:<br><br>According to HTA Regulation (EU) 2021/2282 Article 8 (6) sentence 2 "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." | We refer to our general response to the public consultation of D4.3.2, which also applies here. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | 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 direct and indirect comparisons accepted among all member states should be established. | |
| GSK | 7 | Section 3.1 | During the scoping and consultation process with the HTD, how to ensure that additional MS requirements on top of the requested EU HTA analyses are minimised to avoid duplication of efforts? | This is out of the scope of this guideline. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 9 | 232-239/ 3.2.1. | It would be interesting to clarify the concepts of interaction and effect modification in this section. | See lines 200-217 |
| Mihai Rotaru, EFPIA | 9 | 218-220 | **Assessment of similarity - Evaluation of effect modifiers**<br><br><u>Current wording:</u><br><br>It is essential that the process used to identify relevant effect modifiers is comprehensive and transparently reported. This process should include a comprehensive review of the literature and consultation of healthcare professionals with knowledge of the disease area.<br><br><u>Comment:</u><br><br>EFPIA agrees that effect modifiers should be based on a comprehensive review and clinical input to ensure clinical relevance. Firstly, we would like to highlight the following consideration in the practical guideline that the outcome of the comprehensive review could be to conclude the absence of a relevant effect modifier (leading to unknown bias), or to the insufficient similarity of studies in many cases. Secondly, we would like to highlight that this may not always be possible given the limited time to respond to potential large number of comparators and populations that will derive from the proposed PICOs | Yes, it could be that no valid method for an indirect comparison is possible because some relevant effect modifiers are missing. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | approach. As such, we believe this reinforces the feasibility concerns raised during the scoping process given indirect comparisons will likely be needed to address some PICO but the validity of some indirect treatment comparison methods will be limited (mainly because important effect modifiers will be missing, not reported in the comparator study, or similarity will be insufficient) despite extensive requirements to support these. | |
| Sebastian Werner<br>vfa | 9 | 229 – 231 | *"After assessment of all these aspects, a decision has to be made about whether all studies considered in the evidence synthesis are comparable with respect to possible effect modifiers across all interventions (sufficient similarity) or not (insufficient similarity)."*<br><br>Different Member States may accept varying levels of uncertainty for their decision making with different methodological requirements that could differ compared to EU HTA. Therefore, the opportunity for interaction and scientific discourse between assessor, patients, clinical experts and the HTD is necessary to discuss the available evidence. A Scoping Meeting should be implemented for discussions of the proposed evidence synthesis, incl. how to address the uncertainty and possible conflicting requests between EU HTA and Member States. | This is out of the scope of this guideline. |
| Sebastian Werner<br>vfa | 11 | 336 | *"Inconsistency models: Bayesian NMA"*<br><br>Evidence syntheses based on Bayesian methods are a possible option for JCA. However, some Member States do not accept Bayesian models especially considering that the prior assumptions are insufficient and require evidence syntheses based on frequentist methods only. Therefore, an early methodological exchange between assessor, experts, patients, healthcare professionals and HTD is | This is out of the scope of this guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | required before conducting the analyses for a dossier to resolve conflicting requests between EU HTA and Members States. A Scoping Meeting should be implemented for discussions of the proposed evidence synthesis, incl. how to address the uncertainty and possible conflicting requests between EU HTA and Member States. | |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 20 | 667-671 | Current wording:<br><br>When assessing a population-adjusted indirect comparison, the problem of multiplicity arising from 'researcher degrees of freedom' must be considered. Indeed, the number of methods and potential covariate combinations available to the modeller raises the possibility of selecting the method that produces the most favourable results for the intervention under assessment. For this reason, these methods are often more suitable as an exploratory analysis rather than as the primary analysis.<br><br>Suggested wording:<br><br>~~When assessing a population-adjusted indirect comparison, the problem of multiplicity arising from 'researcher degrees of freedom' must be considered. Indeed, the number of methods and potential covariate combinations available to the modeller raises the possibility of selecting the method that produces the most favourable results for the intervention under assessment. For this reason, these methods are often more suitable as an exploratory analysis rather than as the primary analysis.~~<br><br>Rationale: | We do not exclude the use of population-adjusted methods a priori (see Section 5 and the subsections therein). The Guideline also offers flexibility in the choice of methods. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Takeda strongly disagrees with this statement as it is biased against some among the most advanced methodological approaches in the design and conduct of indirect-treatment comparison and network meta-analyses.<br><br>The critical point that assessors will need to establish, in the context of the JCA, will be the suitability of the methodological approaches adopted to address the relevant research questions from the PICO. Should a given research question be answerable only through adequately justified population-based ITCs, then these should not be dismissed “a priori”, and only considered as exploratory analyses.<br><br>We are concerned that the potential multiple PICOs due to the scoping methodology, combined with the stringent thresholds and proposals in this methodological guideline, will prevent HTDs and assessors from conducting a JCA. The combined impact of these approaches will make it very challenging to provide an evidence package that meets both approaches, which may ultimately adversely impact patient access in the EU.<br><br>Takeda strongly recommends the methodological guideline for direct and indirect comparisons be adjusted to allow for flexibility in the selection the appropriate method(s) for the relevant research question. | |
| Mihai Rotaru, EFPIA | 20 | 667-671 | Current wording:<br><br>When assessing a population-adjusted indirect comparison, the problem of multiplicity arising from ‘researcher degrees of freedom’ must be considered. Indeed, the number of methods and potential covariate combinations available to the modeller raises the possibility of selecting the method that produces the most favourable results for the intervention under assessment. For this reason, these methods are | We do not exclude the use of population-adjusted methods a priori (see Section 5 and the subsections therein). The Guideline also offers |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | often more suitable as an exploratory analysis rather than as the primary analysis.<br><br>Proposed wording:<br><br>~~When assessing a population-adjusted indirect comparison, the problem of multiplicity arising from 'researcher degrees of freedom' must be considered. Indeed, the number of methods and potential covariate combinations available to the modeller raises the possibility of selecting the method that produces the most favourable results for the intervention under assessment. For this reason, these methods are often more suitable as an exploratory analysis rather than as the primary analysis.~~<br><br>[note: ~~strikethrough~~ indicate text recommended to be deleted].<br><br>Rationale:<br><br>EFPIA strongly disagrees with this statement and believes it shows a concerning bias against some among the most advanced methodological approaches in the design and conduct of indirect-treatment comparison and network meta-analyses.<br><br>The critical point that assessors will need to establish, in the context of the JCA, will be the suitability of the methodological approaches adopted to address the relevant research questions reflected in the PICOs. Should a given research question be answerable only through adequately justified population-based ITCs, then these should not be dismissed "a priori", and related to the role of exploratory analyses. | flexibility in the choice of methods. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The requirements for an acceptable indirect comparison as set out in this draft guidance (D4.3.)1 will be stringent and unfeasible for many health technologies, particularly when recognising those included in Phase I and II JCA process (orphan diseases, oncology and ATMPs), given the greater potential for EU marketing authorisation on the basis of assessment with non-RCT evidence (i.e., single arm trials) and dynamic treatment pathways associated with scientific innovation.<br><br>EFPIA is concerned that when combining the potential multiple PICOs and the stringent thresholds and proposals in this methodological guideline, will prevent HTDs and assessors from conducting a JCA on the basis of the most robust comparative effectiveness assessment possible with the available data. As such, not providing benefit to MS as outlined in the EU HTA regulation, '*HTA is able to contribute to the promotion of innovation, which offers the best outcomes for patients and society as a whole, and is an important tool for ensuring proper application and use of health technologies (L 458/1, 3)'.*[1]<br><br>EFPIA believes that the methodological guideline for direct and indirect comparisons should allow for flexibility in the selection the appropriate method(s) for the relevant research question.<br><br><u>Reference:</u><br>1. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance). | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| EFSPI | 20 | 670-671 | Actual wording:<br>"these methods are often more suitable as an exploratory analysis rather than as the primary analysis. ".<br><br>Suggested wording:<br>"these methods require special care in their argumentation and application".<br><br>Rationale:<br>This goes against D4.3.2 (Key points 4a), according to which population adjustment is the preferred alternative: "For cases in which the property of similarity does not hold, the usual methods for direct or indirect comparisons are invalid. In this scenario, population-adjustment methods may be considered as an alternative approach". It is problematic if the practical guideline in case of effect modification will neither accept unadjusted analyses, nor acknowledge adjusted analyses as an appropriate alternative. It is suggested that the guideline is reworded. | Our statement is correct and no change is required. "... often more suitable …" does not mean that this is always the case. |
| Sebastian Werner<br>vfa | 20 | 678 - 679 | *"By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable."*<br><br>There is an agreement, that disconnected networks have limitations compared to RCTs and may increase uncertainty. However, in special situation and to value every evidence, results from disconnected networks might be considered in JCA and might be supported with | It is possible to perform valid analyses also in disconnected networks, see Section 6. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | other sources of evidence. There should be no general rejection, rather a careful discussion. | |
| S. Walleser Autiero,<br><br>Medtronic | 20 | 678-679 | This is a very strong statement. There are a number of limitations and very strong assumptions, but ITCs should not be considered entirely unreliable, and the current published literature will attest to a growth in their use. Please consider re-phrasing this sentence to acknowledge and discuss their limitations and challenges but be open to accepting them if they are done well. | This statement refers not to ITCs in general. It only refers to ITCs applied to disconnected networks. |
| Matias Olsen, EUCOPE | 20 | 678-679 | As noted in another comment, the methods described are, for innovative therapies, in many cases the only way to pool the data as comparisons versus all other treatments are impossible to be made using research clinical trials. The guidance should describe the limitations without precluding their use *per se*.<br><br>There are currently many health technology assessment procedures where these methods are used, correctly interpreted, and have proven their usefulness in reducing uncertainty of a health policy decision regarding the use of a certain therapy.<br><br>Remove:<br><br>"~~By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable.~~". | This statement refers not to ITCs in general. It only refers to ITCs applied to disconnected networks. Possible methods for ITCs in disconnected networks are described in Section 6. |
| Mihai Rotaru, EFPIA | 20 | 678-679 | <u>Current wording:</u> | Possible methods for ITCs in disconnected |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable.<br><br>Proposed wording:<br><br>By describing these methods here, we are not endorsing them~~., and once again reiterate that estimates arising from unanchored ITCs are unreliable.~~ **Estimates from unanchored ITCs tend to rely on strong assumptions, e.g., that all effect modifiers and prognostic factors are accounted for[1]. Despite these limitations, unanchored methods often represent the only viable approach to ITCs, particularly when in presence of single-arm trials.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>EFPIA recognises that, as noted in Phillippo et al, 2016[1], an unanchored MAIC or STC rests on strong, often untestable, assumptions. Nevertheless, the recognition of the limitation of these methods should not lead to their dismissal in the context of a JCA, particularly when the only alternative may be a naïve ITC.<br><br>Reference:<br><br>1. Phillippo DM, Ades AE, Dias S et al. NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submission to NICE. London, UK: National Institute for Health and Care Excellence; 2016. | networks are described in Section 6. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| EFSPI | 20 | 667-670 / 5.1 | The issue of “researcher degree’s of freedom”, or similarly, “assessor’s degrees of freedom” exists, meaning a tradeoff between flexibility and prespecification. We suggest to handle these two types of “degrees of freedom” similarly. Prespecification of the analyses can take place in the scoping phase, therefore a strong interaction between HTD and HTAb is needed in the scoping process to avoid data dredging and cherry picking on both sides. | Thank you for your comment. However, no change is required. We believe that “researcher degree of freedom” is an innate issue with population adjusted indirect comparisons and cannot be entirely mitigated against. In fact, true pre-specification cannot be achieved for evidence synthesis, and it is particularly problematic for population adjustment. Therefore, it is important to highlight the issue in this document. While interaction between the HTD and the assessor is important during the scoping process, it does not solve the problem of “researcher degrees of freedom” and inclusion in the specified paragraph may be confusing for the reader. |
| Roche | 20 | 667-670 / 5.1 | The issue of “researcher degree’s of freedom”, or similarly, “assessor’s degrees of freedom” exists, meaning a tradeoff between flexibility and prespecification. We suggest to handle these two types of “degrees of freedom” similarly, i.e. either researcher and assessor do not | Thank you for your comment. However, no change is required. We believe that “researcher |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | prespecify analyses and therefore have these degrees of freedom, or they prespecify both their analyses and do not deviate from those after seeing the data. | degree of freedom" is an innate issue with population adjusted indirect comparisons and cannot be entirely mitigated against. In fact, true pre-specification cannot be achieved for evidence synthesis, and it is particularly problematic for population adjustment. Therefore, it is important to highlight the issue in this document. While interaction between the HTD and the assessor is important during the scoping process, it does not solve the problem of "researcher degrees of freedom" and inclusion in the specified paragraph may be confusing for the reader. |
| EFSPI | 20 | 678-9 / 5.1 | Current wording:<br>"By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable."<br><br>Suggested wording:<br>~~"By describing these methods here, we are not endorsing them, and~~ | No change; it is important to highlight that estimates arising from unanchored ITCs are unreliable.<br><br>Possible methods for ITCs in disconnected |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~once again reiterate that estimates arising from unanchored ITCs are unreliable."~~<br><br>Reason for change:<br>Such an absolute statement seems neither fair nor correct. Any method will lead to unreliable results if applied incorrectly. Whilst unanchored ITCs require stronger assumptions than some other evidence synthesis approaches, the technique has proven to be useful in some contexts. | networks are described in Section 6. |
| EFSPI | 24 | 852 - 855 | Current wording:<br>In summary, although these methods are often presented as the only way of quantifying a relative treatment effect, this does not mean that the method will be of sufficient standard to confidently estimate a relative treatment effect. A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information<br><br>Suggested wording:<br>~~In summary, although these methods are often presented as the only way of quantifying a relative treatment effect, this does not mean that the method will be of sufficient standard to confidently estimate a relative treatment effect. A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information~~<br><br>Reason for change:<br>Such methods should not be acknowledged to be insufficient. A MAIC may still be considered in situations where aggregate data can be matched to IPD. However, those methods are associated with greater uncertainty due to the strong modelling assumptions. If the uncertainty is accepted for decision making should be left to MS. | No change; it is important to highlight that these methods may be of an insufficient standard. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | 24 | 852 - 855 | Current wording:<br><br>*"In summary, although these methods are often presented as the only way of quantifying a relative treatment effect, this does not mean that the method will be of sufficient standard to confidently estimate a relative treatment effect. A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information."*<br><br>Suggested wording:<br><br>*"~~In summary, although these methods are often presented as the only way of quantifying a relative treatment effect, this does not mean that the method will be of sufficient standard to confidently estimate a relative treatment effect. A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information~~"*<br><br>Reason for change:<br><br>Such methods should not be declared upfront to be insufficient or problematic and any statement that downgrades specific methods should be removed. Instead, methods should be described along with their associated assumptions, strengths and limitations. | No change; it is important to highlight that these methods may be of an insufficient standard. |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Sebastian Vinzens, M.Sc. | 24 | 854, 868 | Original wording:<br><br>"A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information." | That you for your comment. We recognise that full IPD may not be available for all comparator studies, |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| SKC Beratungsgesellschaft mbH | | | and<br><br>"However, in the context of JCAs, it is likely that estimation of the treatment effect in the population of the 'source' (IPD) study is of interest."<br><br>Comment:<br><br>Since it is highly unlikely that HTD will have access to IPD of multiple studies to generate data through evidence synthesis, methods of evidence synthesis that require IPD should in no case be considered the required methodological standard for JCA. As mentioned above, the practical feasibility of methods to be applied in a certain case, rather than the hypothetical highest possible methodological standard, should be the driving factor for the acceptance of data. | which is why methods such as MAIC and STC are described in the guideline. However, the quoted sentences highlight key limitations of these methods, which must be acknowledged in the JCA report if such methods are used to estimate treatment effects. |
| Mihai Rotaru, EFPIA | 26 | 922-924 | **Assessment of comparisons based on non-randomised evidence**<br><br>Current wording:<br><br>All commonly encountered sources of evidence outside of RCTs are non-randomised (i.e., single-arm trials, cohort studies, case-control studies, other observational studies, and the use of historical controls).<br><br>Proposed wording:<br><br>~~All commonly encountered sources of evidence outside of RCTs are non-randomised (i.e., single-arm trials, cohort studies, case-control studies, other observational studies, and the use of historical controls).~~ | We do not think that the proposed change represents "a clearer description of the issue".<br><br>The problem does not only occur in the case of rare diseases and highly targeted patient populations. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **Evidence packages based on single-arm trials and other non-randomised studies (i.e., historical controls, case-control studies, observational studies, cohort studies) are being more commonly encountered by Health Technology Assessment (HTA) agencies, particularly for treatments for rare diseases and highly targeted patients populations[1].**<br><br>[note: ~~strikethrough~~ indicate suggested deleted text and **bold** recommended added text].<br><br><u>Rationale:</u><br><br>The proposed rewording will provide the assessor/co-assessor with a clearer description of the issue that needs to be addressed, which is the potential lack of RCT data for interventions that have been developed to treat rare diseases or highly targeted patient populations.<br><br><u>References:</u><br><br>1. Patel D, Grimson F, Mihaylova E et al. Use of external comparators for Health Technology Assessment submissions based on single-arm trials. Value in Health 2021, 24 98): 1118-1125 | |
| S. Walleser Autiero,<br><br>Medtronic | 26 | 926-928 | This statement is true however it is important to note that many of these sources of data/studies provide evidence that cannot be generated from RCTs. In particular RCTs have low external validity; the treatment effect is an estimate in a very specific population (efficacy) while the non-randomised studies help us understand the effect in a broader more generalisable population (effectiveness). | Thank you for your comment. We acknowledge that external validity of RCTs can be a concern; however, we consider this issue to be beyond |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | the scope of this guideline. |
| Bayer | 26 | 929-930 | Specific comment:<br><br>The general statement regarding careful consideration of validity in case of non-randomised evidence needs a clarification that the assessment of "validity" needs to be limited to the description of factors linked to validity and has to avoid any value judgement. This is a consequence of Regulation (EU) 2021/2282 (Art. 9.1) which makes clear, that the JCA has to refrain from any value judgement. | We refer to our general response to the public consultation of D4.3.2. |
| EFSPI | 26 | 946-8 / 6.1 | Current Wording<br>"These adjustment methods require access to the full IPD information. Aggregated data alone are not sufficient to reliably estimate treatment effects."<br><br>Suggested Wording<br>These adjustment methods require access to the full IPD information, although a MAIC may still be considered in situations where aggregate data can be matched to IPD. Aggregated data when matched (/compared) to other aggregate data alone are not sufficient to reliably estimate treatment effects.<br><br>Reason for Change<br>In general access to IPD is limited by study protocols (patient consent), local, federal or EU General Data Protection regulation. Population-adjusted methods using IPD for one of the trials can generate robust results, and should not be ruled out categorically. | MAIC should generally only be applied in connected networks, although the possibility of unanchored MAIC is discussed in Section 5. Therefore, this addition is not useful. |
| S. Walleser Autiero,<br><br>Medtronic | 26 | 946 | Access to full IPD information is an unrealistic expectation and should not be required in all circumstances. | We agree; in the case of connected networks, access to full IPD information is not |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | required (see Sections 4 and 5). |
| Dr. Norbert Gerbsch<br><br>for IGES Institut GmbH and HealthEcon AG | 27 | 994-1005 /<br><br>Section 6.2.1. | **Comment:**<br><br>The fact that ATE and ATT are proposed as estimands to check the inferential goal is welcomed as proposed in recent years to one of the national HTA bodies. | Thank you for your comment. |
| Sebastian Werner<br><br>vfa | 27 | 989-993 | *"Cut-offs for acceptable absolute standardised difference vary (0.1–0.25) [51]. Therefore, the final conclusion regarding the balance assumption would be left to MS for absolute standardised differences <0.25; if any absolute standardised difference is ≥0.25, violation of the balance assumption should be stated. Doubly robust methods combining propensity scores and outcome regression can be used to reduce bias arising from residual covariate imbalance after matching or weighting."*<br><br>The JCA report should evaluate the validity of submitted comparisons. Therefore, it is necessary that the guidance reflects on the acceptability of methodological approaches, including on the scientific conclusion regarding the balance assumption for absolute standardised differences <0.25. The notion that the assessment would be left to MS is not sufficient in the context of the JCA. The vfa would like to point out, that the JCA report would be inconclusive on the validity assessment, if no decision criterium for <0.25 is stated. The vfa recommends clarifying that if any absolute standardised difference is <0.25, no violation of the balance assumption should be stated. Member States might deviate from this assessment of the JCA report, after considering national complementary clinical analyses. | We maintain that cut-offs for acceptable absolute standardised difference could vary, and should be left to MS evaluation. |
| Prof. Matthias P. Schönermark, M.D., | 27 | 990 | Original wording: | We maintain that cut-offs for acceptable absolute |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Ph.D., Sebastian Vinzens, M.Sc.<br><br>SKC Beratungsgesellschaft mbH | | | "[…] the final conclusion regarding the balance assumption would be left to MS for absolute standardised differences <0.25;"<br><br>Comment:<br><br>As mentioned above, according to HTA Regulation (EU) 2021/2282 Article 8 (6) sentence 2 "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, a specific cut-off value for an acceptable absolute standardised difference should be agreed upon by all MSs. | standardised difference could vary, and should be left to MS evaluation. |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Dr. Lydia Frick<br><br>SKC Beratungsgesellschaft mbH | 20, 24 | 678, 831 | Original wording:<br><br>"By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable" and<br><br>"Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, which is not the goal of the JCA."<br><br>Comment:<br><br>Depending on the number and the extent of the PICO(s) there will likely be constellations in which unanchored ITCs are the only available evidence complying with the requirements according to the respective PICO(s). Clinical evidence of the highest possible evidence level must not be demanded for each PICO as most likely multiple and/or very complex clinical trials would have to be performed to fully comply with the requirements according to every PICO. However, this does not | Thank you for your comment. We recognise that high quality RCT evidence may not be available to inform all comparisons of interest in JCAs. The guideline does not state that unanchored ITCs will not be considered in JCAs. Nonetheless, unanchored ITCs are subject to considerable limitations, which must be acknowledged in the JCA report, regardless of whether or not such a comparison represents |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | seem feasible in practice in terms of the number of patients to be recruited, delay in the generation of clinical evidence and thus in making new drugs available, required resources etc. Therefore, all efforts made by HTDs to generate evidence complying with the requirements of the respective PICO(s) should be appreciated. Methodological limitations should be accepted and should not be the reason why data presented in the dossier is not considered for the JCA. | the best available evidence. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | General | | The real-word data are increasingly used for the assessment of health technologies. Some organisations are working on this topic (e.g., NICE on the real-world evidence framework; CADTH on the real-world evidence for decision-making in the rare disease context). Moreover, it exists methods for the inclusion of real-word evidence in network meta-analysis (see below useful references suggested). Therefore, it would be perhaps useful to cover the methods related to the use of real-word evidence.<br><br>Useful references<br>• Jenkins DA, Hussein H, Martina R, Dequen-O'Byrne P, Abrams KR, Bujkiewicz S. Methods for the inclusion of real-world evidence in network meta-analysis. BMC Med Res Methodol. 2021 Oct 9;21(1):207. doi: 10.1186/s12874-021-01399-3. PMID: 34627166; PMCID: PMC8502389.<br>Lin Z, Zhao D, Lin J, Ni A, Lin J. Statistical methods of indirect comparison with real-world data for survival endpoint under non-proportional hazards. J Biopharm Stat. 2022 Jun 8:1-18. doi: 10.1080/10543406.2022.2080696. Epub ahead of print. PMID: 35675418. | These methods are included in Section 6; see also our general response to the public consultation of D4.3.2. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | General | | The missing data are common during the health technology assessment. They can be related to the presence of publication bias, the outcomes of interest that were not measured, the outcomes of interest that were not reported, and the unavailability of some important independent variables (e.g., prognostic factors). Therefore, it would be helpful to inform the assessors on how to deal and report the presence of missing data during the health technology assessment. | We added a new Section 3.4 regarding missing data. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Useful reference :<br>• Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.<br>Mavridis D, White IR. Dealing with missing outcome data in meta-analysis. Res Synth Methods. 2020 Jan;11(1):2-13. doi: 10.1002/jrsm.1349. Epub 2019 Jun 9. PMID: 30991455; PMCID: PMC7003862. | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | General | | The effect sizes were differently reported in the trials making that their calculation or transformation into the same type of effect size is challenging for assessors when they want to combine them. So, it would be useful to inform the assessors on when it is appropriate to do a transformation of effect sizes.<br><br>Useful reference :<br>Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-analysis. Chichester, UK: Wiley, 2009 | The assessors do not have to combine effect sizes. The HTDs have to do this and the assessors have to assess whether the transformations have been made adequately. These technical details are not described in this Guideline. We refer to the references in D4.3.1 and D4.3.2. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | General | | Being the importance of the evidence quality in the health technology assessment for a decision-making, it would be interesting to add a section on the assessment of evidence quality in network meta-analysis and on how it must be presented. GRADE working group has published a few papers on this topic.<br><br>Useful references:<br>• Yepes-Nuñez JJ, Li SA, Guyatt G, Jack SM, Brozek JL, Beyene J, Murad MH, Rochwerg B, Mbuagbaw L, Zhang Y, Flórez ID, | See Sections 4.2 and 4.3 |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Siemieniuk RA, Sadeghirad B, Mustafa R, Santesso N, Schünemann HJ. Development of the summary of findings table for network meta-analysis. J Clin Epidemiol. 2019 Nov;115:1-13. doi: 10.1016/j.jclinepi.2019.04.018. Epub 2019 May 2. PMID: 31055177.<br>• Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, Hazlewood GS, Alhazzani W, Mustafa RA, Murad MH, Puhan MA, Schünemann HJ, Guyatt GH; GRADE Working Group. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol. 2018 Jan;93:36-44. doi: 10.1016/j.jclinepi.2017.10.005. Epub 2017 Oct 17. Erratum in: J Clin Epidemiol. 2018 Jun;98 :162. PMID: 29051107.<br>• Brignardello-Petersen R, Murad MH, Walter SD, McLeod S, Carrasco-Labra A, Rochwerg B, Schünemann HJ, Tomlinson G, Guyatt GH; GRADE Working Group. GRADE approach to rate the certainty from a network meta-analysis: avoiding spurious judgments of imprecision in sparse networks. J Clin Epidemiol. 2019 Jan;105:60-67. doi: 10.1016/j.jclinepi.2018.08.022. Epub 2018 Sep 22. PMID: 30253217.<br>• Brignardello-Petersen R, Mustafa RA, Siemieniuk RAC, Murad MH, Agoritsas T, Izcovich A, Schünemann HJ, Guyatt GH; GRADE Working Group. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. J Clin Epidemiol. 2019 Apr;108:77-85. doi: 10.1016/j.jclinepi.2018.11.025. Epub 2018 Dec 5. PMID: 30529648.<br>• Brignardello-Petersen R, Izcovich A, Rochwerg B, Florez ID, Hazlewood G, Alhazanni W, Yepes-Nuñez J, Santesso N, Guyatt GH, Schünemann HJ; GRADE working group. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. BMJ. 2020 Nov 10;371:m3907. doi: 10.1136/bmj.m3907. PMID: 33172877.<br>Brignardello-Petersen R, Florez ID, Izcovich A, Santesso N, Hazlewood G, Alhazanni W, Yepes-Nuñez JJ, Tomlinson G, Schünemann HJ, Guyatt GH; GRADE working group. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. BMJ. 2020 Nov 11;371:m3900. doi: 10.1136/bmj.m3900. PMID: 33177059. | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| GSK | General | General | Suggest adding some flowcharts to this document to complement the text and indicate the order of proposed tests or actions. For example, in Section 3.2.2, could be useful to have flowchart showing when studies can be pooled and which test should be applied for a large number of studies. | A flowchart applicable to all data situations would be very complicated. We think that the given descriptions are more helpful. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | General | | Takeda appreciates the objectivity and completeness of this practical guidance which presents the possible methods of conducting ITCs and enables flexibility for the most appropriate method selection given the circumstance – 'the right tool for the right situation' approach.<br><br>However, we believe this draft practical guide would be further strengthened by providing JCA assessors and HTDs further guidance on the context specific factors (for example., data availability) that should be considered when assessing methods(s) used for direct and indirect comparison. | Thank you; we think that context-specific factors are sufficiently included in the Guideline. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | General | | The practical guideline could be further strengthened by including guidelines on the following:<br><br>Best practice methods for treatment comparisons of:<br><br>- New treatment modes (e.g., cell and gene therapies)<br>- How to perform indirect comparisons that include randomised and observational study data in one meta-analysis | This is out of the scope of this Guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Additional guidance on:<br><br>- Minimum reporting requirements<br>- Instructions on the format and content (tables and figures) expected<br>- Templates for data presentation/ network diagrams<br><br>This would enable HTDs to plan and anticipate the required analysis and prepare a comprehensive dossier and reduce the assessor requests for additional analyses after the JCA has commenced. | |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | General | | No information is provided on how to understand or use large confidence intervals in the context of NMA. | There is no need to interpret confidence intervals in the context of NMA in a different way than usual. |
| Mihai Rotaru, EFPIA | General | | The practical guideline Direct and Indirect Comparisons should include additional guidance on the following:<br><br>- Minimum reporting requirements<br>- Instructions on the format and content (tables and figures) expected<br>- Templates for data presentation/ network diagrams<br><br>This would ensure the HTD submits results in a format that aides the EU JCA assessor in their report write-up and reduces need for additional analysis requests. | This is out of the scope of this Guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| Roche | general | | We recommend introducing explicit wording that context specific considerations may be needed. The current draft guideline implicitly acknowledges this need by asking that both statisticians and healthcare professionals should be consulted (line 134) and that "there is an element of subjectivity in the assessment of many assumptions" (line 118). | We think that context-specific factors are sufficiently included in the Guideline. |
| Mihai Rotaru, EFPIA | 16 | 540 | **Requirements for reporting – Bayesian analysis and interpretation**<br><br>EFPIA would like to reiterate a comment provided to the stakeholder consultation to EUnetHTA deliverable 4.3.2 regarding Bayesian analysis and interpretation in the context of the previous EUnetHTA methodological guideline (2015) (extraction below).<br><br>*"In contrast to the frequentist approach, the output of a Bayesian approach can be interpreted in terms of probabilities. (26) Thus Bayesian approaches facilitate the computation of useful measures such as the probability of treatment superiority, the probability that a clinically meaningful difference exists, or the probability of clinical equivalence.(31)"*<br><br>We recommend the insertion of the following statement (or similar) in this practical guideline to support assessors when interpreting indirect treatment comparisons and network meta-analysis from a Bayesian approach in accordance with international best practice guidelines.[1] | We repeat our former reply:<br><br>"We aimed at keeping the Guideline as short and easy as possible. The Guideline is not a statistical textbook (see Objective) and very general statistical and philosophical deliberations are out of the scope." |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Reference:<br>1. Jansen, JP. 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 Health. 2011;14(4)417-28. | |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 26-28 | 6 Assessment of Comparisons Based Upon Non-Randomised Evidence (920-1028) | For the assessment of non-randomized evidence, we might want to mention some of causal inference methodology such as g-computation etc. and recently developed causally interpretable meta-analysis methodology. | G-computation and other causal methods are mentioned in D4.3.2. It is not required to repeat this here. |
| GSK | 26-28 | Section 6.2 | How about Inverse Probability Treatment Weighting (IPTW) as a comparison technique? | IPTW and other causal methods are mentioned in D4.3.2. It is not required to repeat this here. |
| Sebastian Werner<br>vfa | 7 | 128- 140 | Please add further guidance and external references on systematic literature reviews. | See D4.3.2 |
| François Houÿez, EURORDIS | 7 | 149-153 | To best assess possible effect modifiers, it is important to ask trial subjects or representatives of patients (eg Members of Community Advisory Boards) their own experience with the trial. Eurordis is aware | While we agree with your points this is out of the scope of this Guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | of clinical trials where end results did not reflect the reality (true drop-out rate explanation different from clinical study report). | |
| EFSPI | 8 | 186-187 | "The evaluation of similarity should also consider methodological factors that should not differ<br>substantially between studies"<br><br>Estimand, in particular the way intercurrent events are dealt with, should also be mentioned. | See Section 5.3 |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 9 | Section 3.2.2 | Takeda recommends that the section of Assessment of homogeneity also include Bayesian methods to be consistent with the methodological guidance, Methods of Direct Comparison (Section 4.1) which do include Bayesian perspective. | Section 3.2.2 also refers to Bayesian (network) meta-analyses. |
| Mihai Rotaru, EFPIA | 9 | Section 3.2.2 | **Assessment of homogeneity – Bayesian perspective**<br><br>EFPIA believes that similarly to the assessment of consistency (Section 3.2.3), this section would also benefit from the addition of the potential Bayesian perspective alongside the Frequentist approach to assessment (Q and $I^2$). This is especially relevant given Methods of Direct Comparison (Section 4.1) includes Bayesian methods. | Section 3.2.2 also refers to Bayesian (network) meta-analyses. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 10 | 303-306/ 3.2.3 | It would be helpful to add a few examples of methods assessing the consistency for a given network structure. | The available time is insufficient to describe examples. |
| ISPOR – The Professional Society for Health Economics | 10 | 274-287 | For the assessment of homogeneity, recent literature includes fixed effect (or the common effect) model, fixed effects model and random effects model. Should the guideline include discussions on fixed effects model, and choices among the three models? Furthermore, | See Section 3.3 in D4.3.2. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| and Outcomes Research | | | should the guideline discuss the assessment and approaches for small study effect, publication and reporting bias? Right now, it seems to be completely ignored, although that may not be intentional. | |
| Matias Olsen, EUCOPE | 10 | 288-295 | We would suggest to add reporting on both the fixed-effect and the random-effect model so that the public can appreciate how big the difference between those two methods is. | It is not helpful to present the results of both models in all cases. |
| Sebastian Werner<br><br>vfa | 10 | 274 – 278<br><br>288 – 291 | *Requirements for reporting*<br><br>The evaluation of homogeneity is presented as a decision tool to decide regarding the pooling of the included studies. However, heterogeneity can also be seen as a measure of uncertainty. Therefore, guidance on how to deal with increased heterogeneity would be desirable. | See Section 3.2.2. |
| Matias Olsen, EUCOPE | 12 | 374-402 | Consider including fractional polynomial NMA methods. | See Section 4.3.3. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 12 | 374-405/3.3 | The publication bias can seriously affect the pooled intervention effect in meta-analysis. So, it would be very interesting to address this point in the present document.<br><br>Useful references for the pairwise meta-analysis (simple):<br>• Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics. 2018 Sep;74(3):785-794. doi: 10.1111/biom.12817. Epub 2017 Nov 15. PMID: 29141096; PMCID: PMC5953768.<br>• Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000 Jun;56(2):455-63. doi: 10.1111/j.0006-341x.2000.00455.x. PMID: 10877304.<br><br>Useful references for the network meta-analysis: | See Section 3.2 in D4.3.2. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | • Mavridis D, Welton NJ, Sutton A, Salanti G. A selection model for accounting for publication bias in a full network meta-analysis. Stat Med. 2014 Dec 30;33(30):5399-412. doi: 10.1002/sim.6321. Epub 2014 Oct 15. PMID: 25316006.<br>Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. Stat Med. 2013 Jan 15;32(1):51-66. doi: 10.1002/sim.5494. Epub 2012 Jul 17. PMID: 22806991; PMCID: PMC5410995. | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 13 | 406-413/3.3 | In this section, it would be helpful to know how to deal and report the information related to the publication bias. | See Introduction and Section 3.2 in D4.3.2. |
| Mihai Rotaru, EFPIA | 14 | 417-419 | Current wording:<br>Standard approaches for meta-analyses according to the fixed-effect model (with the assumption of a common effect in all included studies) are given by the inverse variance method for continuous data and the Mantel–Haenszel method for binary data [3].<br><br>Comment:<br>Since different methods may be required for binary data (such as GLM and GLMM), and different metrics may be required (e.g. ORR, RR), we recommend broadening the language to allow for these methods/measures as appropriate. | See the sentence directly behind that:<br>"Other useful methods are available …" |
| EFSPI | 14 | 418-419 | Should we add GLM and GLMM to address binary data and different metrics (ORR, RR)? | See Section 4.1 in D4.3.2. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| S. Walleser Autiero,<br><br>Medtronic | 15 | 492-494 | The requirement for reporting does not include requirements for model choices when the set of studies is small although this was discussed at length in the preceding text. | The model choice is more difficult in the case of very few studies, but the reporting requirements are the same. |
| EFSPI | 16 | 540 / 4.2 | The primary outcome from an evidence synthesis done for JCA is the set of relative effect estimates for the new intervention vs all relevant comparators (based on the full body of evidence), along with the associated uncertainties quantified for example as confidence intervals or credible intervals. The reporting requirements should list this output explicitly. | Thank you. We have included an additional reporting requirement. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 16 | 540-556/4.2 | In this section, it would be interesting to report:<br>• the importance to describe the process used for the identification of all relevant intervention comparisons considered in the indirect comparison<br><br>• report the information related to the assessment of publication bias<br>Useful references:<br>• Mavridis D, Welton NJ, Sutton A, Salanti G. A selection model for accounting for publication bias in a full network meta-analysis. Stat Med. 2014 Dec 30;33(30):5399-412. doi: 10.1002/sim.6321. Epub 2014 Oct 15. PMID: 25316006.<br>• Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. Stat Med. 2013 Jan 15;32(1):51-66. doi: 10.1002/sim.5494. Epub 2012 Jul 17. PMID: 22806991; PMCID: PMC5410995. | The relevant comparators are mainly a part of the PICO, which is out of the scope of this Guideline.<br><br>For the issue of publication bias we refer to the Introduction and Section 3.2 of D4.3.2. |
| Mihai Rotaru, EFPIA | 16 | 529 | Current wording: | Thank you. This has been added. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Further considerations must be given to the number and heterogeneity of studies informing each contrast, number of events (rare versus common events), scale (OR, RR, HR, or MD) […]<br><br>Proposed wording:<br>Further considerations must be given to the number and heterogeneity of studies informing each contrast, number of events (rare versus common events), scale (OR, RR, HR, **RD** or MD) […]<br><br>Rationale:<br><br>Risk Difference (RD) could be of particular importance in certain disease areas. | |
| Mihai Rotaru, EFPIA | 16 | 556 | Current wording:<br><br>Convergence of the Markov chains<br><br>Comment.<br><br>It would be helpful to have a description of the minimum requirement for presenting convergence of the Markov chains. | Such technical details are out of the scope of this Guideline. |
| Sebastian Werner<br><br>vfa | 8,9 | 208-217<br><br>222-228 | The Guideline should provide more guidance on the acceptability of e.g., heterogeneity, matching method, number of missing confounders, number of underpowered subgroups, etc. for validity reasons. | We refer to our general response to the public consultation of D4.3.2. |
| Richard Birnie<br><br>Lumanity | 17-18 | Section 4.3.2 and 4.3.3 | Sections 4.3.2 and 4.3.3 describes alternative approaches to evidence synthesis for those cases where the proportional hazards assumption does not hold. There is a trade-off between the different approaches that would be useful to call out explicitly. There are broadly two applications of the outputs from evidence synthesis. Firstly, interpretation of the relative effects themselves to understand | We added the reference in Section 4.3.3. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | comparative efficacy and safety between treatments, and secondly, using the relative effects as inputs to a cost-effectiveness model.<br><br>Comparison of restricted mean survival time is easier to interpret than the outputs from flexible survival models but, to our knowledge, the difference or ratio of RMST would not be applicable in most common cost-effectiveness models. Conversely, flexible survival models are more difficult to interpret clinically but easier to use in a cost-effectiveness model. It would be helpful to add a sentence or two to each section to identify the different applications and highlight where the different methods may be more appropriate.<br><br>A recent publication by Freeman et al discusses a range of possible methods for NMA of time to event outcomes in the presence of non-PH. All of the approaches mentioned would be appropriate in some situations and should be considered, rather than giving undue emphasis to any one method. The paper is freely available here https://pubmed.ncbi.nlm.nih.gov/35044255/ | |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 18 | 609 – 614 | In addition to AIC, BIC and DIC, it would be useful to discuss the recently developed WAIC and LOOIC statistics. | Thank you for this suggestion – we suggest AIC or BIC and others as possible examples but it is not intended to be an exclusive list of possible methods. |
| GSK | 21 | 704-706 | Suggest guidance regarding appropriate methods for quantifying bias and its impact (e.g., through QBA methods) be provided. | There are many methods papers describing how this can be done and therefore not efficient or |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | feasible to repeat these here. |
| EFSPI | 24 | 852-862 | The section refers to IPD-based methods (e.g., propensity score (PS) matching), however, IPD-based methods could be either model-based methods or PS methods. It would be good to distinguish between the 2 types of methods and comment on their use (e.g., is adjustment trough model-based method can also be considered in this guidelines). Also, can you comment on different known propensity score methods (matching, inverse probability weighting, stratification)? | More guidance is provided in the references of Section 6. It is not the intention to provide extensive descriptions of the methodology within this practical guideline for assessors. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 26 | Section 6 | Section 6 discusses using non-randomised evidence for comparators where direct evidence is not available. Although single-arm trials belong to this section, single arm trials also have unique features compared to other observational studies. Single arm trials are valuable in new drug development on rare disease and highly targeted patient populations. In these cases, RCTs could be either unethical or unpractical. In addition, single arm trials can also be used for earlier phase trials to understand whether patient would benefit from the new treatment [1]. Although with its limitations, an unanchored MAIC or STC, may be used for single-arm trials with a thorough discussion to the limitations of these approaches.<br><br>In Section 6.2.1, three assumptions are discussed for the use of an unanchored comparison. It is implausible for a single arm trial to meet the positivity assumption. Therefore, when discussing how to consider non-RCT for comparison, a separate section on single arm trials is suggested for clarity.<br><br>[1] Patel D. et al. (2021) Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials. Value in Health; 24 (8); 1118-1125 | From a methodological point of view, there is no justification to separate these two types of evidence when checking assumptions validity. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 27 | 976 - 984 | Section: Checking the overlap assumption<br><br>Comment:<br><br>Suggest adding some text that explains how trimming may be used where complete overlap cannot be achieved. Such as<br><br>"Restricting the study population through trimming provides a way to reduce the variance of weighted estimates and to reduce bias from confounding in the tails of the PS distribution."<br><br>Rationale:<br><br>This section raises the use of trimming but does not explain what it is and how it is used/applied. | Trimming is mentioned in D4.3.2. It is not required to repeat this here. |
| | | | | |
| | general | | Prespecification of methods for evidence synthesis<br><br>There is common agreement that pre-specification of analysis is good statistical practice for evidence generation and also evidence synthesis. Statistical Analysis Plans should always be finalized before the analyses are conducted. For HTA purposes, the analysis planning is determined by the PICOs, which might not be completely addressed in the clinical study protocols and which will be defined during the EU HTA Scoping process.<br><br>We recommend that it is clarified in the guideline that Pre-specification, therefore, may not only refer to analyses specified in the protocol and/or SAP of a trial, but also to evidence synthesis and complementary statistical analysis that a sponsor may specify in a separate SAP due to the request by the HTA bodies in the scoping phase. | We added an explanation for prespecification in the context of JCA in the first paragraph of Section 3 |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The traditional (regulatory) definition of pre-specification refers to formally defined SAP that can only be modified prior to study read-out. However, it may be helpful to acknowledge that in the case of the EU HTA guidelines, the "pre-specification" refered to is in the form of defining an HTA SAP that is in response to the PICOS specified (and hence, is defined post database lock)<br><br>In considering the implementation of this for the HTA regulation, we strongly recommend further consideration of the opportunity for interaction between the Assessor/Co-Assessor and the HTD, in order to foster alignment on sound choices of data sources and appropriate statistical methods for ITC. | |
| Sebastian Werner<br>vfa | General | | Clear and unambiguous definitions are needed for the words 'a priori' and 'pre-specified' to avoid any misinterpretation by the assessors and co-assessors. We request definitions for these words are provided in Section 1.1. The vfa would like to point out that analyses requested by the HTA authorities regarding specific PICO questions, might 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). | We added an explanation in the first paragraph of Section 3. |
| Mihai Rotaru, EFPIA | General | | EFPIA request further definitions for key concepts and assumptions described in the practical guide are added to Section 1.1 (e.g., similarity, homogeneity, consistency). The definitions should be sourced from textbooks and best practice guidelines and referenced accordingly. For illustration, the ISPOR Task Force guide1 'Interpreting indirect comparison and network meta-analysis for health-care decision making' recommends the term 'network meta-analysis' should be used when the evidence base consists of two RCTs connecting more than two interventions. If the network consists of at least one | 1) We refer to the Section "Scope and Terminology" of D4.3.2. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | closed loop, the term MTC could be used and for an open-loop network the term ITC would be appropriate to use.<br><br>We feel it is important to ensure consistency in terminology is used throughout the document as well as across EUnetHTA 21 methods guidelines (e.g., D4.3.2 Direct and Indirect Comparison) to avoid any misunderstanding or misinterpretation amongst assessors and co-assessors.<br><br>We would like to see clear and unambiguous definitions are provided in Section 1.1 for the words 'a priori' and 'pre-specified'.<br><br>Finally, if HTA-related analysis plans are required, for example for trial analyses or for indirect comparisons, it would be helpful to understand when these should be finalised to be considered a priori by the EU assessors and co-assessors.<br><br>Reference:<br><br>1. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC. 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 Health. 2011 Jun;14(4):417-28. | 2) We added an explanation in the first paragraph of Section 3. |

**EUnetHTA 21 Public Consultation Comments and Responses**
**of D4.3.1 on Practical Guideline Direct and Indirect Comparisons**

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| GSK | General | General | Is there any specific practical guidance on how to approach ITC with results from single arm trials? | See Sections 5.5 and 6 |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 7-13 | 3 General Considerations (127-413) | In addition to randomized trials, it is unclear whether JCAs can use results from single-arm trials, historical controls, and non-randomized multiple arm trials to strengthen the evidence synthesis. | See Sections 5.5 and 6 |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 20-25 | 5 Assessment of Population-Adjusted Methods (650-919) | It is unclear whether Bayesian methods are considered acceptable for population-adjusted indirection comparison. | See Section 3.4 in D4.3.2. |
| GSK | 12-13 | Section 3.3 | During the consultation, will the assessor recommend a preferred approach for the HTD to consider, in circumstances where the exchangeability assumptions are violated? | It is not within the scope of this guideline to describe the individual interactions between HTDs and HTAbs. |
| GSK | 14-15 | Section 4.1.3 | Are there any exceptions to this guidance? For rare disease and oncology indications, it is rare to have more than five studies. | Section 4.1.3 explains how to perform meta-analyses in the case of less than 5 studies. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 15-16 | 509-539/4.2 | In this section, it has been stated that important domains in assessment of the credibility of indirect comparison methods are described. However, the information related to the identification of all relevant intervention comparisons is missing. This step should be based on a systematic review and a consultation of content experts confirming the relevance of intervention comparisons identified. Therefore, it would be helpful to remind, here, the importance to | The definition of the research question (PICO) is out of the scope of this guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | describe the process used for the identification of all relevant intervention comparisons considered in the indirect comparison. | |
| | | | | |
| Matias Olsen, EUCOPE | 7 | 127-140 | These general considerations seem to forget the pooling of observational studies with research clinical trials that is mentioned in section 6. We suggest also referring to it here. | We changed the word “trials” to “studies”. |
| BAH | 7 | 142-148 | In order to be able to ensure the complete evidence, a systematic literature search must be carried out for the evidence synthesis (see Deliverable 4.3.2 Methods Guideline. | See lines 129-130 |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 7 | 135-138 | From the patient’s perspective, it is difficult to trade-off relative treatment effects among multiple outcomes. In addition to relative effects, should treatment-specific absolute effects be also provided as additional information? In addition, pooled relative treatment effects are conditional effects (i.e. weighted average of trial-specific effects), generally not equal to the marginal or population-averaged effects. This is particularly an issue for non-collapsible effect measures such as the odds ratio or hazard ratio. Absolute effects for all outcomes over all studies included in a research synthesis are important for patients and caregivers to trade-off potential benefits and harms, while one cannot make such decisions based solely on relative effects. For pooling absolute effects, we understand that there is some debate on the risk of break randomization when some studies use unequal group size randomization. However, as long as we use a study-specific weight (as compared to study-treatment-specific weights) for all intervention groups within a study, there is no such risk of break randomization. One can easily validate this argument for a direct comparison meta-analysis. | This is a misunderstanding.<br><br>With relative and absolute treatment effects we do not mean relative (e.g., RR) and absolute effect measures (e.g., RD). As explained in the brackets, we mean a comparison of 2 groups vs. Results from one group only. |
| GSK | 7 | 154-156 | Please clarify if the requirements for reporting should also include a summary of the studies that don’t match the established PICO and the reasons for this. | We refer to EUnetHTA 21 D5.2 *“Assessment Report Guidance”* |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Roche | 8 | 184-5 / 3.2.1 | *"an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation."*<br><br>The statement seems confusing as it mixes the terms 'I = Intervention' and 'C = Comparator' from the PICO. The exact meaning should be clarified.<br><br>The new intervention being assessed (identified via 'I' in the PICO) will not yet have received EMA marketing authorisation at the time of JCA. And the selection of comparators ('C') should follow a structured approach ensuring a final list of comparators that is reasonable and concise. The priority should be given to established licensed medicines with published robust clinical data, followed by those recommended in European clinical guidelines. | We agree and deleted the sentence. |
| Sebastian Werner<br><br>vfa | 8 | 218-221 | *"It is essential that the process used to identify relevant effect modifiers is comprehensive and transparently reported. This process should include a comprehensive review of the literature and consultation of healthcare professionals with knowledge of the disease area. The set of all potentially relevant effect modifiers should be reported in the submission."*<br><br>The request of "all potentially relevant effect modifiers" should be changed to "known effect modifiers". The list of effect modifiers should be limited to clearly known effect modifiers and not to "all potentially relevant effect modifies", as this definition might include effect modifiers that are not relevant. | We think that the current wording is adequate. The definition does not include effect modifiers, which are clearly not relevant. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The vfa supports the proposal of involving health care professionals in the identification process of known effect modifiers. The inclusion of clinician's perspective is necessary to evaluate the importance of effect modifiers for e.g., treatment decisions. However, the compressed timelines of EU-HTA are a big hurdle for the identification process and the involvement of health care professionals during the dossier preparation. Further, large number of PICO questions make the process not feasible. Hence, the identification process of effect modifiers needs to be prepared well in advance of the JCA based on JSC. Therefore, JSC consultations on this aspect need to be offered on a regular basis with sufficient capacity to satisfy the needs of all HTD and with the possibility to discuss all other relevant aspects of the product development. In addition, opportunities for interaction between HTD and HTA bodies must be provided, including a PICO meeting. | |
| Matias Olsen, EUCOPE | 8 | 183-185 | The choice of appropriate comparators is vital for the outcome of HTA procedures. Since the JSC slots are limited, companies may not be able to get a consultation meeting for their product, meaning that they might be left in the dark when it comes to the appropriate comparators from an early stage. Furthermore, the draft guidance provided so far has not described the opportunity for HTD to participate in a scoping meeting with the Assessors, meaning that even at the start of the assessment there would be no opportunity to align on appropriate comparators for the assessment.<br><br>It is of utmost importance that the process for choosing an appropriate comparative therapy is very clearly defined to minimise the chance of inappropriate comparators, that lead to reduced quality of the assessments, or to discontinuation of the procedure. The HTD should therefore be invited to participate in a scoping meeting with the assessors to discuss the draft consolidated PICO(s). As noted in comments on deliverables D7.1.1 "practical guideline on the interaction | This text part was modified; the choice of comparators is out of the scope of this Guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | between HTD and HTAb" and D4.2 "Scoping Guideline", there needs to be clear guidance on appropriate comparators, and these should be limited to products holding an EMA marketing authorisation in the indication of interest when an authorised alternative exists. Above all, the choice of comparator must be evidence-based.<br><br>Replace:<br><br>"Characterisation of the intervention (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation."<br><br>With:<br><br>"Characterisation of the intervention (e.g., dosage or application, concomitant treatments): an intervention ~~should~~ **can, subject to existing guidance at the time of the assessment,** be considered as a comparator even if it has not yet been granted European Medicines ~~Authority~~ **Agency** (EMA) marketing authorisation. **However, as a general rule it must have a marketing authorisation for that indication and line of treatment and above all the choice of comparator must be based on available clinical evidence.**" | |
| GSK | 8 | 183-185 | The I in PICO is the Intervention to be assessed, the C in PICO are the comparators which the Intervention is to be compared against. The current text suggests the intervention should be considered a comparator but we think the intent was to note an investigational product that has not yet been granted EMA marketing authorization could be considered a comparator for the intervention. Suggest rewording | The corresponding sentence was deleted. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | intervention in this context to e,g. health technology or investigational product. | |
| GSK | 8 | 183-185 | Please clarify why an investigational product not approved by EMA should be considered a comparator and clarify what are the implications to any assessments of relative treatment effects of the intervention if the comparator is not approved by EMA and/or the MAA is withdrawn by the sponsor. | The corresponding sentence was deleted. |
| GSK | 8 | 183-185 | What timeline should be taken into consideration for identifying possible investigational products which may not be EMA authorised but that should be used as comparators? How far into the future should one consider? Must an MAA have been submitted to EMA for an investigational product to be considered a comparator? | The corresponding sentence was deleted. |
| Mihai Rotaru, EFPIA | 8 | 183-185 | **Assessment of exchangeability – Characteristics of the intervention**<br><br>Current wording:<br><br>Characteristics of the intervention (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation.<br><br>Proposed wording:<br><br>Characteristics of the intervention (**incl. comparators**): dosage, application, concomitant treatments, line of treatment, etc. a~~n intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation~~ | The corresponding sentence was deleted and the remaining text was modified. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively]<br><br>Rationale:<br><br>EFPIA is unclear why this sentence is included since it doesn't explain the characteristics of the intervention. Whilst the intervention under assessment in the JCA will not yet have EMA marketing authorisation in the submitted indication, comparators are specifically those outlined in the PICO(s). Treatments that are new technologies that have not yet received marketing authorisations will not be relevant comparators as they will not represent current European standard of care.<br><br>EFPIA also believes that factors such as the line of treatment can be an important factor that should be added. | |
| François Houÿez, EURORDIS | 8 | 183-185 | "not yet" implies a Marketing Authorisation Application is under review by the EMA. The guidance does not address situations where no positive opinion has been made by the EMA, but national authorisations might exist.<br><br>The European Commission, not the EMA, can grant a marketing authorisation.<br><br>EMA stands for European Medicines Agency (not authority). | The corresponding sentence was deleted. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested rewording:<br><br>2. Characteristics of the intervention (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if no marketing authorisation application has been submitted to the European Medicines Agency (EMA) marketing authorisation. | |
| Sarah Smith, Lumanity | 8 | 183/184 | Suggest it is worth clarifying whether the manufacturer only needs to look at interventions included in the PICO in the following sentence: "An intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation." | The corresponding sentence was deleted. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 8 | 183 | Current wording:<br><br>**"Characteristics of the intervention** (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation."<br><br>Suggested wording:<br><br>**"Characteristics of the intervention** (e.g., dosage or application, concomitant treatments): an intervention ~~should~~ may be considered as a comparator, ~~even if it has~~ that has not yet been granted European Medicines Authority (EMA) marketing authorization if it is recommended by pan-European clinical guidelines and is the standard of care in multiple EU countries."<br><br>Comment: | The corresponding sentence was deleted. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Medicines that do not have a marketing authorization in the EU should only be considered comparators if they are the standard of care across multiple EU Member States and are recommended by pan-EU clinical guidelines.<br><br>As there is likely to be poor evidence base for unlicensed medicines in the indication or population in question, please clarify how clinical trial participation or treatment with therapies approved and used outside the EU or off label should be considered? | |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Elisa Zavatta, M.A.<br><br>SKC Beratungsgesellschaft mbH | 8 | 183 | Original wording:<br><br>“Characteristics of the intervention (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation.”<br><br>Comment:<br><br>Firstly, the original wording does not clearly describe the characteristics of an intervention that should be taken into consideration. For clarification and completeness, the characteristics of an intervention to be considered should be precisely described. Secondly, an intervention should be considered as a comparator if it is part of the standard of care in the respective indication, as it is common practice in HTA. Drugs for which EMA marketing authorisation in the respective indication has not been granted yet are generally to be considered off-label use. Such drugs should only be considered as comparators if they are part of the standard of care, i.e., if they are routinely used in clinical practice in the respective indication. | The corresponding sentence was deleted |
| | | | | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Mihai Rotaru, EFPIA | 8 | 176-177 | **Assessment of exchangeability – Study and patient characteristics (1)**<br><br>Current wording:<br><br>Study and patient characteristics (including duration of follow-up): a list of potential effect modifiers should be drawn up a priori.<br><br>Proposed wording:<br><br>Study and patient characteristics (including duration of follow-up): a list of potential effect modifiers should be drawn up a priori **(i.e., before analyses are undertaken in the comparison).**<br><br>[note: **bold** denotes proposed inclusion]<br><br>Rationale:<br><br>EFPIA believes that the meaning of “a priori” is unclear in this context and should be described in more detail. If it refers to “before the start of the evidence synthesis” there must be a clear guidance (e.g., a JCA statistical analysis plan (SAP) or another required document). However, even if an SAP is written in advance, the underlying studies are still usually completed and there is no safeguard against data driven decisions.<br><br>If “a priori” refers to the individual studies, it should be kept in mind that a study is usually not designed for evidence synthesis purposes, so | We included a similar explanation of “a priori”. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | the list of potential effect modifiers considered might not be comprehensive. | |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 8 | 189 | Current wording:<br>**"Characteristics of outcomes** (e.g., definitions of outcomes): an *a priori* definition of what is considered sufficiently similar for each characteristic will usually be difficult. It will often also depend on what is present in the studies included;"<br><br>Comment:<br>Please illustrate what characteristics of outcomes are important to assess relative to an outcome if as stated above 'an *a priori* definition will be difficult' and in consideration of the importance of data availability.<br><br>Rationale:<br>Knowledge of the perceived importance of relationship with the outcome variable is critical in an acceptable study design. | The definitions of the considered outcome in the included studies has to be described and compared, see the example in brackets. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 8 | 181 – 182 | Current wording:<br><br>"Effect modifiers can be identified through a literature search, input from healthcare professionals and, other methods."<br><br>Suggested wording:<br>"Effect modifiers can be identified through a literature search, input from healthcare professionals, and other non-data-driven methods,"<br><br>Rationale:<br><br>The added wording provides clarity by explicitly states that effect modifiers will not be identified based on the data, and places emphasis that these must be identified *a priori.* | We think this is not required due to the clear statement in lines 176-177. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| EFSPI | 9 | 256-259/3.2.2 | The agency suggests $I^2$ boundaries for the interpretation of heterogeneity between studies. Can the agency comment on whether the synthesised evidence will be acceptable in the presence of boundaries of high heterogeneity (it is possible that after the inspection of studies and effort to eliminate heterogeneity, it will still be present)? | See the explanations in Section 3.2.2. |
| Mihai Rotaru, EFPIA | 9 | 256-259 | **Assessment of homogeneity (4)**<br><br>Current wording:<br><br>As a rough guide for the interpretation of $I^2$, the following overlapping categories were proposed [9]:<br><br>– 0–40%: might not be important.<br>– 30–60%: might represent moderate heterogeneity.<br>– 50–90%: might represent substantial heterogeneity.<br>– 75–100%: considerable heterogeneity.<br><br>Comment:<br><br>This section of the practical guideline includes proposed $I^2$ boundaries for the interpretation of heterogeneity between studies. EFPIA recommends the addition of clarity on whether synthesised evidence will be acceptable in the presence of boundaries of high heterogeneity. We think it is possible that there may be instances where this may occur, despite approaches to eliminate heterogeneity by the HTD. | See the explanations in Section 3.2.2. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 16 | 540-556 | Section: **Requirements for reporting**<br><br>Comment: | Yes, for each different PICO a different analysis is required, which means that for one PICO a fixed-effect model is appropriate and for |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | To meet the reporting requirements many judgement calls are required to be made by both the HTD in preparing the dossier and the JCA assessors in reviewing the dossier.<br><br>As detailed in the data applicability and availability documents evidence required to be presented may vary by national HTA agencies as they decide a specific PICO is more appropriate.<br><br>Takeda seeks clarification to the following questions: How will post-hoc analyses for individual Member States HTAs be handled? Will the entire approach change (e.g., random to fixed effects models)? Furthermore, will these be considered against the alpha for effect estimate? | another PICO a random-effects model. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 16 | 541-542/ 4.2 | Trials often evaluate more than one outcome of interest for a given intervention comparison. However, the meta-analysis can be possible for some outcomes but not for others from the same trial. Therefore, it would be more exact to say:<br>Determination of whether pooling of the intervention effects of interest from studies under consideration is meaningful, and justification for this determination | All reporting requirements refer to one PICO, i.e., to one outcome. |
| Mihai Rotaru, EFPIA | 26 | 947-948 | <u>Current wording:</u><br><br>A statistical analysis plan (SAP) is required to describe the methods planned to adjust for confounding."<br><br><u>Proposed wording:</u><br><br>A statistical analysis plan (SAP) is required to describe the methods planned to adjust for confounding, **whenever possible**. | 'Required' is not coherent with the use of 'whenever possible'. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | [note: **bold denotes** recommended added text].<br><br><u>Rationale:</u><br><br>Pre-specification of analyses is good statistical practice. However, it should also be noted that this isn't always possible, particularly when some of the analyses required (e.g. sub populations, sub-groups and comparators) will only be known on the publication of PICOs. | |
| Sebastian Werner<br><br>vfa | 26 | S 6<br><br>920-1028 | In section 6 it is mentioned several times that methods, analysis aspects and selection of covariates for comparisons based upon non-randomized evidence are required to be planned and described in a statistical analysis plan (SAP).<br><br>There is an agreement that pre-planning analysis in an SAP is good statistical practice. However, it is not clear whether the guidance refers to a "clinical trial SAP" or a possible "HTA SAP". The guideline should clarify that the analyses for HTA purposes can be specified in a separate "HTA SAP" to complement the clinical trial SAP. Especially those analyses that require different approaches due to the specification of the PICO research questions by HTA bodies should be specified in a complementary "HTA SAP".<br><br>Due to the timing of the future EU HTA scoping process (as described in the Guideline "Scoping process") in relation to the EU Market Authorization process, it is inevitable that not all information required for the EU HTA submission (especially those for evidence synthesis as described in this guideline) might be available to the HTD prior to the initial data release of the pivotal trials.<br><br>In this context pre-specification, therefore, 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 | It could be both: clinical trial SAP or a distinct HTA SAP. This will be updated. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | complementary analyses as well as any statistical analysis that a HTD may specify in a separate "HTA SAP", before conducting the analysis to meet a HTA body's request. The guideline should clarify that the analyses for HTA purposes can be specified in a separate "HTA SAP" to complement the clinical trial SAP. | |
| EFSPI | 27 | 1002 | "An assessment of the SAP with the propensity score methods used to adjust for confounding;"<br><br>Please precise that the SAP can be also a separate SAP from the study SAP for HTA purposes, which needs to be developed after knowledge of the PICOs, and before data analysis. | We added an explanation. |
| EFSPI | 7, 9, 10, 11, 12, 14, 15 | 165, 266 – 267, 283, 331, 386, 445, 454, 463, 465, 487 | **General comment on precision regarding number of trials (few versus large)**<br><br>We recognize it is difficult to numerically define the number of studies that are generally required for *all* assessments, given that this can be context-dependent. In that school of thought, we recommend to note that context-dependency in the text, and generally recommend that the Assessor and Co-Assessor can make more explicitly recommendations on this threshold in the context of a specific assessment - preferably in dialogue/consultation with the HTD about the availability and suitability of additional applicable data sources | We think that the Guideline makes it sufficiently clear that a context-specific assessment is required. |
| Roche | 7, 9, 10, 11, 12, 14, 15 | 165, 266 – 267, 283, 331, 386, 445, 454, 463, 465, 487 | Please give a precise number instead of "few studies" and "large number of studies" based on established rules of thumb from the scientific literature. | See Section 4.1.3 |
| S. Walleser Autiero,<br><br>Medtronic | 26 | 922-926 | Reconsider where this information is placed. The topic of historical controls should be included in the methods section (i.e., in reference to its use in single arm studies or observational registry data) | It is not entirely clear what the commentator is advising. We consider that section 6 is the right place for this information |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| ISPOR – The Professional Society for Health Economics and Outcomes Research | General | | For any research synthesis involving direct and direct comparisons, it may be helpful to discuss the study population first. It is common that a meta-analysis includes multiple studies with different subpopulations with different baseline risks. When we compare multiple treatments, it is important to make sure that relative effects are derived from the same population, i.e. comparing the counterfactual absolute effects if all subjects included in a research synthesis analysis were treated with a particular treatment versus that under another treatment. In addition, it is also important to specify which type of treatment effects are of primary interests: the (weighted) average of conditional (or study-specific) effects, or the marginal effect over all subjects include in the meta-analysis. It may implicitly produce systematic bias when we assume a specific scale of relative effects is transportable or transitive across population with different effect modifiers. For example, the commonly used assumption, that odds ratios are transportable, is not valid as shown in a recent controversy and debate based on 40,243 meta-analyses from the Cochrane Database of Systematic Reviews (https://pubmed.ncbi.nlm.nih.gov/34384876/ and https://pubmed.ncbi.nlm.nih.gov/34390790/) | Thank you, we refer to Section 3.1, where we added a reference to ICH E9 (R1). |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 7-13 | 158-413 | As pointed out in lines 212-217, effect modification depends on the scale on which the treatment effect is measured, it suggests that the assessment of exchangeability, similarity, homogeneity, and consistency also depend on the scale of effect measurement. However, it is unclear how to decide which relative effects to use in practice. For example, when the baseline risks are high, the choice of RR versus OR can lead to a substantial difference in the assessments of exchangeability, homogeneity, and consistency. It may be useful to advocate using multiple scales of effect measurements for those assessments, or to provide some guidelines on the choice of effect scales in practice. | This is part of the PICO and the estimand framework and not within the scope of this guideline. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 14-19 | 4 Methods Applicable to Direct or Indirect Comparisons (414-649) | In addition to methods mentioned in Line 426-434, bivariate generalized linear mixed models (BGLMM) have been frequently used in pairwise meta-analysis estimating marginal risk difference and relative risk, and meta-analysis of diagnostics tests. They are shown to perform better compared to traditional two-step approaches (https://pubmed.ncbi.nlm.nih.gov/21177306/). In the presence of double-zero-event studies, Peto and other traditional method should be avoided. Bayesian BGLMM and exact methods are recommended (https://pubmed.ncbi.nlm.nih.gov/30887438/), except when none of the included studies have an event in one or both treatment arms. | Meta-analysis of studies investigating diagnostic tests are out of the scope of this Guideline.<br><br>The Guideline is not a statistical textbook covering all available methods. |
| S. Walleser Autiero,<br><br>Medtronic | 26-27 | 962-967 | The two assumptions are that PSM requires “strong ignorability” whereby, 1) treatment assignment is independent of the potential outcomes and conditional on the observed baseline covariates and 2) every subject has a non-zero probability to receive either treatment. That first treatment. The first condition is very important and not discussed here. Suggest adding references such as Rosenbaum and Reubin. Biometrika, Volume 70, Issue 1, April 1983, Pages 41–55. | The 3 assumptions were already mentioned in the methodological guideline, and will remain the same for this guideline. |
| Bayer | 7-9 | Section 3.1 and Section 3.2.1 | The assessment of effect modifier status is not only specific to the treatments being compared (lines 202-211), but also to the scale used to measure the relative effect (as stated in lines 212-217). The guideline document recommends drawing potential effect modifiers a priori, from a literature search or recommendations from health care professionals (lines 176-182 and requirements for reporting in page 9). It is unclear how to do this without pre-specifying a relative effect measure of interest (using estimands terminology, a “population-level summary effect measure”) in 3.1 (“Initial Feasibility Questions”). | We agree; therefore, the list of effect modifiers refers to a certain effect measure. |
| Matias Olsen, EUCOPE | 7 | 149-151 | There is no point in highlighting certain potential effect modifiers over others, as any given modifier could be more important for a particular treatment than another. | We think it is useful to present some examples, which are frequently important in practice. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Replace:<br><br>"~~H~~However, patient characteristics, such as distributions of age, sex, disease duration, measurement, and operationalisation of the outcome of interest, and features of the experimental design still need to be assessed in detail.".<br><br>With:<br><br>"~~However, patient characteristics, such as distributions of age, sex, disease duration, measurement, and operationalisation of the outcome of interest, and features of the experimental design still need to be assessed in detail~~. **All effect modifiers should be assessed in detail**". | |
| EFSPI | 8 | 176 / 3.2.1 | "including duration of follow-up"<br>'Duration of follow-up' is a concept that is typically not properly defined. The Practical Guideline D4.3.1 should be made more precise. We recommend replacing the term 'duration of follow-up' with a more exact definition of what is requested. See https://arxiv.org/abs/2206.05216 for a discussion. | It is not intended that this Guideline would provide such a level of detail. |
| Mihai Rotaru, EFPIA | 8 | 178-179 | **Assessment of exchangeability – Study and patient characteristics (2)**<br><br><u>Current wording:</u><br><br>The following characteristics are generally relevant: age, sex, disease severity, region, and study duration. | We modified the sentence slightly. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br>~~The following characteristics are generally relevant: age, sex, disease severity, region, and study duration.~~<br><br>[note: ~~strikethrough~~ denotes proposed deletion]<br><br>Rationale:<br>EFPIA believes that the term (*generally relevant*) is unclear in this specific context. For example, this could be interpreted as meaning these characteristics are typically effect modifiers, which we think would be inaccurate. EFPIA therefore requests deletion of this sentence. | |
| GSK | 8 | 178-9 | Not clear what "generally relevant" means in this context. It could be interpreted as meaning these characteristics are typically effect modifiers, which we do not believe is accurate. Suggest either removal of sentence or further clarification. | We modified the sentence slightly. |
| EFSPI | 8 | 179 | "are generally relevant: age, sex, disease severity, region, and **study duration**"<br><br>While we agree that study duration is an important parameter, we would suggest to treat it differently from the other variables mentioned as study duration cannot be adjusted for. | We disagree; it is possible to adjust for study duration in data analyses. |
| Tanja Podkonjak, | 8 | 196-199 | Current wording:<br>**"Observed values of relevant outcomes at baseline:** an examination of the observed values of relevant outcomes at baseline | We think it is clear that "observed values at baseline" are the values |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Takeda Pharmaceuticals International AG | | | can provide information on the similarity of the individual studies, especially the study arms in which the comparator is used. However, to determine similarity, it is not a standard prerequisite that the observed values have to be identical, because the distribution of prognostic variables might well differ between studies."<br><br>Comment:<br>Please clarify what is meant by defining baseline values for outcomes. Are these measures (e.g., tumor size) or some other factor which may be impacted by an intercurrent event? This section follows effect modifiers by patient characteristics and intervention characteristics. And after this section, important things to note about effect modification is then discussed. Suggest moving this section after effect modification (lines 200-221). | of outcomes before randomisation. We see no need to move the section behind the explanations regarding effect modification. |
| EFSPI | 8 | 198 / 3.2.1 | Current wording:<br>"If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead."<br><br>Suggested wording:<br>~~"If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead."~~<br><br>Reason for change:<br>Use of post-baseline variables bears the risk of introducing lots of issues and biases, e.g. unclear causality, immortal bias. This could only be an option for factors like e.g. gender, race etc, that do not change over the course of the trial. | This is true if the goal is the estimation of the treatment effect. However, here the goal is the assessment of similarity. We wonder why you want to delete this option, because if the values at baseline for important outcomes are not available, it is impossible to assess similarity and all corresponding evidence syntheses are questionable. We added a note that differences can of course be due to treatment effects. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 8 | 200-217/ 3.2.1. | The concept of interaction and the one of effect modification are interchangeably used in clinical trial reports. However, these are two different concepts. The effect modification means a causal effect of an exposure (e.g., technology) that differs across the levels of a second exposure (e.g., prognostic factor) while the interaction means a joint causal effect of two exposures (e.g., technology and prognostic factor) that differs from their expected combined effect. Therefore, it would be very interesting to:<br>1. clarify these two concepts;<br>2. explicitly name the additive and multiplicative scales on which they should be assessed; and<br>3. list the relevant information for their reporting<br><br>Useful reference :<br>• Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. 2012 Apr;41(2):514-20. doi: 10.1093/ije/dyr218. Epub 2012 Jan 9. PMID: 22253321; PMCID: PMC3324457.<br>• VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology. 2009 Nov;20(6):863-71. doi: 10.1097/EDE.0b013e3181ba333c. Erratum in: Epidemiology. 2010 Jan;21(1):162. Erratum in: Epidemiology. 2011 Sep;22(5):752. PMID: 19806059.<br>• Bours MJL. Tutorial: A nontechnical explanation of the counterfactual definition of effect modification and interaction. J Clin Epidemiol. 2021 Jun;134:113-124. doi: 10.1016/j.jclinepi.2021.01.022. Epub 2021 Feb 4. PMID: 33548464. | Interaction is a more general term, because, e.g., in a regression model there may also be an interaction between two patient characteristics. Nevertheless, effect modification can be investigated by means of interaction terms. We see no need for precise definitions in this Guideline, which is not a statistical textbook. |
| Dr. Norbert Gerbsch<br><br>for IGES Institut GmbH and HealthEcon AG | 8 | 212-217 / Section 3.2.1 | **Comment:**<br><br>DRAFT GUIDANCE: “The status of a variable as an effect modifier, and the magnitude and direction of this effect, is specific to the scale on which the treatment effect is measured. For example, in a hypothetical placebo-controlled study of an influenza vaccine, female participants experience a reduction in risk from 10% to 5% and male participants from 6% to 3%, with vaccination compared with placebo. | Thank you for your comment. We disagree that this example is unhelpful; indeed, its purpose is to illustrate that effect-modification is specific to the effect measure used. This has |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | On the relative risk scale, sex is not an effect modifier [relative risk (RR)=0.5 in both groups], but it is on the risk-difference scale (–5% for females versus –3% for males)."<br><br>The cited example is not helpful. It suggests that an effect modification for the same subgroup characteristic and binary endpoint depends on the effect measure used.<br><br>Regardless of the substantive complications of such a statement, there is no practical relevance because one usually examines effect modification via an interaction term in a GLM (generalized linear model) and interprets the resulting p-value regardless of the risks for the type 1 and type 2 error. The effect measure is merely a consequence of the distribution and link function used.<br><br>**Suggestion:**<br><br>We recommend deleting the example lines 212-217 without replacement. | major practical relevance when assessing the similarity assumption in evidence synthesis, since we require similarity of the distributions of those variables that are effect modifiers with respect to the treatment effect measure being synthesised. |
| Mihai Rotaru, EFPIA | 8 | 196-198 | **Assessment of similarity - Observed values of relevant outcomes at baseline (1)**<br><br><u>Current wording:</u><br><br>Nevertheless, extreme differences that even lead to floor or ceiling effects regarding the range of possible outcome values should not exist. | The explanation is given in the same sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Comment:<br>EFPIA recommends further clarification in the practical guideline is required for the appropriate interpretation of *extreme differences*. | |
| GSK | 8 | 198-199 | Given the values recorded during the study could be influenced by treatment, we don't believe they should be used as a replacement for values recorded at baseline. | We wonder, why you want to delete this option, because if the values at baseline for important outcomes are not available it is impossible to assess similarity and all corresponding evidence syntheses are questionable. We added a note that differences can of course be due to treatment effects. |
| Mihai Rotaru, EFPIA | 8 | 198-199 | **Assessment of similarity - Observed values of relevant outcomes at baseline (2)**<br>Current wording:<br>If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead.<br>Proposed wording | We wonder, why you want to delete this option, because if the values at baseline for important outcomes are not available it is impossible to assess similarity and all corresponding evidence syntheses are questionable. We added a note that differences can of course be due to treatment effects. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | If ~~the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead~~ **information on a treatment effect modifier is not available from the comparator study at baseline, this has to be acknowledged as a limitation of the analysis. Post-baseline data (if available) should be used with caution, as it is likely to confound the association between imbalances in characteristics across treatment groups and outcome, given the potential impact of treatment on these post-baseline characteristics.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed inclusion and deletion, respectively]<br><br><u>Rationale</u><br><br>Patient characteristics are likely to change and evolve post-baseline. These may also be influenced by the treatment received (post-baseline), which is being tested through the indirect treatment comparison – thereby conflating effects. | |
| Bayer | 8 | 201 | <u>Current wording:</u><br><br>“Not all prognostic variables are effect modifiers”.<br><br><u>Proposed wording</u>:<br><br>“not all prognostic variables are effect modifiers on a specific scale”. | We agree, however this is discussed in bullet point three. We believe that including it in the first bullet point would confuse the reader. We have cited the suggested reference, thank you. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Rationale:<br><br>Effect modifier status is tied to the scale used to define the relative effect. A prognostic covariate that is not an effect modifier on one scale will likely be an effect modifier on another scale.<br><br>Reference: Brumback B, Berg A. On effect-measure modification: relationships among changes in the relative risk, odds ratio, and risk difference. Stat Med. 2008; 27(18): 3453- 3465. | |
| EFSPI | 9 | 249-50 / 3.2.2 | Current wording:<br>"The heterogeneity between the studies has to be assessed to determine whether a pooling of the results is meaningful at all and to choose between the fixed-effect and random-effects approach for the evidence synthesis."<br><br>Suggested wording:<br>"The heterogeneity between the studies has to be assessed to ~~determine whether a pooling of the results is meaningful at all~~ to inform the synthesis process and to choose between the fixed-effect and random-effects approach for the evidence synthesis."<br><br>Reason for change:<br>Heterogeneity is in itself not a reason for not pooling studies in evidence synthesis as heterogeneity can be accounted for via random effects. Large heterogeneity will lead to larger uncertainty in the results and, therefore, weaker conclusions. Also, the Practical Guideline D431 correctly points out that inconsistency must be assessed.<br>Heterogeneity may lead to inconsistency, therefore, problematic levels of heterogeneity - leading to inconsistency - would be detected and the corresponding synthesis be flagged as problematic.<br>Ruling out the pooling of studies purely based on heterogeneity would unduly penalise certain disease areas, where for example endpoints | We disagree; it is a broad agreement that the pooling of studies is not useful if heterogeneity is too large. We added a reference to this statement. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | are inherently variable (such as for example diseases leading to cognitive impairment, which is more difficult to quantify than overall survival, say). | |
| EFSPI | 9 | 262 / 3.2.2 | Suggest to add the following text:<br>"It is important to note that I2 is not an absolute measure of heterogeneity. I2 is the proportion of the total (observed) variance that is due to the between-study variance. In other words, it tells 'what proportion of the observed variance would remain if one could eliminate the sampling error.' [1]"<br>[1] Borenstein et al., "Basics of Meta-Analysis: I2 Is Not an Absolute Measure of Heterogeneity," Research Synthesis Methods 8, no. 1 (March 2017): 5–18, https://doi.org/10.1002/jrsm.1230. | We think that this issue is clear from the given explanations and references. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 9 | 263-264/ 3.2.1 | This statement '' One easy and objective criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test ($p$ <0.05) '' must be nuanced because it is not true when the heterogeneity can be explained by modifying factors.<br><br>Here is a suggestion:<br>One easy and objective criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test ($p$ <0.05) *with an unexplained heterogeneity*. | See Section 3.3 |
| EFSPI | 9 | 263-73 / 3.2.2 | Recommend to reword the paragraph: The Q-test is not a useful test in practice - the paragraph itself illustrates the limitations. Therefore it cannot serve as an "easy and objective criterion to decide whether the studies should not be pooled". | We disagree; with the limitations in mind, the Q-test is useful in practice. |
| Mihai Rotaru, EFPIA | 9 | 223-225 | **Assessment of similarity - Subgroups**<br><br><u>Current wording:</u> | It is unclear what you mean with "univariate" and "multivariate models" in the context of subgroup analyses.<br><br>It is not an option to simply ignore potential |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | However, statistical tests for effect modification using subgroup data from clinical trials (e.g., testing for the significance of interaction terms) will often be underpowered and suffer from issues with multiplicity.<br><br>Proposed wording:<br><br>However, statistical tests for effect modification using subgroup data from clinical trials (e.g.,testing for the significance of interaction terms) will often be underpowered and suffer from issues with multiplicity. **In such cases, the application of univariate models should be considered as compared to multivariate models.**<br><br>[note: **bold** denotes proposed inclusion]<br><br>Rationale:<br><br>EFPIA recommends additional clarification and inclusion of the potential application of univariate models. We agree there will be instances where low power may be expected and, in such cases, the application of univariate models might be an alternative. Furthermore, we wish to reiterate that the proposed PICO process may also drive the occurrence of this issue (e.g., sub-population or sub-groups) which may not be supported by clinical or biological plausibility. | effect modification because statistical tests for interaction are underpowered. |
| Mihai Rotaru, EFPIA | 9 | 249-251 | **Assessment of homogeneity (1)**<br><br>Current wording:<br><br>The heterogeneity between the studies has to be assessed to determine whether a pooling of the results is meaningful at all and to | We disagree; it is a broad agreement that the pooling of studies is not useful if heterogeneity is too large. We added a |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | choose between the fixed-effect and random-effects approach for the evidence synthesis.<br><br>Proposed wording:<br><br>The heterogeneity between the studies has to be assessed ~~to determine whether a pooling of the results is meaningful at all~~ **to inform the evidence synthesis process** and to choose between the fixed-effect and random-effects approach for the evidence synthesis.<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>EFPIA believes that heterogeneity is in itself not a reason for not pooling studies in evidence synthesis as heterogeneity can be accounted for via random effects. Large heterogeneity will lead to larger uncertainty in the results and, therefore, weaker conclusions. Also, the Practical Guideline (D4.3.1) correctly points out that inconsistency must be assessed. Heterogeneity may lead to inconsistency, therefore, problematic levels of heterogeneity - leading to inconsistency - would be detected and the corresponding synthesis be flagged as problematic.<br><br>Ruling out the pooling of studies purely based on heterogeneity would unduly penalise certain disease areas, where for example endpoints are inherently variable (such as for example diseases leading to cognitive impairment, which is more difficult to quantify than overall survival, for example). | reference to this statement. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| Mihai Rotaru, EFPIA | 9 | 260-262 | **Assessment of homogeneity – $I^2$ value**<br><br>Current wording:<br><br>However, the importance of observed $I^2$ values depends on the magnitude and direction of treatment effects and the strength of evidence for heterogeneity (p-value from the Q-test, uncertainty of the $I^2$, or number of studies).<br><br>Proposed wording:<br><br>However, the importance of observed $I^2$ values depends on the magnitude and direction of treatment effects and the strength of evidence for heterogeneity (p-value from the Q-test, uncertainty of the $I^2$, or number of studies). **It is also important to note that $I^2$ is not an absolute measure of heterogeneity. $I^2$ is the proportion of the total (observed) variance that is due to the between-study variance. In other words, it tells us 'what proportion of the observed variance would remain if one could eliminate the sampling error.' [1]**<br><br>[note: **bold** denotes proposed insertion]<br><br>Rationale:<br><br>EFPIA recommends the addition of the proposed text regarding the measure of heterogeneity ($I^2$) for additional clarify in the practical guideline. | We think that this issue is clear from the given explanations and references. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | It should be noted that I2 measures the proportion of total variability that can be attributed to between-study heterogeneity, which means I2 is a relative measure. In addition, an estimate of the absolute level of between-study variability on a meaningful scale should be provided. Therefore, we recommend that for RE models the RE standard deviation should always be reported.<br><br>In practice, the I2 and the RE SD are more meaningful tools to judge the importance of heterogeneity than the Q-test and p-values. Therefore, EFPIA recommends dropping the Q-test and the p-value of the heterogeneity test (given the limitations described in lines 263-273 of the Practical Guideline D4.3.1 itself).<br><br><u>Reference:</u><br>1. Borenstein et al., "Basics of Meta-Analysis: I2 Is Not an Absolute Measure of Heterogeneity," Research Synthesis Methods 8, no. 1 (March 2017): 5–18, https://doi.org/10.1002/jrsm.1230. | |
| S. Walleser Autiero,<br><br>Medtronic | 9 | 246-247 | It is not uncommon for meta-analyses of medical devices to have less than 5 RCTs included for each outcome of interest. The statement that analyses with less than 5 trials are not reliable is not reflective of current and past meta-analyses for non-pharmaceutical technologies and should be revised to reflect diversity in the evidence base. | See Section 4.1.3 |
| Mihai Rotaru, EFPIA | 9 | 251-252 | **Assessment of homogeneity (2)**<br><br><u>Current wording:</u> | Minor change to text made. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | It is important to use statistical methods as well as design features of the included studies to assess heterogeneity.<br><br>Proposed wording:<br><br>It is important to use statistical methods **to estimate heterogeneity** as well as **to report** design features of the included studies **that may lead** to ~~assess~~ heterogeneity.<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>EFPIA recommends that the document clearly distinguishes between qualitative and quantitative assessments of heterogeneity. | |
| Mihai Rotaru, EFPIA | 9 | 253-254 | **Assessment of homogeneity (3)**<br><br>Current wording:<br><br>Two widely used statistical approaches to assess heterogeneity are given by the statistical test based on the Q statistic (Q-test) [8,53] and the heterogeneity measure $I^2$ [8,25].<br><br>Proposed wording: | Thank you, we added a slightly modified sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Two widely used statistical approaches to assess heterogeneity are given by the statistical test based on the Q statistic (Q-test) [8,53] and the heterogeneity measure $I^2$ [8,25]. **In the case of IPD meta-analysis, other approaches can be more appropriate.**<br><br>[note: **bold** denotes proposed inclusion]<br><br><u>Rationale:</u><br>EFPIA recommends that the situation in which the HTD has access to IPD should also be mentioned here. In this case, an individual patient meta-analysis can be preferable, an in this case, other tests of heterogeneity might be a better choice than Q or $I^2$. | |
| S. Walleser Autiero,<br>Medtronic | 9 | 258-259 | The overlap between percentages in these 2 lines should be reconsidered. Greater than 75-80% should be considered as substantial or considerable heterogeneity. | This is the proposal in the Cochrane Handbook, which is used as reference. |
| EFSPI | 9 | 263-264 | "One easy and objective criterion to decide whether the studies should not be pooled is given by the<br>statistical significance of the Q-test ($p < 0.05$)"<br><br>Given all the caveats that are given afterwards, we suggest to rephrase to "One criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test ($p < 0.05$)" | We see no need to modify this sentence. |
| Mihai Rotaru, EFPIA | 9 | 263-264 | **Assessment of homogeneity – Q-test**<br><br><u>Current wording:</u> | We see no need to modify this sentence, which should be used in |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | One easy and objective criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test ($p$ <0.05).<br><br>Proposed wording:<br><br>One ~~easy and objective~~ criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test ($p$ <0.05).<br><br>[note: ~~strikethrough~~ denotes proposed deletion]<br><br>Rationale:<br><br>EFPIA recommends the proposed wording or rewording of the paragraph due to the limitations of the Q-test in practice as described in the text. As such, it may not serve as an easy and objective criterion to decide whether studies should not be pooled. | combination with the following statement. |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 9 | 252-262 | While the Q and I2 statistics are useful measure of heterogeneity, they also have some limitations. For example, I2 statistics can be influenced by a few outlying studies, robust version of I2 statistics based median rather than mean have been proposed (https://pubmed.ncbi.nlm.nih.gov/27167143/), and has been shown to perform better (https://pubmed.ncbi.nlm.nih.gov/29847495/). In addition, I2 statistics is not an absolute measure of heterogeneity (https://pubmed.ncbi.nlm.nih.gov/28058794/). | Thank you but this level of detail is beyond the scope of the guideline. |
| Tanja Podkonjak, | 9 | 263-273 | Current wording: | We see no need to modify this sentence, |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Takeda Pharmaceuticals International AG | | | "One easy and objective criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test (*p* <0.05)."<br><br>Suggested wording:<br>"One ~~easy and objective~~ criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test (*p* <0.05)."<br><br>Rationale:<br>Due to the limitations of the Q-test as described in the following text, we recommed the adjectives 'easy and objective' be removed. | which should be used in combination with the following statement. |
| Mihai Rotaru, EFPIA | 9 | 222-223 | **Assessment of similarity – Patient covariates**<br><br>Current wording:<br>The assessment of similarity should include a quantitative analysis of the impact on all observed patient covariates.<br><br>Proposed wording:<br>The assessment of similarity should include a quantitative analysis of the impact on all observed patient covariates. **Once the assessment is made, the significant effects need to be cross-checked against availability from the competitor trial. If certain relevant effect modifiers are unavailable this should be acknowledged as a limitation. Proxies or potential effect modifiers that use different** | Thank you for this suggestion. We have added some text to this effect elsewhere in this section. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **units to that collected on the study (e.g., lab values) should not be discounted, but should be consulted on their suitability for inclusion with healthcare professionals.**<br><br>[note: **bold** denotes proposed insertion]<br><br><u>Rationale:</u><br>EFPIA wishes to highlight that a quantitative assessment of all potential effect modifiers against all observed patient covariates may not be feasible or sensible in all circumstances. Additionally, we recommend the consideration of proxies or potential effect modifiers for consultation with healthcare professionals. | |
| Sebastian Werner<br>vfa | 10 | 274 - 287 | *[..] "If it can be decided that there is sufficient homogeneity and it is meaningful to pool the included studies, it has to be determined whether a fixed-effect or a random-effects model should be used for the evidence synthesis" [..]*<br><br>It would be more informative to quantitatively specify "few studies" and "large number of studies" to give better guidance on when to apply fixed or random effects models. This would ensure that assessors and HTD can precisely follow these aspects. In case of few studies a fixed effects model should be preferred compared to a random effect model as there are methodological limitations due to the small number of studies that limit the meaningful use. It can be expected that for most new medicines under evaluation only a very limited number of appropriate studies in the respective indication will exist. Therefore, this section would benefit from further specifications and discussions about use of Fixed effects analyses. | See Section 4.1.3 |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| GSK | 10 | 288-295 | What kind of response can the HTD expect from the assessor's evaluation, will there be a clear recommendation on the approach to take? | This depends on the submitted dossier and the corresponding data situation. |
| GSK | 10 | 282-284 | How many studies should be included to make a fixed-effect model appropriate? | Ultimately, the choice of fixed effect versus random effects depends on the degree of homogeneity between the studies and is independent of the number of included studies. |
| GSK | 10 | 286-287 | What are the criteria for the random-effects model to be "not feasible in practice"? Is that based on computational/convergence or other considerations? | Typically, this occurs when there are too few studies to reliably estimate the between-study variance parameter. We believe that this issue is discussed in sufficient depth in Section 4. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 10 | 276-278/3.2.1 | It is not appropriate to use the heterogeneity test results for the choice of the statistical model. Indeed, this choice has to be done a priori depending on the assumption of interest.<br><br>Useful reference:<br>Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010 Apr;1(2):97-111. doi: 10.1002/jrsm.12. Epub 2010 Nov 21. PMID: 26061376. | Ideally, you are right. However, heterogeneity may be unexpectedly very large. There is a broad agreement that heterogeneity has to be thoroughly investigated in any evidence synthesis, and the |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | corresponding results may impact on the model choice. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 10 | 288-295/3.2.1 | It would be interesting to add the following relevant points in this section:<br><br>• The choice of statistical model has to be done a priori on the assumption tested.<br>• The subgroup analyses have to be performed to explain the presence of heterogeneity using predetermined variables. So, the methods used to explain the heterogeneity as well as the results associated have to be reported.<br><br>Useful reference :<br>Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010 Apr;1(2):97-111. doi: 10.1002/jrsm.12. Epub 2010 Nov 21. PMID: 26061376. | This first addition is not useful because this is frequently not possible in practice.<br><br>Regarding the second point, we refer to Section 3.3. |
| EFSPI | 10 | 289-91 / 3.2.2 | Rather than testing for homogeneity, the focus should be on quantifying between-study-heterogeneity, which is typically achieved by estimating the random effects standard deviation. An estimate of the RE SD (or Var) should be added.<br>One should also consider dropping the p-value of the heterogeneity test given its limitations. | We disagree; the *p*-value should be reported. However, of course, its limitations have to be considered. |
| EFSPI | 10 | 299-302 | Current wording:<br>“Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts.”<br>Suggested wording:<br>~~“Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus,~~ | Thank you, we have amended the text. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts."~~<br>Reason for change:<br>We consider the current text potentially misleading. Inconsistency could in principle also arise for groups of very homogeneous trials, but where direct and indirect evidence are (systematically) in conflict. Inconsistency and heterogeneity are related, but still different concepts (and, therefore, we recommend not using the formulation that "inconsistency is a form of heterogeneity".) | |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Dr. Lydia Frick<br><br>SKC Beratungsgesellschaft mbH | 10 | 290 | Original wording:<br>"[…] (including the forest plots, the p-values for the heterogeneity test, and the $I^2$ values) […]"<br><br>Comment:<br>According to D4.3.1, "[…] typically at least five studies are required for a reliable assessment" and "if only one study is available for each pairwise comparison, the homogeneity assumption cannot be tested". This implies that the heterogeneity test cannot be performed in many cases. Hence, p-values for the heterogeneity test and the $I^2$ values should not be required for reporting in general.<br><br>Suggestion for rewording:<br>"[…] (including the forest plots, *and, where applicable,* the p-values for the heterogeneity test and the $I^2$ values) […]" | Of course, the requirement for reporting refers only to situations where the corresponding analyses are possible and applied. For example, the reporting box of Section 4.3 has only relevance if an NMA was undertaken. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 11 | 323-335/3.2.3 | Since the multi-arm trials can be often included in the network meta-analysis, it would be interesting to inform on how to address the inconsistency in a network with multi-arm trials.<br><br>Useful references suggested: | We refer to D4.3.2 |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | • Higgins JP, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods. 2012 Jun;3(2):98-110. doi: 10.1002/jrsm.1044. PMID: 26062084; PMCID: PMC4433772.<br>Jackson D, Barrett JK, Rice S, White IR, Higgins JP. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. Stat Med. 2014 Sep 20;33(21):3639-54. doi: 10.1002/sim.6188. Epub 2014 Apr 29. PMID: 24777711; PMCID: PMC4285290. | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 12 | 362-373/3.2.3 | It would be interesting to add the following point given the particularity of the three-arm trials:<br>State if or not trials with more than two arms were included in the analysis. | This is inherently included in Section 4.2 |
| S. Walleser Autiero,<br><br>Medtronic | 12 | 396-401 | Indirect methods require stronger assumptions and therefore it is more accurate that these methods may be used if the similarity assumption is violated and other assumptions that are required are not. | See Section 5 |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 12 | 377-380 | Current wording:<br><br>"1. Splitting into subgroups: if dissimilarity is shown for a potential effect modifier or heterogeneity is shown that can be explained by the effect modifier, it might be useful to divide the entire study pool into several subpools and draw separate conclusions (e.g., for men and women). The limitations of subgroup analyses based upon aggregated data should be taken into account [14];"<br><br>Comment:<br><br>Takeda is concerned around the proposed approach of disaggregation of a study population into several sub-pools in the event of heterogeneity. Although we understand the rationale behind the proposal, we are concerned about breaking randomisation and analysing small underpowered populations. Aligned to our feedback | The splitting into subgroups regarding variables measured before randomisation does not break randomisation. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | given in the scoping consultation, we believe sub-population analyses should be limited to those that are pre-specified or have a strong clinical and policy (i.e. PICO) rationale. Please add additional clarification and parameters of when this approach and to what extent subgroups and subpools are appropriate. | |
| Bayer | 12 | 381-383 | Current wording:<br><br>"Use of (network) meta-regression: (...) requires a sufficient number of data points (= number of studies) so that all parameters can be estimated in the model."<br><br>Proposed wording:<br><br>Delete.<br><br>Rationale:<br><br>While a few studies are likely required to investigate heterogeneity, a network meta-regression can be used for covariate adjustment with as little as two studies (with the relevant number of data points being the number of subjects). For instance, multilevel network meta-regression was developed as an alternative to MAIC and STC, which are typically used for covariate adjustment in a two-study scenario. Reference: Phillippo, D.M., Dias, S., Ades, A.E., Belger, M., Brnabic, A., Schacht, A., Saure, D., Kadziola, Z. and Welton, N.J., 2020. Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society: Series A (Statistics in Society), 183(3), pp.1189-1210. | This statement refers to the situation with aggregated data.<br><br>ML-NMR, MAIC and STC are mentioned in point 5. |
| Tanja Podkonjak, | 12 | 400-402 | Current wording: | Thank you for your comment. We agree that population adjustment may be an appropriate |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Takeda Pharmaceuticals International AG | | | "However, these methods have numerous limitations and might not generate results that are applicable to the research question."<br><br>Suggested wording:<br><br>"~~However, these methods have numerous limitations and might not generate results that are applicable to the research question.~~"<br><br>Rationale:<br><br>There may be circumstances where the application of population-adjusted indirect comparison may be the most appropriate way to address the research question identified by the PICO and the current language is dimissive of this approach. We recommend it be removed so that the presentation of all methods is balanced and without judgement. | way to deal with a violation of similarity in some situations; however the limitations of these methods must be acknowledged. We do not believe that the current language is dismissive of this approach. |
| Matias Olsen, EUCOPE | 12 | 400-402 | The methods described (MAIC, STC, ML-NMR) are, for innovative therapies in many cases the only way to establish comparative efficacy versus all relevant treatments. The guidance should just mention that its limitations should be addressed, the same way as it was done with the Bucher method or the Bayesian NMA.<br><br>Remove:<br><br>"~~However, these methods have numerous limitations and might not generate results that are applicable to the research question (see Section 5).~~". | Thank you for your comment. We agree that population adjustment may be an appropriate way to deal with a violation of similarity in some situations; however the limitations of these methods must be acknowledged. We do not believe that the current language is dismissive of this approach. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Mihai Rotaru, EFPIA | 12 | 400-402 | **Possible approaches when the assumptions are violated – Population-adjusted indirect comparisons**<br><br>Current wording:<br>However, these methods have numerous limitations and might not generate results that are applicable to the research question (see Section 5).<br><br>Proposed wording:<br>~~However, these methods have numerous limitations and might not generate results that are applicable to the research question (see Section 5).~~<br><br>[note: ~~strikethrough~~ denotes proposed deletion]<br><br>Rationale:<br>EFPIA requests the removal of this last sentence with regards the *numerous limitations and might not generate applicable results*. We think that there should be recognition that there will be circumstances where the application of population-adjusted indirect comparison may be the most appropriate way to address the mandated research question in the context of PICO. | Thank you for your comment. We agree that population adjustment may be an appropriate way to deal with a violation of similarity in some situations; however the limitations of these methods must be acknowledged. We do not believe that the current language is dismissive of this approach. |
| EFSPI | 12 | 400 - 401 | Current wording:<br>However, these methods have numerous limitations and might not generate results that are applicable to the research question. | Thank you for your comment. We agree that population adjustment |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested wording:<br>~~However, these methods have numerous limitations and might not generate results that are applicable to the research question.~~<br>Reason for change:<br>Statements that downgrade specific methods should be removed. Such statements do not support assessors with their assessment. | may be an appropriate way to deal with a violation of similarity in some situations; however the limitations of these methods must be acknowledged. We do not believe that the current language is dismissive of this approach. |
| EFSPI | 14 | 433-4 / 4.1.1 | The statement that Bayesian methods can be used should be a separate paragraph. The rest of the paragraph discusses “other methods for special situations”. The Bayesian approach represents an inferential framework applicable in all situations (line 433: “as a general alternative to frequentist methods”), not only in the special circumstances (such as rare events, and double-zero studies) discussed in the current paragraph (lines 426-32). | Thank you for your comment; we have made the suggested change. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 14 | 457-458/4.1.3 | The information reported here does not seem exact. The choice of statistical model must mainly depend on the prior hypothesis done (see previous comments). | We disagree; see the references given in Section 4. |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 14 | 444-455 | DerSimonian-Laird (DSL) method has been shown to produce biased estimates with falsely high precision (https://pubmed.ncbi.nlm.nih.gov/24727843/). Thus results from the Knapp-Hartung method should not be compared with DSL method - it might be better to be compared with random effects model using the REML method. | No, in this case a comparison with DSL (only to check whether the ad hoc variance correction is required) is appropriate; see the given references. |
| Tanja Podkonjak, | 14 | 463 - 468 | Current wording: | We disagree with this suggestion; see Section |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Takeda Pharmaceuticals International AG | | | "In general, a random-effects model should be applied even for meta-analyses with very few studies. However, the chance that the assumption of the fixed-effect approach is valid is greater in the case of very few studies compared with situations with a large number of studies. Especially in the situation with only two studies, it might be justified to apply the fixed-effect model by default. This means that the fixed-effect model should always be applied when there are only two studies, unless there are clear reasons against its use."<br><br>Suggested wording:<br><br>"In general, a random-effects model should be applied despite requiring more data to achieve the same statistical power as fixed effects models. If there is very little variations between studies (low I square), a fixed effects model might be appropriate."<br><br>Rationale:<br><br>Random-effects models requires more data to achieve the same statistical power as fixed-effects models. Random-effects models should not be used with sparse datasets and needs expert statistical guidance [Ref]. As the guidance suggested that Q-stat or I square need to be used for checking heterogeneity. It is not the number of studies that determines the appropriateness of fixed or random effects models, but the variation between studies.<br><br>Ref: ttps://www.statsdirect.com/help/meta_analysis/heterogeneity.htm | 4 and the references therein. |
| Sebastian Werner | 14 | 459 - 462 | *"Second, the standard random-effects KH approach has frequently very low power. The power might be so low that the KH confidence* | See Section 4.1.3 |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| vfa | | | *interval is wider than the union of all confidence intervals of the included studies [48]. In such cases, the KH method is not useful because the results are non-informative and, thus, alternative approaches are required."*<br><br>The issue of the low power of the KH approach is still existing, even if the KH confidence interval is not wider than the union of all confidence intervals of the included studies (e.g., KH is only slightly narrower). Therefore, KH approach should always be used careful in situations with less than 5 studies and fixed effect models might be more reasonable. | |
| | | | | |
| Matias Olsen, EUCOPE | 14 | 450-451 | The possibility of a comparison with DerSimonian-Laid (DSL) is appreciated. As we have noted in comments on deliverable D4.3.2 "Guideline on comparators and comparisons", the method of Knapp-Hartung is only suitable if more than 5 studies are available, something which is usually not the case for new products. As this is a case which could frequently come up in assessments, we recommend also specifying in the more general guideline this need for a comparison with the confidence interval calculated by means of DSL to assess whether an *ad hoc* variance correction is warranted in deliverable D4.3.2., Section 4.1. | The two Guidelines complement each other; there is no need for duplication. |
| Bayer | 14 | 457 | <u>Current wording:</u><br><br>"Meta-analyses with fewer than five studies are problematic in most cases".<br><br><u>Proposed wording:</u> | The statement is correct. We added a reference. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Delete<br><br>Rationale:<br><br>Pooling studies with sufficient homogeneity (in terms of similar outcomes/endpoints, treatment implementations when one of the treatments is evaluated in multiple studies, identical study designs, consistency with PICO properties, etc.) is not necessarily problematic if there are few studies. If only a few studies are available, meta-analysis is an appropriate and valid option to answer certain questions by increasing the number of subjects that are studied, power, etc. Furthermore, meta-analyses be aimed at increasing sample size (e.g., for statistical power) may be appropriate. Given that other conditions are satisfied, this could be reasonably achieved with two suitable studies. This is supported by Cochrane (Ryan R; Cochrane Consumers and Communication Review Group. 'Cochrane Consumers and Communication Group: meta-analysis. http://cccrg.cochrane.org, December 2016 (accessed 04/08/2022)). In reality, around launch of a new intervention it is not regularly to be assumed that a higher number of individual comparative trials will be available. Due to this fact this draft guideline schould focus on assessments which are derived out of a number of comparative trials that are usually available at this time. The unique focus on level 1a evidence is not fit for purpose for the assessment of new interventions around time of launch and needs adjustment to also consider other evidence levels than 1a only.<br>This is also relevant for rare indications or ATMPs, where it is even at later time-points hardly possible to have higher numbers of trials. | |
| EFSPI | 15 | 485-7 / 4.1.3 | "Alternatively, a random-effects Bayesian meta-analysis with weakly informative prior distribution for the heterogeneity parameter might be useful in the case of very few studies, because external heterogeneity information decreases the problem of estimating heterogeneity with insufficient data."<br>We agree and suggest adding references to the informative priors | Thank you for your comment, we agree and have added the suggested references. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | based on empirical evidence by Turner et al.:<br>Turner et al., "Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data from the Cochrane Database of Systematic Reviews," Int J Epidemiol 41, no. 3 (June 2012): 818–27, https://doi.org/10.1093/ije/dys041.<br>M. Turner et al., "Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis," Statistics in Medicine 34, no. 6 (March 15, 2015): 984–98, https://doi.org/10.1002/sim.6381.<br>Turner et al., "Incorporating External Evidence on Between-Trial Heterogeneity in Network Meta-Analysis," Statistics in Medicine 38, no. 8 (2019): 1321–35, https://doi.org/10.1002/sim.8044. | |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 15 | 485-490 | For the choice of prior information, particularly for the variance parameters, one may consider to use empirical distribution based on large number of CDSR meta-analyses (e.g. https://pubmed.ncbi.nlm.nih.gov/22461129/). Again, small study effect, publication, reporting bias should be discussed. | Thank you for your comment, we agree and have added the suggested reference. |
| Sebastian Werner<br><br>vfa | 15 | S 4.2<br><br>508-556 | *"4.2 Indirect comparisons"*<br><br>Bucher method is discussed in detail here, it is usually not recommended in ITC (e.g., Petto 2019; Alternative Weighting Approaches for Anchored Matching-Adjusted Indirect Comparisons<br><br>via a Common Comparator; Value in Health 22 (2019). The authors also indicated the limitation of Bucher, e.g., it can only be used for specific evidence network, does not allow adjustment and strong underlying assumptions. Many conclusions for the direct comparison may also be applicable here, e.g., random-effects model and Bayesian framework. | Thank you for your comment. We do not see any reason to recommend against the Bucher method, provided the assumptions are satisfied. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 15 | 457 | Current wording:<br><br>Meta-analyses with fewer than five studies are problematic in most cases.<br><br>Suggested wording:<br><br>Meta-analyses with fewer than five studies **may be** ~~are~~ problematic ~~in most cases~~, **considered where warranted, particularly in rare diseases and oncology indications where the number of available studies may be limited.**<br><br>Rationale:<br><br>The current statement is overly restrictive and does not consider circumstances where there is paucity of data available. In scenarios where there is low data availability, <5 studies available, the recommendation should still be for the evidence to be synthesised and presented and HTD encouraged to perform a meta-analysis. The guidance should also request the pros and cons of the meta-analysis with a small number of studies be clearly described and discussed in the submission dossier. These situations occur in the therapy areas including in the first two waves of JCAs, oncology and rare diseases, therefore we recommend the language be tweaked to add flexibility for these situations. | The statement is correct. We added a reference.<br><br>If a method has problems, this does not mean that it should not be applied (see Section 4.1.3). |
| Mihai Rotaru, EFPIA | 15 | 457 | Current wording:<br><br>Meta-analyses with fewer than five studies are problematic in most cases. | The statement is correct. We added a reference.<br><br>If a method has problems, this does not mean that it should not |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br><br>Meta-analyses with fewer than five studies **can be** ~~are~~ problematic ~~in most cases~~, **but can be exceptionally considered where warranted, particularly in rare diseases and oncology indications where the number of available studies may be limited.**<br><br>[note: ~~strikethrough~~ indicate suggested deleted text and **bold** recommended added text].<br><br>Rationale:<br><br>We assume it is generally unclear how often meta-analyses with fewer than five studies are "problematic". This statement reads too restrictive. As it might be often the case that <5 studies are available, one should still be aiming to perform a meta-analysis with all pros/cons in mind. | be applied (see Section 4.1.3). |
| ISPOR – The Professional Society for Health Economics and Outcomes Research | 16 | 533-539 | While the original method of Lumley [32] and the 'arm-based' NMA introduced by Hong et al. [26] make different assumptions, they can provide very similar goodness-of-fit to the data, and thus should be considered as alternative or sensitivity analyses to the contrast-based NMA. Furthermore, for binary data, the contrast-based NMA primarily use OR as the scale of effect measure. However, OR is a non-collapsible measure, the (weighted) average of OR from multiple studies, although mathematically attractive and valid, do not have a good interpretation as it does not apply to any population or study. In addition, for NMA of binary data, Bayesian methods should be preferred over frequentist approach as the latter typically has convergence issues when dealing with >3 or 4 dimension of random effects for a generalized linear mixed model. Computing relative effects with a normal approximation and subsequently fit linear mixed | Thank you for your comment. We recognise that these methods can be valid approaches in some situations. However, it is not feasible to provide detailed guidance on all available methods of NMA. We believe that the current text provides sufficient flexibility to allow for the application of these methods, while |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | model can have issues for rare outcomes as it involves continuity correction. | as the same time highlighting that the fundamental assumptions of these approaches are different and require careful assessment. |
| Mihai Rotaru, EFPIA | 16 | 530-532 | Current wording:<br><br>In networks with large discrepancies in the number of studies informing each contrast, the estimated treatment effect might change from being significant in a clinical trial study to non-significant in the evidence synthesis.<br><br>Comment:<br><br>We recommend reformulating the sentence since evidence synthesis results typically differ from single RCT results if there is a) large heterogeneity, or b) discrepancy between direct and indirect evidence. But a discrepancy in the number of studies informing each contrast is, per se, not an issue (though a more balanced evidence base is preferable). | We deleted the corresponding sentence. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 16 | 550-551 | Current wording:<br><br>If possible, assessment of rankograms [surface under the cumulative ranking curve (SUCRA), cumulative probability curves, and probability of being the best treatment]<br><br>Suggested wording:<br><br>If possible, assessment of rankograms [surface under the cumulative ranking curve (SUCRA), cumulative probability curves. ~~and probability~~ | We do not believe that presentation of these outputs constitutes a value-judgement, since they are quite clearly outputs of the NMA rather than conclusions of the JCA. We acknowledge that this could potentially be misunderstood by some |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~of being the best treatment].~~ A**ny SUCRA-ranking should be accompanied with an assessment of the limitations associated with such rankings and should never lead to an implicit or explicit value judgement of the technology.**<br><br>Rationale:<br><br>The reporting of rankograms must be accompanied with a careful assessment of the limitations of this activity to avoid misinterpertation and therefore patient access decisions based on SUCRA-rankings without full consideration of the limitations.<br><br>Furthermore, as the final value rating of a technology is not to be concluded at the EU JCA but instead left to each Member State, SUCRA ranking should never lead the assessor/co-assessor to reach conclusions regarding the value of a given technology. By stating in the JCA report the probability that a technology is the best treatment, the JCA assessors may be seen as issuing a value judgement which is not supported by the HTA Regulation text. We strong recommedn the text in this statement be changed. | readers of the JCA report. The reporting requirements include an assessment of these outputs, which is not the same as a presentation of these outputs without further comment. Therefore, we have not changed the highlighted text. However, we have included text acknowledging their limitations in the main body of Section 4.2. |
| Mihai Rotaru, EFPIA | 16 | 550-551 | Current wording:<br><br>If possible, assessment of rankograms [surface under the cumulative ranking curve (SUCRA), cumulative probability curves, and probability of being the best treatment]<br><br>Proposed wording: | We do not believe that presentation of these outputs constitutes a value-judgement, since they are quite clearly outputs of the NMA rather than conclusions of the JCA. We acknowledge that this could potentially be |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | If possible, assessment of rankograms [surface under the cumulative ranking curve (SUCRA), cumulative probability curves. ~~and probability of being the best treatment].~~ **However, any SUCRA-ranking should be accompanied with an assessment of the limitations associated with such rankings and should never lead to an implicit or explicit value judgement of the technology.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br><u>Rationale:</u><br><br>EFPIA believes that the reporting of rankograms must be accompanied with a careful assessment of the limitations of such exercise. It has been acknowledged[1], that SUCRA ratings may be misleading of the relative efficacy or safety of relevant technologies being compared, and that this would in turn lead decision-makers to reach erroneous conclusions when assessing a given intervention. In particular, SUCRA rankings do not consider the magnitude of differences in effects between treatments. In addition, it has been noted that[1]: "[…] *exactly the same set of ratings may arise from a small body of studies with major limitations in risk of bias (unconcealed randomization, lack of blinding, large loss to follow-up), imprecision (wide confidence intervals or small number of events), inconsistency in results, indirectness (for instance, studies enrolling a sample of patients that differ from the population of interest, or measuring outcomes differently, such as with shorter follow-up), and publication bias—and thus warrant only low or very low certainty*". . | misunderstood by some readers of the JCA report. The reporting requirements include an assessment of these outputs, which is not the same as a presentation of these outputs without further comment; therefore, we have not changed the highlighted text. However, we have included text acknowledging their limitations in the main body of Section 4.2. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Most importantly, EFPIA strongly recommends that the reporting of SUCRA ranking should never lead the assessor/co-assessor to reach conclusions regarding the value of a given technology.<br><br>Reference:<br>1. Mbuagbaw L, Rochwerg B, Jaeschke R, Heels-Andsell D, Alhazzani W, Thabane L, Guyatt GH. Approaches to interpreting and choosing the best treatments in network meta-analyses. Syst Rev. 2017 Apr 12;6(1):79. doi: 10.1186/s13643-017-0473-z. PMID: 28403893; PMCID: PMC5389085. | |
| | | | | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 16 | 546/4.2 | Line 546 should number of patients be also mentioned? | Thank you, we have added this. |
| | | | | |
| EFSPI | 16 | 530-2 / 4.2 | We recommend reformulating the sentence since evidence synthesis results typically differ from single RCT results if there is a) large heterogeneity, or b) discrepancy between direct and indirect evidence. But a discrepancy in the number of studies informing each contrast is, per se, not an issue (though a more balanced evidence base is preferable). | We deleted this sentence |
| EFSPI | 16 | 550-1 / 4.2 | We recommend removing rankograms and SUCRAs from the list of required outputs. Alternatively, rankograms and SUCRAs should be clearly flagged as optional. These outputs have many limitations and can be misleading.<br>For additional justifications, see: Mbuagbaw et al., “Approaches to Interpreting and Choosing the Best Treatments in Network Meta- | The reporting requirements include an assessment of these outputs, which is not the same as a presentation of these outputs without |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Analyses," Systematic Reviews 6 (2017): 79, https://doi.org/10.1186/s13643-017-0473-z. | further comment; therefore, we have not changed the highlighted text. However, we have included text acknowledging their limitations in the main body of Section 4.2. |
| EFSPI | 17 | 559-64 and 577-85 / 4.3.1 | The current Section 4.3.1 implies that when performing a network meta-analysis for time-to-event data, the PH assumption always needs to be assessed for the network via construction of pseudo-IPD from digitised Kaplan-Meier curves (for competitor trials). This requirement is meaningful when the validity of the PH assumption is a real concern, such as in immuno-oncology for example. However, we recommend to allow for more flexibility as there may be cases where the PH assumption is generally accepted for the given evidence base. In cases with broad agreement that PH is a valid assumption for the given setup, the steps outlined in Section 4.3.1 should not be needed. Flexibility and context specific considerations should explicitly be acknowledged in the Section 4.3.1 overall and in the box in lines 577-85. | Thank you; we modified the sentence |
| EFSPI | 17 | 560-4 / 4.3.1 | "This means that the assumption must be tested"<br>Testing for PH (understood as a formal hypothesis test) has the same drawback as, e.g., testing for heterogeneity: the null hypothesis is actually what you want to show (namely PH).<br>Proposal: rather "explore PH" assumption (e.g. through diagnostic plots) rather than use hypothesis test. We suggest to merge / align the text here with the one in Line 577ff. | Thank you; we modified the sentence. |
| EFSPI | 17 | 573-6 / 4.3.1 | Current wording:<br>"In this scenario, there are two alternative approaches that may be undertaken:" | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested wording:<br>"~~In this scenario, there are two alternative approaches that may be undertaken:~~<br>In this scenario, other effect measures should be considered, e.g.:"<br><br>Reason for change:<br>There are other options than the two approaches mentioned.<br>We also invite consideration of adding further potential effect measures to the list such as milestone comparison, weighted HR, average HR or HRs based on parametric models. | Thank you; we modified the sentence. |
| Bayer | 17 | 559-564 | Current wording:<br><br>"A (network) meta-analysis of HRs requires that the proportional hazards (PH) assumption holds for all pairwise comparisons in the network."<br><br>Proposed wording:<br><br>Delete<br><br>Rationale:<br><br>Given that the proportional hazards assumption will never "exactly" hold, this recommendation is c0onsidered to be too restrictive in complex networks including large numbers of studies/treatments. | The wording "requires that the assumption holds" does not mean that the assumption holds "exactly". |
| Mihai Rotaru, EFPIA | 17 | 560-563 | Current wording:<br><br>This means that the assumption must be tested for all included studies, which requires either access to individual patient-level data (IPD) for all studies or the construction of pseudo-IPD from digitised | Thank you; we modified the sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Kaplan–Meier curves (e.g., by using the algorithm proposed by Guyot [22]).<br><br><u>Comment.</u><br><br>Examining the PH assumption based on digitised KM curves ignores the possible stratification used in the study and an apparent deviation from PH can be detected incorrectly. There are no solutions unless KM curves are provided by stratum (or combination of strata). We suggest to include a sentence highlighting the limitation. | |
| GSK | 17 | 571-572 | Can the accelerated failure time model with Weibull distribution be used if PH assumption violated? | We consider this very specific question outside the scope of the commentary and cannot provide an answer without further context. |
| Richard Birnie<br><br>Lumanity | 17 | 571-572 | We believe the following statement is too strong and should be moderated “if the PH assumption is deemed to be implausible for one or more comparisons in the network, then (network) meta-analysis of HRs should not be carried out”<br><br>We agree that using HRs in a meta-analysis when the PH assumption does not hold does not make sense. However, if only one or two studies fail to show PH then abandoning the entire meta-analysis seems excessive. It would be more useful to perform sensitivity analysis excluding those studies where PH does not hold. This would at least allow estimates of relative effects to be derived making use of the evidence where the assumptions do hold. We recognise that this approach still has limitations but excluding one or two studies still seems preferable to discarding the whole evidence base. Of course, if | Thank you; we modified the sentence |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | the study that breaks PH is the main study for the new intervention being assessed then this approach would not be useful. Assuming that a connected network can still be formed then performing an analysis of those studies where PH holds alongside alternative methods such as splines, fractional polynomials or RMST would be appropriate in many cases.<br><br>We would suggest the following wording as an alternative ""if the PH assumption is deemed to be implausible for one or more comparisons in the network, then sensitivity analysis excluding those studies where PH does not hold should be performed and alternative methods of synthesis considered"" | |
| GSK | 17 | 582-583 | If evidence, such as Schoenfeld residuals and PH assumption statistical test, contradict each, can either evidence be chosen? Are there any rules for triangulation of evidence? | We consider this very specific question outside the scope of the commentary and cannot provide an answer without further context. |
| Mihai Rotaru, EFPIA | 17 | 584-585 | Current wording:<br><br>Any opinions from healthcare professionals received on the plausibility of the PH assumption; for example, if a delayed treatment effect is expected, then PH might not hold.<br><br>Proposed wording:<br><br>**Relevant external data in comparable treatment settings and** any opinions from healthcare professionals ~~received on the plausibility of the PH assumption;~~, for example, if a delayed treatment effect is expected, ~~then PH might not hold~~ **should be used to assess whether the PH assumption is likely to hold[1,2]. If the PH assumption is** | Thank you; we modified the sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **violated, a strategy should be adopted that is similar to violations of other NMA assumptions (e.g., homogeneity and consistency). The size of the violation, the ramifications of ignoring it and the data available to undertake more complex NMA methods must be taken into consideration.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br><u>Rationale:</u><br><br>A systematic and evidence-based approach is needed to justify the use of flexible survival models for NMA of time-to-event outcomes. In addition to expert opinion, it is also important to consider external evidence (for example, early phase and/or observational studies of relevance to the research question and/or target population) when assessing whether the PH assumption might hold. The application of more complex non-PH NMA should be carefully considered as it may introduce additional challenges in terms of the interpretation of results and completeness of the networks.<br><br><u>References:</u><br><br>1. NICE DSU Technical Support Document 21: Flexible methods for survival analysis. January 2021<br>2. Palmer S, Borget I, Friede T et al. A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. Value in Health 2022. Online access. | |
| Roche | 17 | 559-64 and<br><br>577-85 / 4.3.1 | The current Section 4.3.1 implies that when performing a network meta-analysis for time-to-event data, the PH assumption always needs to be assessed for the network via construction of pseudo-IPD from | Thank you; we modified the sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | digitised Kaplan-Meier curves (for competitor trials). This requirement is meaningful when the validity of the PH assumption is a real concern, such as in immuno-oncology for example. However, we recommend to allow for more flexibility as there may be cases where the PH assumption is generally accepted for the given evidence base. In cases with broad agreement that PH is a valid assumption for the given setup, the steps outlined in Section 4.3.1 should not be needed. Flexibility and context specific considerations should explicitly be acknowledged in the Section 4.3.1 overall and in the box in lines 577-85. | |
| EFSPI | 17 | 562 | Examining the PH assumption based on digitised KM curves ignores the possible stratification used in the study and an apparent deviation from PH can be detected incorrectly. there are no solutions unless KM curves are provided by stratum (or combination of strata). We suggest to include a sentence highlighting the limitation. | Thank you; we modified the sentence. |
| EFSPI | 18 | 599 / 4.3.3 | The original text reads as if this is a comprehensive list of alternatives to use in case of NPH, though there are other options than fractional polynomials (FP) or piecewise exponential models. Such alternative approaches should be mentioned too. | Thank you; we modified the sentence. |
| Richard Birnie<br>Lumanity | 18 | 604-614 | The framing of the discussion in this section seems to give undue emphasis to FP models with cubic splines appearing to be a bit of an afterthought. We assume that this is unintended but we would suggest amending the wording to give FP, piecewise models and splines equal emphasis. This could be achieved by referring to “flexible survival models” in the general text and citing splines, FP and piecewise models as alternatives that are generally equally credible. See also the paper by Freeman referenced above which includes other options that may be applicable here. | Thank you; we made a corresponding addition. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The limitations of FP, piecewise models and splines are largely similar. We agree that the interpretation of the relative effects obtained from such models is difficult. In many cases the synthesis of relative effects and the extrapolation of clinical outcomes cannot be considered in isolation. The additional flexibility obtained from a polynomial function may allow a better fit to the observed data but can lead to implausible extrapolations beyond the trial period. Although this is true of all three approaches to some extent FPs are more prone to this than the alternatives. | |
| Mihai Rotaru, EFPIA | 18 | 607-608 | Current wording:<br>A similar approach is possible using restricted cubic spline models [20].<br><br>Proposed wording:<br>A similar approach to FP (network) meta-analysis is to use restricted cubic spline models [20] which are functions of time that can mirror the complex shapes of some hazard functions. A restricted cubic spline is a series of polynomial functions. For this method, knots are placed at equally spaced percentiles of the uncensored survival times based on the (pseudo)-IPD for each trial, with boundary knot at the minimum and maximum values of the censored survival times. The number and location of the knots determines to complexity and flexibility of the RCS model[1].<br><br>Rationale:<br>EFPIA proposes that the current sentence is moved to line 615 and expanded to provide assessors/co-assessors with more details of the | Thank you; we made a corresponding addition. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | use of restricted cubic spline (RCS) models for NMA of time-to-event outcomes where the PH assumption does not hold. The main reason for this request is that there is some evidence to suggest this method may have advantages over fractional polynomial models and piecewise exponential models in terms of computational practicality and fit to the observed data[1].<br><br>Reference:<br><br>1. Freeman SC, Cooper NJ, Sutton AJ et al. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: Application to a melanoma network. Stat Methods Med Res 2022; 31(5): 839-861. https://dx.doi.org/10.1177/09622802211070253 | |
| Bayer | 20 | 659-666 | Specific comment:<br><br>MAIC could be used to reweight the IPD study with respect to another sample/population, not necessarily that of the AgD study. The issue would be performing the indirect treatment comparison outside the AgD study. So is the case for STC. While the fitted outcome regression model can be extrapolated to any population, the indirect comparison cannot be conducted outside the AgD study. Only ML-NMR can conduct the indirect comparison in any target population. | Thank you for your comment. We agree that the original wording was inaccurate and have amended the text accordingly. |
| GSK | 20 | 670-673 | What is the preferred way to show that bias will be reduced? | Thank you for highlighting this. We have included text to clarify how this may be achieved. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| S. Walleser Autiero,<br>Medtronic | 20 | 668-670 | The process to mitigate this concern is to prepare and publish a transparent protocol prior to starting the analysis, which is similar to best practices for systematic reviews, RWD and other studies. | Yes, but even with a detailed SAP a number of 'degrees of freedom' remain. |
| Mihai Rotaru, EFPIA | 20 | 671-673 | Current wording<br>In the case of anchored comparisons, it should be demonstrated that bias will be reduced by the use of a population-adjusted methods.<br><br>Proposed wording:<br>In the case of anchored comparisons, it should be demonstrated that bias will be reduced by the use of a population-adjusted method~~s~~. **This requires (i) showing there are grounds for believing one or more of the available covariates is an effect modifier, and (ii) showing that there is sufficient imbalance in those effect modifiers to result in a material bias, in relation to the observed relative treatment effect.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br>EFPIA believes the statement, in its current wording, is vague and can be misinterpreted. EFPIA suggests including additional wording on the issue, based on the publication by Phillippo et all, 2016.<br><br>Reference: | Thank you for your comment. We agree that there is a need for clarity here and have included text along the lines that you have proposed. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | 1. Phillippo DM, Ades AE, Dias S et al. NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submission to NICE. London, UK: National Institute for Health and Care Excellence; 2016. | |
| Richard Birnie<br><br>Lumanity | 20 | 666 | The statement that "ML-NMR can be applied to any connected network" is slightly misleading. We are not aware of any published method for ML-NMR for time to event outcomes. Therefore, ML-NMR cannot be applied to connected networks with time to event outcomes. This is a substantial limitation of the method as it prevents the use of the method in key areas, particularly oncology which makes up a large proportion of HTAs. If there are published methods for time to event that we are not aware of then it would be useful to cite those papers. | Thank you for highlighting this, to our knowledge ML-NMR cannot be applied to time-to-event outcomes. We have added a sentence to this effect. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 21 | 707 - 716 | Current wording:<br><br>"Assessors should be aware that, when population-adjusted indirect comparisons are carried out despite relevant covariates being unavailable, bias in the estimated treatment effects could be increased as a result of adjustment compared with the results of a standard NMA."<br><br>Suggested wording:<br><br>"Assessors should be aware that, when population-adjusted indirect comparisons are carried out despite relevant covariates being unavailable, bias in the estimated treatment effects could be *present* as a result of adjustment compared with the results of a standard NMA." | Thank you for your comment; we have reworded this sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Rationale:<br><br>The opposite could also be true- bias could also decrease, data would be missing non-differentially between the study groups for either a positive or negative bias to the hypothesis. Therefore, we recommend this sentence be modified. | |
| S. Walleser Autiero,<br><br>Medtronic | 21 | 717-721 | This was also included in the methods guide without providing details or references. Please consider adding best practice references for implementation of this method. | We added a reference |
| Sebastian Werner<br><br>vfa | 21 | 717-721 | *"To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified."*<br><br>The proposed approach of testing shifted hypotheses against a null hypothesis of some non-zero relative treatment effect of a magnitude large enough should be removed. This approach (testing of shifted hypothesis) is not representative of international guidelines on evidence-based medicines. Furthermore, existing guidance from leading international organisations for HTA such as ISPOR and Cochrane do not reference the 'testing of shifted hypothesis' method. As it is not an internationally recommended approach by neither other EU HTA bodies (accept IQWIG) nor academic societies, it is not appropriate to include this approach in the EU guideline. In addition, there is no standard rule on which these thresholds should be based, and what is acceptable can largely vary depending on disease area and/or outcomes. It is better suited to local complementary analyses of Member States than for JCA. | The use of shifted hypothesis testing is described as an option.<br><br>Other methods are mentioned in Section 6 of D4.3.2.<br><br>See also our replies to the public consultation of D4.3.2 regarding this issue. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The authors should give guidance on methods following the international standards of evidence-based medicine, discussing context-specific approaches (with best fit for the data situation) to evaluate the degree of certainty based on the best available evidence, incl. state-of-the-art methodology, such as propensity scoring, and matching-adjusted analyse. Different levels of uncertainty should be discussed in specific contexts using flexible methods and recommending further analyses to assess the robustness of results. | |
| Mihai Rotaru, EFPIA | 21 | 717-721 | **Test of shifted null hypothesis**.<br><br>Current wording:<br><br>To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified.<br><br>Proposed wording:<br><br>~~To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising~~ | The use of shifted hypothesis testing is described as an option. Other methods are possible, but convincing evidence has to be provided that the applied methods sufficiently account for risk of bias. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified.~~<br><br>[~~strikethrough~~ indicates proposed deletion]<br><br><u>Rationale:</u><br><br>EFPIA would like to reiterate our concerns over the use of the test of shifted null hypothesis to judge the acceptability of conclusions drawn from population-adjusted indirect comparisons. We believe the use of this approach goes beyond the scope of the EUnetHTA21 methods guides as it provides instructions to assessors and co-assessors on the level of uncertainty that will be acceptable for pan-EU decision making. The extent of uncertainty deemed acceptable for a specific decision problem implicitly represents a value judgement, and as such, remains the responsibility of each EU MS to consider as part of its national HTA process. This is stated in the EU HTA regulation, '*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*.'[1]<br><br>We are aware similar concerns have been expressed by multiple stakeholders in the public consultation to 'D4.3.2 – Methods guideline on comparators and comparisons'[2], including the International Society of Pharmacoeconomic Research (ISPOR). Furthermore, this approach is not a recognised international standard of evidence-based medicine, as required in the Regulation (Recital 24). | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | We feel EUnetHTA 21's response to these legitimate public concerns: "T*he use of shifted hypothesis testing is an accepted method to ensure formally that a treatment effect exceeds a certain threshold (e.g., in noninferiority trials*)." provides insufficient justification to warrant its inclusion in D.4.3.1 and D.4.3.2. We therefore request that EUnetHTA reconsiders the use of the test of shifted null hypothesis and remove it from these methods guides.<br><br>References:<br>1. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance).<br>2. EUnetHTA 21 Public Consultation of D4.3.2 – Methods Guideline on comparators and comparisons. | |
| EFSPI | 21 | 717-720 | Current wording:<br>"To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. "<br><br>Proposed wording:<br>~~To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising~~ | The use of shifted hypothesis testing is described as an option.<br>Other methods are mentioned in Section 6 of D4.3.2.<br>See also our replies to the public consultation of D4.3.2 regarding this issue. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified.~~<br><br>"A thorough discussion of the possible risk of bias (RoB) because of missing or unknown effect modifiers needs to be performed. Several approaches exist to balance the estimated treatment effect versus bias because of missing or unknown effect modifiers, One approach is the e-value, which gives a direct estimate about the magnitude of bias that would be needed to fully negate the estimated treatment effect. Another one is using a threshold for the treatment effect that is dependend on the magnitude of bias. The CI for the estimated treatment effect must not include the predefined threshold. More complex approaches like negative controls are available to estimate the magnitude of bias by unknown/unmeasured confounding."<br><br><u>Rationale:</u><br><br>Shifted hypotheses are only one of possible methods to ensure that the estimated treatment effect is large enough that it is not just a consequence of bias. There are no common thresholds, and the test is not scientifically established to balance RoB versus unmeasured confounding.<br><br>The concept is used for other applications like non-inferiority trials, but the non-inferiority margin, which the CI must not include to proof non-inferiority, needs to be prespecified upfront, based on the underlying disease and endpoint. This application is not linked to balance risk of bias and measured treatment effects.<br><br>We therefore either recommend to delete the complete sentence, or to add a list of possible methodological options, as the magnitude of the estimated treatment effect is only one option. | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Mihai Rotaru, EFPIA | 21 | 704-706 | **Assessing covariate selection – Bias estimation**<br><br>Current wording:<br><br>Assessors should highlight the potential for residual bias in the resulting estimate and give an indication of the size and direction of that bias where possible.<br><br>Proposed wording:<br><br>Assessors should highlight the potential for residual bias in the resulting estimate and give an indication of the ~~size and~~ direction of that bias where possible.<br><br>[note: ~~strikethrough~~ denotes proposed deletion]<br><br>Rationale:<br><br>EFPIA believes it is unrealistic for assessors to estimate the size of the potential bias.<br><br>Furthermore, EFPIA recommends that clear guidance should be given on the relevant methods that assessors could apply to estimate such residual bias. | We accept that in many situations it may not be possible to estimate the size of potential bias, and note that the wording 'where possible' acknowledges this. Nonetheless, it may still be possible to comment on the likely strength of effect-modification and/or the degree of imbalance between studies. We have therefore left this sentence unchanged. |
| Richard Birnie<br><br>Lumanity | 21 | 731 | Following from the comment above, this statement is slightly misleading "The STC and ML-NMR methods involve fitting an outcome regression model (e.g., a generalised linear model or Cox PH model)" | Thank you, we agree and have amended the text accordingly. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | We are not aware of an ML-NMR method that works with Cox models | |
| Mihai Rotaru, EFPIA | 21 | 699 to 706 | Comment:<br><br>EFPIA recommends that the guideline elaborates on the issue regarding the trade-offs between the number of treatment effect modifiers and the available sample size, which can be an issue particularly in rare diseases. | Thank you for your suggestion, we have included some text to this effect. |
| Bayer | 22 | 770-772 | Current wording:<br><br>"These approaches generally require additional assumptions to estimate the joint covariate distribution from the AgD study".<br><br>Proposed wording:<br><br>Delete<br><br>Rationale:<br><br>Arguably, they are not additional assumptions but simply different assumptions. It is true that the "covariate simulation"-based methods (marginalized STC and ML-NMR) make explicit parametric assumptions about the joint covariate distribution of the AgD population. However, all methods require assumptions to characterize this population due to limited IPD.<br><br>For MAIC and the (centered) version of STC, as stated in the NICE Decision Support Unit technical support document: "when covariate correlations are not available from the (AgD) population, and therefore cannot be balanced by inclusion in the weighting model, they are assumed to be equal to the correlations amongst covariates in the pseudo-population formed by weighting the (IPD) population." In | Thank you for highlighting this, we agree that these are different assumptions rather than additional assumptions. We have edited the wording accordingly. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | typical usage, MAIC only balances moments of the marginal distributions of selected baseline characteristics, not the multidimensional joint covariate distributions, due to the lack of published correlation data for the AgD study. In the "plug-in" version of STC, the correlations between the AgD covariates are assumed to be equal to the correlations between covariates in the IPD study.<br><br>In the marginalized version of STC and in ML-NMR, the joint distribution of the AgD covariates (the combination of their marginal distributions and correlation structure) is specified explicitly, by assigning parametric forms to the marginal distributions and specifying pairwise correlations. While assumptions are made more explicitly, all population-adjusted indirect comparisons require assumptions to approximate the joint distribution of covariates in the AgD trial. | |
| EFSPI | 22 | 748-9 / 5.3 | "In this case, standard measures of model fit, such as AIC/BIC, residual deviance, and so on, can be used to select these additional covariates."<br>Such model selection procedures typically again introduce bias for the resulting estimates. We suggest rewording such that these covariates are to be selected based on substantive-matter considerations. | Thank you for your comment. We are not aware of any reason why this would be the case and note that the use of AIC/DIC is recommended for this purpose in NICE TSD 18. We have revised the text slightly to highlight that this recommendation has been sourced from that report. |
| Bayer | 23 | 813-815 | Current wording:<br><br>"This could be problematic in the context of a JCA because it is likely that the source population is of greater interest to the assessment than the target population (see<br><br>815 also Section 5.6)." | Thank you for highlighting this, we have rephrased this sentence as suggested. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br>Use different term than “target population” i.e. “AgD population”, if this is what is meant.<br><br>Rationale:<br>This would completely depend on what the “target population” is. For HTA decision-making it is undoubtedly the population of greatest interest, particularly if the “source” trial lacks external validity. | |
| GSK | 23 | 810-811 | What is the smallest proportion ESS / "original sample size" that should be accepted? What smallest absolute ESS is acceptable? | We consider acceptability of an analysis to be beyond the scope of this guideline. A low ESS will result in a wide confidence interval for the treatment effect, which should be highlighted as an area of uncertainty in the JCA report. |
| Bayer | 23 | 812-813 | Current wording:<br>“Low ESS also indicates that the target population of the MAIC is considerably different from the source population”.<br><br>Proposed wording: | We agree and have amended the text accordingly. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | "Large reductions in ESS also indicates that the target population of the MAIC is considerably different from the source population".<br><br>Rationale:<br><br>A low ESS could simply be due to a low original sample size in the first place, even if the source and target populations are similar. | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 23 | 821-822/5.4 | It would be important to remind that the balance assessment must be done without hypothesis tests. Here is a suggestion:<br><br>• Assessment of covariate balance achieved after matching **without the use of hypothesis tests**, and of potential impact of any residual imbalance on the results (if this can be estimated) | Thank you; we agree and have inserted the suggested text. |
| EFSPI | 24 | | Current wording:<br><br>"Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, which is not the goal of the JCA."<br><br>Actual wording:<br><br>"~~Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, which is not the goal of the JCA~~."<br><br>Rationale:<br><br>We suggest to delete this sentence since it represents a strong value statement about the relevance of unanchored comparisons. A proper comparison of treatments in an unanchored setting relies on statistical | We deleted the sentence |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | modelling to approximate treatment effects expected in a head-to-head comparison. Accordingly, it does produce relative effects needed for JCA. However, they are associated with greater uncertainty due to the strong modelling assumptions. If those uncertainty is accepted should be left to MS. | |
| Mihai Rotaru, EFPIA | 24 | 831-832 | Current wording:<br><br>Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, which is not the goal of the JCA.<br><br>Proposed wording:<br><br>Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, ~~which is not the goal of the JCA.~~ **In particular contexts, like in treatments for some rare diseases, such unanchored comparisons might be the only feasible option while more data is generated and the comparison can be further refined.**<br><br>[[note: ~~strikethrough~~ indicate suggested deleted text and **bold** recommended added text].<br><br>Rationale:<br><br>EFPIA believes that the findings of unanchored ITCs are still relevant for decision makers and should still be presented and described in the context of the JCA, albeit with the necessary clarifications regarding the limitation of the approach and interpretation of the results. In addition, absolute treatment effects can still be important to decision- | We deleted the sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | makers and, thus, MS, and should not be dismissed in the context of a JCA. | |
| Bayer | 24 | 873-876 | Specific comment:<br><br>The method of moments in MAIC could also be used to reweight treatment effects from the IPD study to any population (assuming there is overlap), not necessarily that of the IPD study. STC and ML-NMR can also transport effects by extrapolating the outcome regression model fitted to the IPD. The challenge of MAIC and STC is performing the indirect comparison in any target that is not the AgD study. ML-NMR can potentially perform the indirect comparison in any target population. | Thank you for highlighting this, we have amended the text accordingly. |
| GSK | 24 | 849-851 | Is there a preference by therapeutic area, e.g. oncology should favour MAIC? Or endpoints? | There is ongoing debate in the scientific literature as to which of MAIC or STC/ML-NMR should be preferred for HTA in general. We are not aware of any specific preferences by therapeutic area. |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Elisa Zavatta, M.A.<br><br>SKC Beratungsgesellschaft mbH | 24 | 845 | Original wording:<br><br>“There will inevitably be differences in the trials other than patient characteristics. Interventions will be administered under different conditions and endpoints might be recorded in different ways (e.g., investigator versus independent assessment of tumour progression). Again, these differences typically have a greater impact on unanchored comparisons compared with anchored comparisons, because absolute effects are being compared.” | We have not made any comment in the guideline on whether or not such comparisons should be considered for JCA. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Comment:<br><br>According to D4.3.1, anchored comparisons are the preferred method compared to unanchored comparisons and differences in the trials other than patient characteristics such as differences in the administration of interventions or recording of endpoints typically have a greater impact on unanchored comparisons compared with anchored comparisons. Hence, all anchored comparisons complying with the methodological standards described in D4.3.1 should be considered for JCA. Moreover, differences between trials included in an anchored comparison other than patient characteristics should not be the reason why the results of an anchored comparison are not considered for JCA. | |
| GSK | 25 | 891-894 | Does that mean MAIC should be preferred over STC as a marginal population treatment effect looks more generalizable than an effect more specific to the covariates of the competitor trial? | We have not proposed that either method should be preferred, although this question has been debated in the literature. The JCA assessors should be aware of the different estimands targeted by each method and ensure that this is clearly stated in the JCA report. |
| Bayer | 25 | 902-903 | "When inference is made on the basis of population-adjusted comparisons, assessors should take into account that these comparisons are typically underpowered." Generally, studies are not powered for indirect treatment comparisons. This is something to bear in mind for indirect treatment comparisons in general, not exclusively for population-adjusted indirect comparisons. | We agree with this observation, however, we note that the loss of power is typically greater when adjustment is carried out. Therefore, we believe it is important |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | to retain this sentence in its current section. |
| Bayer | 26 | 941-943 | Specific comment:<br><br>Effect modifiers are here not directly relevant. The topic of interest is rather the lack of exchangeability between treatment arms. The requirement that confounders are measured is sufficient and therefore "effect modifiers" should be removed here. | We rather think that effect modifiers are relevant, because if there is an interaction between a confounder and the treatment, than this confounder is also an effect modifier. |
| Mihai Rotaru, EFPIA | 26 | 944-946 | Current wording:<br><br>The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present.<br><br>Proposed wording:<br><br>The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present, **however, the bias can be mitigated to a certain extent if the disease natural progression is predictable, and the observed effect size is large.**"<br><br>[note: **bold denotes** recommended added text].<br><br>Rationale:<br><br>Non-randomised studies should not be completely dismissed and there are situations where such studies can provide valid evidence for | We prefer to keep the original wording without listing potential exceptions. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | treatment effect either separately or combined and might even be needed in rare disease settings when RCTs are unethical or less feasible. | |
| EFSPI | 26 | 922 / 6.1 | "All commonly encountered sources of evidence outside of RCTs are non-randomised."<br><br>We see "randomization" vs. "non-randomization" not as a binary concept, but rather a continuum. As an example, we can complement an underpowered RCT with external data using dynamic borrowing.<br><br>While there is increased risk of bias for a trial that deviates from a fully powered RCT we would welcome acknowledgment that in absence of a fully powered RCT there is a range of designs that is at different risk of bias. | Sentence already concerned by a previous comment, see related answer. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 26 | 937-940/6.1 | To avoid any confusion, please do a clear distinction between observational studies (e.g., cohort studies, case-control studies) that can be evaluated with ROBINS-E tool and non-randomised studies with intervention that can be evaluated with ROBINS-I tool.<br><br>Useful reference for ROBINS-E:<br>• ROBINS-E Development Group (Higgins J, Morgan R, Rooney A, Taylor K, Thayer K, Silva R, Lemeris C, Akl A, Arroyave W, Bateson T, Berkman N, Demers P, Forastiere F, Glenn B, Hróbjartsson A, Kirrane E, LaKind J, Luben T, Lunn R, McAleenan A, McGuinness L, Meerpohl J, Mehta S, Nachman R, Obbagy J, O'Connor A, Radke E, Savović J, Schubauer-Berigan M, Schwingl P, Schunemann H, Shea B, Steenland K, Stewart T, Straif K, Tilling K, Verbeek V, Vermeulen R, Viswanathan M, Zahm S, Sterne J). Risk Of Bias In Non-randomized Studies - of Exposure (ROBINS-E). Launch version, 1 June 2022. Available from: https://www.riskofbias.info/welcome/robins-e-tool. | For RoB tools, see guideline 4.6 Validity of clinical studies. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| BAH | 26 | 942 | "However, this requires that all confounders and effect modifiers relevant for adjustment are measured and that the model and covariate selection strategies for adjustment are prespecified and based upon transparent criteria."<br><br>Practicability as well as scientific rigour must fit. Therefore, we propose to change the wording to:<br><br>"However, this requires that confounders and effect modifiers **with the highest confounding potential** should be measured (…)" | Thank you for your comment. No change required. Although we agree that this may not be possible in practice, theoretically *all* confounders and effect modifiers must be considered. In fact, it is for this reason that most ad hoc techniques and non-randomised evidence cannot replace RCTs and it is important that this is highlighted in this guideline. This is also the reason why a large effect is required for a clear-cut decision. We added this to the text (to be consistent with D4.3.2). |
| EFSPI | 26 | 944-6 / 6.1 | Current wording:<br>"The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present."<br><br>Suggested wording:<br>~~"The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present."~~<br><br>Reason for change: | See previous comment. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Such an absolute statement seems neither fair nor correct. The statement should be removed unless it can be substantiated with sufficient and robust evidence. | |
| Bayer | 27 | 965-1005 | Specific comment:<br><br>The distinction between overlap and positivity is not obviously. Typically, these terms are used interchangeably. Poor overlap leads to a violation of positivity. For instance, in lines 969-975: "The positivity assumption means that patients in both groups must be theoretically eligible for both treatments of interest", and there is a suspected violation of the positivity assumption with the "inclusion of patients in one treatment group, with a contraindication to the other treatment". In these cases, there would be non-overlap between treatment groups in the baseline characteristics determining different inclusion/exclusion criteria. | We prefer keeping overlap and positivity as separate entities. This was already the case in the methodological guideline. |
| EFSPI | 28 | 1013-5 / 6.2.2 | Current wording:<br>“Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality.”<br><br>Suggested wording:<br>~~“Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality.”~~<br><br>Reason for change:<br>Propensity score based methods do allow for causal inference when applied properly, though they may achieve this using assumptions that are stronger and/or more difficult to meet compared to other approaches. The Practical Guideline D431 should provide recommendations and best practices on how to use such methods. In | We disagree with deleting this sentence. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | general access to IPD is limited by study protocols (patient consent), local, federal or EU General Data Protection regulation. Population-adjusted methods using IPD for one of the trials can also generate robust results, if well applied. | |
| EFSPI | 28 | 1026-1027 | Actual wording:<br>"If shifted hypothesis testing has been used, an assessment whether this is sufficient to account 1026 for the likely magnitude of residual bias arising from missing covariates."<br><br>Proposed wording:<br>"~~If shifted hypothesis testing has been used, an assessment whether this is sufficient to account for the likely magnitude of residual bias arising from missing covariates~~."<br><br>Rationale:<br>Shifted hypotheses are only one of possible methods to ensure that the estimated treatment effect is large enough that it is not just a consequence of bias. There are no common thresholds, and the test is not scientifically established to balance RoB versus unmeasured confounding. A thorough discussion on confounding is good scientific practice. | The use of shifted hypothesis testing is described as an option.<br>Other methods are mentioned in Section 6 of D4.3.2.<br>See also our replies to the public consultation of D4.3.2 regarding this issue. |
| S. Walleser Autiero,<br>Medtronic | 28 | 1014-1014 | If the assumptions noted in the previous comment are true than the results of the PSM can be considered causal. | The text was modified. |
| Mihai Rotaru, EFPIA | 28 | 1026-1027 | Current wording: | The use of shifted hypothesis testing is described as an option. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | If shifted hypothesis testing has been used, an assessment whether this is sufficient to account for the likely magnitude of residual bias arising from missing covariates.<br><br>Suggested wording:<br><br>~~If shifted hypothesis testing has been used, an assessment whether this is sufficient to account for the likely magnitude of residual bias arising from missing covariates.~~<br><br>[~~strikethrough~~ indicates proposed deletion]<br><br>Rationale:<br><br>EFPIA would like to reiterate our concerns over the use of the test of shifted null hypothesis to judge the acceptability of conclusions drawn from population-adjusted indirect comparisons. We believe the use of this approach goes beyond the scope of the EUnetHTA21 methods guides as it provides instructions to assessors and co-assessors on the level of uncertainty that will be acceptable for pan-EU decision making. The extent of uncertainty deemed acceptable for a specific decision problem implicitly represents a value judgement, and as such, remains the responsibility of each EU MS to consider as part of its national HTA process. This is stated in the EU HTA regulation, '*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*.'[1] | Other methods are mentioned in Section 6 of D4.3.2.<br><br>See also our replies to the public consultation of D4.3.2 regarding this issue. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | We are aware similar concerns have been expressed by multiple stakeholders in the public consultation to 'D4.3.2 – Methods guideline on comparators and comparisons'[2], including the International Society of Pharmacoeconomic Research (ISPOR). Furthermore, this approach is not a recognised international standard of evidence-based medicine, as required in the Regulation (Recital 24).<br><br>We feel EUnetHTA 21's response to these legitimate public concerns: "T*he use of shifted hypothesis testing is an accepted method to ensure formally that a treatment effect exceeds a certain threshold (e.g., in noninferiority trials*)." provides insufficient justification to warrant its inclusion in D.4.3.1 and D.4.3.2. We therefore request that EUnetHTA reconsiders the use of the test of shifted null hypothesis and remove it from these methods guides.<br><br><u>References:</u><br>1. Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU (Text with EEA relevance).<br>2. EUnetHTA 21 Public Consultation of D4.3.2 – Methods Guideline on comparators and comparisons. | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 4 | 61/List of abbreviations | The effective sample size abbreviation (ESS) seems to be missing from the list. The abbreviation is not specified when the expression is first used in the text body (line 810; abbreviation used in line 812). Further, it appears to be the only time the ESS abbreviation is used; the other mentions are using the full expression (line 823, 916). | Thank you, we have corrected this. |
| Mihai Rotaru, EFPIA | 5 | 94-96 | <u>Current wording:</u><br><br>Population-adjusted method for indirect comparisons: method for indirect comparisons with adjustment for imbalances in effect modifiers | Thank you for the suggestion, we have amended the text along |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | between studies and additionally, in the case of disconnected networks, for imbalances in prognostic variables between studies.<br><br>Proposed wording:<br><br>Population-adjusted ~~method for~~ indirect comparisons: ~~method for indirect comparisons with adjustment for imbalances in effect modifiers between studies and additionally, in the case of disconnected networks, for imbalances in prognostic variables between studies.~~ **indirect treatment comparisons in which individual patient data from one (or several) trial(s) is used to adjust for population characteristics that differ between studies and that are expected to modify the treatment effect and influence the outcome.**<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>The definition of Population-adjusted indirect comparisons (PAIC) should focus on the primary characteristic of the approach, which is adjustment for imbalances in important characteristics between populations. Themes such as effect modifiers, prognostic factors, and connected/disconnected networks are not strictly needed. Therefore, these terms may make the definition unnecessarily complex and distract from the key concept. | similar lines to what you have proposed. |
| Mihai Rotaru, EFPIA | 5 | 86-87 | Current wording | Thank you, we have amended the text but avoided to explain the |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Indirect comparison: evidence synthesis in which inference about the relative effectiveness of two treatments is made without the use of trials comparing both treatments head-to-head<br><br>Proposed wording:<br><br>Indirect comparison: a broad term to refer to any evidence synthesis incorporating indirect evidence, which therefore includes NMA.<br><br>Rationale:<br><br>This definition is not aligned with the definition in the final Guideline D4.3.2, which states: "we use the term indirect comparison as the broadest term to refer to any evidence synthesis incorporating indirect evidence, which therefore includes NMA" (page 9 of final deliverable D4.3.2). This means indirect comparisons in D4.3.2 is used in a broader sense that can also contain data from direct comparisons (head-to-head trials). The terminology in the two documents needs to be consistent. We therefore propose that D.4,3,1 adopts the same definition for indirect comparison as used in the final Guideline D.4.3.2 (see P9 Section 1.2 Scoping and terminology). | term "indirect comparison" with the term "indirect evidence". |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 5 | 74/1.1 | Some of the terms used throughout the present document would be useful to define: average treatment effect (ATE); average treatment effect among treated (ATT); effect-modifiers; absolute effects; relative effects. | This is not a statistical textbook explaining all relevant terms. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Prof. Matthias P. Schönermark, M.D., Ph.D., Elisa Zavatta, M.A.<br><br>SKC Beratungsgesellschaft mbH | 5 | 80 | Original wording:<br>"Effectiveness: describes how well a treatment works in practice; […]"<br><br>Comment:<br>The original wording implies that the effectiveness of a new drug is assessed in clinical practice. However, effectiveness is mainly evaluated in the setting of clinical trials which do not necessarily reflect clinical practice in all aspects.<br><br>Suggestion for rewording:<br>"Effectiveness: describes how well a treatment works in patients; […]" | Thank you, we have made the suggested change. |
| Mihai Rotaru, EFPIA | 5 | 81 | Current wording:<br>Exchangeability: if patients from one treatment group are substituted to another, the same treatment effect is expected; contains the components similarity, homogeneity, and, in the case of indirect comparisons, consistency<br><br>Proposed wording:<br>Exchangeability: if patients from one treatment group ~~are~~ **were** substituted to another, the same treatment effect would be expected; contains the components similarity, homogeneity, and, in the case of indirect comparisons, consistency | Thank you, changed. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively]<br><br>Rationale:<br><br>Grammar (refers to completed studies, so this is an assumption, not a fact). | |
| Mihai Rotaru, EFPIA | 6 | 119-122 | **Exceptional circumstances for applying population-adjusted methods of evidence synthesis despite uncertainty or concern over their validity**<br><br>Current wording:<br><br>There might be exceptional circumstances in which methods of evidence synthesis will need to be applied despite uncertainty or doubt as to their validity. We believe that these should be kept to a minimum and only used in circumstances in which there is a lack of other options to produce an estimate of relative treatment effect.<br><br>Proposed wording:<br><br>There will be ~~exceptional~~ circumstances in which **population-adjusted** methods of evidence synthesis will need to be applied despite uncertainty or concern over their validity (**for example: when only single-arm trial or observational data are available or there is no connection with a comparator of interest or when significant imbalances in treatment effect modifiers across trials have been identified**). We believe that these methods should ~~be kept to a~~ | We think our original text is adequate. We do not want to give the expression that exceptional cases are the norm in JCAs. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | ~~minimum and~~ only used in circumstances in which there is a lack of other options to produce a reliable estimate of relative treatment effect.<br><br>[note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively]<br><br>Rationale:<br><br>We believe the proposed wording will provide assessors and co-assessors with clearer guidance on when such evidence synthesis methods will be needed in the absence of other options. Also, given that cancer treatments and ATMPs will be the first to undergo JCA, the need for such evidence synthesis methods may be more frequent than is suggested in the current wording. | |
| Mihai Rotaru, EFPIA | 6 | 103-105 | Current wording:<br><br>This Practical Guideline describes how to deal in practice with evidence syntheses in JCA reports and provides guidance for assessors and co-assessors dealing with submitted results of direct and indirect treatment comparisons from HTDs.<br><br>Proposed wording:<br><br>This Practical Guideline describes how to deal in practice with evidence syntheses in JCA reports and provides guidance for assessors and co-assessors dealing with submitted results of direct and indirect treatment comparisons from HTDs. **This Guideline also specifies requirements and recommendations to allow for a** | Thank you for your comment. The 'requirements for reporting' are aimed at assessors when authoring the final JCA report. Reporting requirements for HTDs are out of the scope of this guideline.<br><br>Therefore, we have not made the proposed change. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **thorough assessment of evidence syntheses; providing information to guide HTDs in their JCA submissions**.<br><br>[note: **bold** denotes proposed inclusion of text]<br><br><u>Rationale</u>:<br>The "requirements for reporting" sections are also of value to HTDs as they describe what is expected in the JCA submission. This should be reflected in the guideline objective, and more detail on expectations for reporting would be helpful throughout the guideline. | |
| GSK | 6 | 124-126 | Suggest to reword as:<br>"...is insufficient **<u>as the only reason</u>** to demonstrate…" | We do not think that this addition is appropriate. Other arguments are required to demonstrate the validity of results |
| Mihai Rotaru, EFPIA | 6 | 105-106 | <u>Current wording:</u><br>Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which a direct or indirect treatment comparison was submitted.<br><br><u>Proposed wording:</u><br>Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which ~~a direct or indirect treatment comparison was~~ **quantitative evidence synthesis results were** submitted. | Thank you for your comment. We have amended the text to clarify that these guidelines only apply to submissions informed by evidence synthesis results. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | [note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br>The current formulation would imply that the guideline also applies to purely RCT based dossiers. One should clarify that only evidence synthesis results are in scope. | |
| Mihai Rotaru, EFPIA | 6 | 116-117 | Current wording:<br>The aim of this Guideline is to enable HTA assessors and developers to identify potential issues and reduce bias and uncertainty as much as possible.<br><br>Proposed wording:<br>The aim of this Guideline is to enable HTA assessors and developers to identify potential issues and ~~reduce~~ **address** bias and uncertainty as much as possible.<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br>Bias and uncertainty cannot necessarily be reduced retrospectively – but they can be addressed by appropriate analysis methods and interpretation. | Thank you, we agree and have changed the text as suggested. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| Mihai Rotaru, EFPIA | 7 | 142-144 | **General considerations – Initial feasibility questions**<br><br>Current wording:<br><br>For direct and indirect comparisons by means of evidence syntheses, the aspects of the population, intervention, control, outcome (PICO) framework and the study design of the included studies have to be examined.<br><br>Proposed wording:<br><br>For direct and indirect comparisons by means of evidence syntheses, the aspects of the population, intervention, ~~control~~, **comparator**, outcome (PICO) framework and the study design of the included studies have to be examined.<br><br>[note: **bold** and ~~strikethrough~~ denote proposed insertion and deletion, respectively]<br><br>Rationale:<br><br>EFPIA wishes to indicate that PICO is described here as population, intervention, *control*, outcome. In all other documents, PICO is defined as population, intervention, *comparator*, outcome. | Thank you, we changed this. |
| GSK | 7 | 137-138 | Suggested rewording to account for particular situation of single-arm trials: | This addition is not appropriate. A |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | "(i.e., **except for single arm trials,** regarding the mean outcome in one group only)". | comparison is always required, even in the case of single-arm trials.<br><br>Possible methods are described in Section 6. |
| Matias Olsen, EUCOPE | 7 | 155-156 | Replace:<br><br>"A determination that the studies included in the evidence synthesis match the established PICO based on all information described above."<br><br>With:<br><br>"~~A determination that~~ **The extent to which** the studies included in the evidence synthesis match the established PICO based on all information described above." | Thank you, we have changed the text along the suggested lines. |
| Mihai Rotaru, EFPIA | 7 | 155-156 | **Initial feasibility questions – Requirements for reporting:**<br><br>Current wording:<br><br>A determination that the studies included in the evidence synthesis match the established PICO based on all information described above.<br><br>Proposed wording: | Thank you, we have changed the text along the suggested lines. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | A determination that the studies included in the evidence synthesis **reflect** ~~match~~ the established PICO based on all information described above.<br><br>[note: **bold** and ~~strikethrough~~ denote proposed inclusion and deletion, respectively]<br><br>Rationale:<br>EFPIA suggests an alternative given the current word ("match") implies a level of exactness that may not be intended. | |
| GSK | 7 | 136 | Suggested rewording to account for particular situation of single-arm trials:<br>"(i.e., compared with an appropriate comparator, **except for single arm trials**)". | This addition is not appropriate. A comparison is always required, even in the case of single-arm trials.<br>Possible methods are described in Section 6. |
| GSK | 7 | 155 | Suggest a word or phrase other than "match" (e.g. "correspond to" or "are consistent with"), as this currently implies a level of exactness that may not be intended. | Thank you, we have changed the text along the suggested lines. |
| EFSPI | 9 | 251-2 / 3.2.2 | Current wording:<br>"It is important to use statistical methods as well as design features of the included studies to assess heterogeneity."<br><br>Suggested wording: | Thank you for your comment; we have made a minor change to the text. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | "It is important to use statistical methods to estimate heterogeneity as well as to report design features of the included studies that may lead to ~~to assess~~ heterogeneity."<br><br>Reason for change:<br>Clearly distinguish between qualitative and quantitative assessments. | |
| Prof. Matthias P. Schönermark, M.D., Ph.D., Dr. Lydia Frick<br><br>SKC Beratungsgesellschaft mbH | 9 | 233 | Original wording:<br>• "Description of methodology used to identify potential effect modifiers and whether it sufficiently captures all possible effect modifiers;<br>• […] whether this list [of all potential effect modifiers] is likely to be complete; […]"<br><br>Comment:<br>It is not clear what the difference between the two assessments is because both require the assessor and co-assessor to evaluate whether all potential effect modifiers were identified.<br><br>Suggestion for rewording:<br>"Description of methodology used to identify potential effect modifiers and whether *the methodology is suitable to capture* all possible effect modifiers;" | Thank you for your comment; we have made the suggested change. |
| Sebastian Werner<br>vfa | 14 | S 4.1<br>415-507 | *"4.1 Methods for direct comparisons"*<br><br>Although trivial, it may be added that there are also situations where only one single comparative study is available and that these situations are out of scope for Section 4.1, especially as "direct comparisons" are defined in section 1.1 as "comparison of treatments | Indeed, this is trivial.<br>Moreover, in principle, a Bayesian meta-analysis is possible even if there is only 1 study. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | either by means of a single comparative study or a pairwise meta-analysis or other method for synthesis of comparative studies without indirect comparisons" (L77-79). | |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 15 | 505/4.1.3 | The term "qualitative summary" seems inappropriate. Please, use the term "narrative summary". | We disagree and think that this term is appropriate. |
| Mihai Rotaru, EFPIA | 16 | 545-546 | Current wording<br><br>Assessment of the graphical and tabular presentations of the evidence network, including the information on the number of randomised controlled trials (RCTs) per contrast;<br><br>Proposed wording:<br><br>Assessment of the graphical and tabular presentations of the evidence network, including the information on the number of randomised controlled trials (RCTs) **and RCTs identifiers (acronym, main publication)** per contrast;<br><br>[note: **bold** recommended added text].<br><br>Rationale:<br><br>Study identifiers are very useful as well, as not all trials have the same number of patients and/or quality and reporting only a number of included RCTs per contrast can be visually misleading. | Although RCT identifiers would be useful in tables or graphs, the addition is inappropriate here, because the RCT identifier do not have to be assessed. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | | |
| EFSPI | 16 | 532 / 4.2 | Current wording:<br>"[...] the estimated treatment effect might change from being significant in a clinical trial study to non-significant in the evidence synthesis."<br>Suggested wording:<br>"[...] the estimated treatment effect might change from being significant in a clinical trial ~~study~~ to non-significant in the evidence synthesis." | We deleted the whole sentence. |
| Sebastian Werner<br>vfa | 16 | 524 | *"…but is not limitedto.."*<br><br>Please insert a blank space ("limited to") | Thanks, changed |
| Mihai Rotaru, EFPIA | 16 | 547 | Comment:<br>The output described here is commonly referred to as 'league tables' [1]. We recommend using this terminology here as well for clarity.<br><br>Reference<br>1. Hutton et al., "The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-Analyses of Health Care Interventions: Checklist and Explanations," *Annals of Internal Medicine* 162, no. 11 (June 2, 2015): 777, https://doi.org/10.7326/M14-2385. | A league table refers mainly to the results of the NMA with full data, and not to the separate results from direct and indirect comparisons. |
| Mihai Rotaru, EFPIA | 17 | 575-576 | Current wording:<br>(Network) meta-analysis of flexible survival models [fractional polynomials (FPs) or piecewise exponential models] (see Section 4.3.3). | We modified the corresponding text. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br><br>(Network) meta-analysis of flexible survival models [fractional polynomials (FPs), **restricted cubic spline models** or piecewise exponential models] (see Section 4.3.3).<br><br>[note: **bold** denotes proposed inclusion]<br><br>Rationale:<br><br>Restricted cubic splines (RCS) represent an alternative approach for undertaking NMA of time-to-event outcomes in the presence of non-proportional hazard and therefore warrant inclusion in the bullet. Also, this would improve consistency since RCS are mentioned in Section 4.3.3. | |
| Sarah Smith, Lumanity | 18 | 602 | Should the word “digitalised” be changed to “digitized”? | Thank you, corrected. |
| Sebastian Werner<br><br>vfa | 18 | 604 | Please introduce the abbreviation “FP” at first use. | This is not the first use of the abbreviation FP (it occurs at the end of Section 4.3.1). |
| Sebastian Werner<br><br>vfa | 23 | 810 | *“effective samples size”*<br><br>Please introduce the abbreviation “ESS” here, especially as it is used again later in the text. | Thank you, change made. |
| Hervé Tchala Vignon Zomahoun; Richard Bisaillon; François Désy / INESSS | 24 | 837-838/5.5 | In this sentence, replace absolute outcomes and relative outcomes by absolute effects and relative effects respectively. Here is a suggestion: | Thank you for your suggestion. We have changed ‘relative outcomes’ to ‘relative |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Differences in patient characteristics are typically more likely to affect **absolute effects** than they are **relative effects**, which means that more covariates must be included in the adjustment model to obtain an unbiased estimate the treatment effect. | effects' as suggested, but prefer 'absolute outcomes' to 'absolute effects' since the latter is sometimes also used for treatment effects measured on the risk difference scale. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 26 | 926 - 929 | Current wording:<br><br>A key concern is that the underlying assumption of exchangeability is unlikely to hold because there is a very high risk of confounding bias, meaning that the association between intervention and outcome differs from its causal effect. Therefore, treatment comparisons based upon non-randomised evidence require careful consideration of its validity.<br><br>Suggested wording:<br><br>A key concern is that the underlying assumption of exchangeability *may not hold* because there is a very high risk of confounding bias, meaning that observed association between intervention and outcome differs from the true causal effect. *Therefore, careful consideration of the study design, population selection, and data is needed to mitigate bias.*<br><br>Rationale:<br><br>The suggested text attempts to provide clarity between the observed confounded association vs. the true exposure-outcome association. | We prefer to keep the broader original wording. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The next paragraph describes considerations for study validity. While the following paragraph describes rigorous adjustment (assuming this is statistical). It is important to note that the study design is the first step to address potential confounding, before applying statistical adjustment. | |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 26 | 941-944 | Current wording:<br><br>In some cases, it might be that the lack of randomisation can be compensated for by rigorous adjustment for confounding. However, this requires that all confounders and effect modifiers relevant for adjustment are measured and that the model and covariate selection strategies for adjustment are prespecified and based upon transparent criteria.<br><br>Suggested wording:<br><br>The lack of randomisation may be compensated for by rigorous adjustment for confounding. However, this requires that all confounders relevant for adjustment are measured and that the model and covariate selection strategies for adjustment are prespecified and based upon transparent criteria.<br><br>Rationale:<br><br>Effect modification was removed from this sentence as “Adjustment” of a variable implies statistical adjustment which is what is used for confounding variables. However, effect modifiers are distinctly different and not “adjusted” for. | We disagree with the proposed suggestion, and maintain our original wording. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Recommend discussing effect modification in a separate sentence outside the context of “adjustment”. | |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 27 | 969-970 | Current wording:<br><br>The positivity assumption means that patients in both groups must be theoretically eligible for both treatments of interest.<br><br>Suggested wording:<br><br>The positivity assumption means that each patient has a nonzero probability to receive their treatment. In other words, each patient must be theoretically eligible for both treatments of interest.<br><br>Rationale:<br><br>Suggested wording provides a little more clarity from what is meant by “theoretically” eligible. | There is no need to formulate the definition in other words. |
| Tanja Podkonjak,<br><br>Takeda Pharmaceuticals International AG | 27 | 986 | Current wording:<br><br>The populations in the compared groups must be sufficiently balanced after adjustment for confounding.<br><br>Suggested wording:<br><br>Characteristics and potential confounding factors between the two treatment groups must be sufficiently balanced after adjustment for confounding. | We prefer to keep the broader initial wording with populations |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Rationale:<br><br>The suggested wording seeks to add clarity to what specifically needs to be sufficiently balanced between two treatment groups. | |
| Roche | 5 | 86-7 / 1.1 | *"Indirect comparison: evidence synthesis in which inference about the relative effectiveness of two treatments is made without the use of trials comparing both treatments head-to-head"*<br><br>This definition is not aligned with the definition in the final Guideline D4.3.2, which states: *"we use the term indirect comparison as the broadest term to refer to any evidence synthesis incorporating indirect evidence, which therefore includes NMA"* (page 9 of final deliverable D4.3.2). This means indirect comparisons in D4.3.2 is used in a broader sense that can also contain data from direct comparisons (head-to-head trials).<br><br>The terminology should be harmonised by using the more general definition from D4.3.2 in both documents, D4.3.1 and D4.3.2. | We refer to our reply to the same comment above. |
| Roche | 5 | 94-6 / 1.1 | *Current wording:*<br><br>*"Population-adjusted method for indirect comparisons: method for indirect comparisons with adjustment for imbalances in effect modifiers between studies and additionally, in the case of disconnected networks, for imbalances in prognostic variables between studies."*<br><br>*Suggested wording:*<br><br>*"Population-adjusted indirect comparisons: method for indirect comparisons with adjustment for imbalances in effect modifiers* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | *between studies and additionally, in the case of disconnected networks, for imbalances in prognostic variables between studies.*<br><br>*indirect treatment comparisons in which individual patient data from one (or several) trial is used to adjust for population characteristics that differ between studies and that are expected to influence outcome."*<br><br>*Reason for change:*<br><br>*The definition of PAIC should focus on the primary feature of the approach, which is adjusting for imbalances in important characteristics between populations. Themes such as effect modifiers, prognostic factors, and connected/disconnected networks are important contextual considerations, but these themes are not needed to define PAIC itself. Therefore, these topics make the definition unnecessarily complex and distract from the key concept.* | |
| EFSPI | 6 | 105-6 / 2 | Current wording:<br>"Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which a direct or indirect treatment comparison was submitted."<br><br>Suggested wording:<br>"Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which ~~a direct or indirect treatment comparison was~~ evidence synthesis results were submitted."<br><br>Reason for change:<br>The current formulation would imply that the guideline also applies to | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | purely RCT based dossiers. One should clarify that only evidence synthesis results are in scope. | |
| Roche | 6 | 105-6 / 2 | Current wording:<br><br>*"Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which a direct or indirect treatment comparison was submitted."*<br><br>Suggested wording:<br><br>*"Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which ~~a direct or indirect treatment comparison was~~* ***evidence synthesis results were*** *submitted."*<br><br>Reason for change:<br><br>The current formulation would imply that the guideline also applies to purely RCT based dossiers. One should clarify that only evidence synthesis results are in scope. | We refer to our reply to the same comment above. |
| Roche | 6 | 107-10 / 2 | Current wording<br><br>*"It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so."* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested wording:<br><br>*"~~It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so.~~*<br><br>***The Practical Guideline D4.3.1 should foster the adherence to (recognized) best practices in the conduct of direct and indirect treatment comparisons. The guideline therefore discusses the assumptions, strengths and limitations of given approaches based on the available literature, which will help assessors and co-assessors to evaluate the scientific validity of presented indirect comparisons. Therefore, JCA reports should include a detailed discussion and firm conclusions on the scientific validity of presented indirect comparisons, their uncertainty, and the strengths and limitations of the underlying evidence base and/or methods.****"*<br><br>Reason for change:<br><br>The EU HTA Regulation states that JCA reports shall contain a description of the relative effects of the new intervention, the degree of certainty of the relative effects, taking into account the strengths and limitations of the available evidence (Article 9 (1)). | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The scientific assessment presented in the JCA report will then form the bases for decision making at the member state level. The current text in lines 107-10 of the draft guideline D4.3.1 undermines the objective of a joint scientific assessment at the EU level. How a direct or indirect estimate from a JCA is incorporated into decision making should indeed be left to member states, but the JCA should include a summary of the scientific validity of the estimate itself. | |
| Roche | 8 | 176 / 3.2.1 | *"including duration of follow-up"*<br><br>'Duration of follow-up' is a concept that is typically not properly defined. The Practical Guideline D4.3.1 should be made more precise. We recommend replacing the term 'duration of follow-up' with a more exact definition of what is requested. See https://arxiv.org/abs/2206.05216 for a discussion. | We refer to our reply to the same comment above. |
| EFSPI | 8 | 184-5 / 3.2.1 | "an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation."<br>The statement seems confusing as it mixes the terms 'I = Intervention' and 'C = Comparator' from the PICO. The exact meaning should be clarified. | We refer to our reply to the same comment above. |
| Roche | 8 | 198 / 3.2.1 | <u>Current wording:</u><br><br>*"If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead."* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested wording:<br><br>*"~~If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead.~~"*<br><br>Reason for change:<br><br>Use of post-baseline variables bears the risk of introducing lots of issues and biases, e.g. unclear causality, immortal bias. | |
| Roche | 9 | 249-50 / 3.2.2 | Current wording:<br><br>*"The heterogeneity between the studies has to be assessed to determine whether a pooling of the results is meaningful at all and to choose between the fixed-effect and random-effects approach for the evidence synthesis."*<br><br>Suggested wording:<br><br>*"The heterogeneity between the studies has to be assessed ~~to determine whether a pooling of the results is meaningful at all~~* ***to inform the synthesis process*** *and to choose between the fixed-effect and random-effects approach for the evidence synthesis."*<br><br>Reason for change:<br><br>Heterogeneity is in itself not a reason for not pooling studies in evidence synthesis as heterogeneity can be accounted for via random effects. Large heterogeneity will lead to larger uncertainty in the results and, therefore, weaker conclusions. Also, the Practical Guideline D431 | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | correctly points out that inconsistency must be assessed. Heterogeneity may lead to inconsistency, therefore, problematic levels of heterogeneity - leading to inconsistency - would be detected and the corresponding synthesis be flagged as problematic.<br><br>Ruling out the pooling of studies purely based on heterogeneity would unduly penalise certain disease areas, where for example endpoints are inherently variable (such as for example diseases leading to cognitive impairment, which is more difficult to quantify than overall survival, say). | |
| Roche | 9 | 251-2 / 3.2.2 | Current wording:<br><br>*"It is important to use statistical methods as well as design features of the included studies to assess heterogeneity."*<br><br>Suggested wording:<br><br>*"It is important to use statistical methods* ***to estimate heterogeneity*** *as well as* ***to report*** *design features of the included studies* ***that may lead to*** ~~*to assess*~~ *heterogeneity."*<br><br>Reason for change:<br><br>Clearly distinguish between qualitative and quantitative assessments. | We refer to our reply to the same comment above. |
| Roche | 9 | 262 / 3.2.2 | Suggest to add the following text:<br><br>***"It is important to note that $I^2$ is not an absolute measure of heterogeneity. $I^2$ is the proportion of the total (observed) variance that is due to the between-study variance. In other words, it tells*** | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **'what proportion of the observed variance would remain if one could eliminate the sampling error.'** ***[1]***" <br><br> [1] Borenstein et al., "Basics of Meta-Analysis: I2 Is Not an Absolute Measure of Heterogeneity," Research Synthesis Methods 8, no. 1 (March 2017): 5–18, https://doi.org/10.1002/jrsm.1230. | |
| Roche | 9 | 263-73 / 3.2.2 | Recommend to reword the paragraph: The Q-test is not a useful test in practice - the paragraph itself illustrates the limitations. Therefore it cannot serve as an "easy and objective criterion to decide whether the studies should not be pooled". | We refer to our reply to the same comment above. |
| | | | | |
| Roche | 10 | 289-91 / 3.2.2 | Rather than testing for homogeneity, the focus should be on quantifying between-study-heterogeneity, which is typically achieved by estimating the random effects standard deviation. An estimate of the RE SD (or Var) should be added. <br><br> One should also consider dropping the p-value of the heterogeneity test given its limitations. | We refer to our reply to the same comment above. |
| Mihai Rotaru, EFPIA | 10 | 299-302 | **Assessment of consistency** <br><br> <u>Current wording:</u> | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts.<br><br>Proposed wording:<br><br>~~Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts.~~<br><br>[note: ~~strikethrough~~ denotes proposed deletion]<br><br>Rationale:<br><br>EFPIA considers the current text potentially misleading. Inconsistency could in principle also arise for groups of very homogeneous trials, but where direct and indirect evidence are (systematically) in conflict.<br><br>Inconsistency and heterogeneity are related, but are still different concepts (and, therefore, we recommend not using the formulation that "*inconsistency is a form of heterogeneity*". | |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Roche | 10 | 299-302 | Current wording:<br><br>*"Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts."*<br><br>Suggested wording:<br><br>*"~~Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts.~~"*<br><br>Reason for change:<br><br>We consider the current text potentially misleading. Inconsistency could in principle also arise for groups of very homogeneous trials, but where direct and indirect evidence are (systematically) in conflict.<br><br>Inconsistency and heterogeneity are related, but still different concepts (and, therefore, we recommend not using the formulation that "*inconsistency is a form of heterogeneity*".) | We refer to our reply to the same comment above. |
| Mihai Rotaru, EFPIA | 12 | 377-380 | **Possible approaches when the assumptions are violated – splitting into subgroups**<br><br>Current wording: | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | **1. Splitting into subgroups**: if dissimilarity is shown for a potential effect modifier or heterogeneity is shown that can be explained by the effect modifier, it might be useful to divide the entire study pool into several subpools and draw separate conclusions (e.g., for men and women). The limitations of subgroup analyses based upon aggregated data should be taken into account [14];<br><br>Comment:<br><br>EFPIA is concerned around the proposed possible approach of disaggregation of a study population into several sub-pools in the event of dissimilarity for a potential effect modifier or heterogeneity. Whilst we acknowledge the rationale, we think the assessment of sub-populations should be wholly driven by the sub-populations identified in the PICO. As such, we recommend the addition of clarification to this section of the practical guideline. | |
| | 12 | 400 - 401 | Current wording:<br><br>*"However, these methods have numerous limitations and might not generate results that are applicable to the research question."*<br><br>Suggested wording:<br><br>*"~~However, these methods have numerous limitations and might not generate results that are applicable to the research question.~~"*<br><br>Reason for change: | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Statements that downgrade specific methods should be removed. Such statements do not support assessors with their assessment. | |
| Mihai Rotaru, EFPIA | 14 | 433-434 | Current wording:<br><br>As a general alternative to frequentist methods, a Bayesian approach can be used for meta-<br><br>analysis provided that the required prior distributions are available and can be justified [52].<br><br>Comment:<br><br>The statement that Bayesian methods can be used should be a separate paragraph. The rest of the paragraph discusses "*other methods for special situations*". The Bayesian approach represents an inferential framework applicable in all situations (line 433: "*as a general alternative to frequentist methods*"), not only in the special circumstances (such as rare events, and double-zero studies) discussed in the current paragraph (lines 426-32). | We refer to our reply to the same comment above. |
| Roche | 14 | 433-4 / 4.1.1 | The statement that Bayesian methods can be used should be a separate paragraph. The rest of the paragraph discusses "*other methods for special situations*". The Bayesian approach represents an inferential framework applicable in all situations (line 433: "*as a general alternative to frequentist methods*"), not only in the special circumstances (such as rare events, and double-zero studies) discussed in the current paragraph (lines 426-32). | We refer to our reply to the same comment above. |
| Roche | 15 | 485-7 / 4.1.3 | *"Alternatively, a random-effects Bayesian meta-analysis with weakly informative prior distribution for the heterogeneity parameter might be useful in the case of very few studies, because external heterogeneity* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | *information decreases the problem of estimating heterogeneity with insufficient data."*<br><br>We agree and suggest adding references to the informative priors based on empirical evidence by Turner et al.:<br><br>Turner et al., "Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data from the Cochrane Database of Systematic Reviews," *Int J Epidemiol* 41, no. 3 (June 2012): 818–27, https://doi.org/10.1093/ije/dys041.<br><br>M. Turner et al., "Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis," *Statistics in Medicine* 34, no. 6 (March 15, 2015): 984–98, https://doi.org/10.1002/sim.6381.<br><br>Turner et al., "Incorporating External Evidence on Between-Trial Heterogeneity in Network Meta-Analysis," *Statistics in Medicine* 38, no. 8 (2019): 1321–35, https://doi.org/10.1002/sim.8044. | |
| Mihai Rotaru, EFPIA | 15 | 485-487 | Current wording:<br><br>Alternatively, a random-effects Bayesian meta-analysis with weakly informative prior distribution for the heterogeneity parameter might be useful in the case of very few studies, because external heterogeneity | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | information decreases the problem of estimating heterogeneity with insufficient data.<br><br>Comment:<br><br>We agree and suggest adding references to the informative priors based on empirical evidence by Turner et al.:<br><br>1. Turner et al., "Predicting the Extent of Heterogeneity in Meta-Analysis, Using Empirical Data from the Cochrane Database of Systematic Reviews," *Int J Epidemiol* 41, no. 3 (June 2012): 818–27, https://doi.org/10.1093/ije/dys041.<br><br>2. M. Turner et al., "Predictive Distributions for Between-Study Heterogeneity and Simple Methods for Their Application in Bayesian Meta-Analysis," *Statistics in Medicine* 34, no. 6 (March 15, 2015): 984–98, https://doi.org/10.1002/sim.6381.<br><br>3. Turner et al., "Incorporating External Evidence on Between-Trial Heterogeneity in Network Meta-Analysis," *Statistics in Medicine* 38, no. 8 (2019): 1321–35, https://doi.org/10.1002/sim.8044. | |
| Roche | 16 | 532 / 4.2 | Current wording:<br><br>*"[...] the estimated treatment effect might change from being significant in a clinical trial study to non-significant in the evidence synthesis."*<br><br>Suggested wording: | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | *"[...] the estimated treatment effect might change from being significant in a clinical trial ~~study~~ to non-significant in the evidence synthesis."* | |
| Roche | 16 | 540 / 4.2 | The primary outcome from an evidence synthesis done for JCA is the set of relative effect estimates for the new intervention vs all relevant comparators (based on the full body of evidence), along with the associated uncertainties quantified for example as confidence intervals or credible intervals. The reporting requirements should list this output explicitly. | We refer to our reply to the same comment above. |
| Roche | 16 | 547 / 4.2 | The output described here is commonly referred to as 'league tables' [1]. We recommend using this terminology here as well for claridy.<br><br>[1] Hutton et al., "The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-Analyses of Health Care Interventions: Checklist and Explanations," *Annals of Internal Medicine* 162, no. 11 (June 2, 2015): 777, https://doi.org/10.7326/M14-2385. | We refer to our reply to the same comment above. |
| Roche | 16 | 530-2 / 4.2 | We recommend reformulating the sentence since evidence synthesis results typically differ from single RCT results if there is a) large heterogeneity, or b) discrepancy between direct and indirect evidence. But a discrepancy in the number of studies informing each contrast is, per se, not an issue (though a more balanced evidence base is preferable). | We refer to our reply to the same comment above. |
| Mihai Rotaru, EFPIA | 16 | 532 | Current wording:<br><br>[...] the estimated treatment effect might change from being significant in a clinical trial study to non-significant in the evidence synthesis. | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Proposed wording:<br><br>[...] the estimated treatment effect might change from being significant in a clinical trial ~~study~~ to non-significant in the evidence synthesis.<br><br>[note: ~~strikethrough~~ indicate suggested deleted text]. | |
| Mihai Rotaru, EFPIA | 16 | 540 | Comment:<br><br>The primary outcome from an evidence synthesis done for JCA is the set of relative effect estimates for the new intervention vs all relevant comparators (based on the full body of evidence), along with the associated uncertainties quantified for example as confidence intervals or credible intervals. The reporting requirements should list this output explicitly. | We refer to our reply to the same comment above. |
| EFSPI | 16 | 547 / 4.2 | The output described here is commonly referred to as 'league tables' [1]. We recommend using this terminology here as well for clarity.<br>[1] Hutton et al., "The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-Analyses of Health Care Interventions: Checklist and Explanations," Annals of Internal Medicine 162, no. 11 (June 2, 2015): 777, https://doi.org/10.7326/M14-2385. | We refer to our reply to the same comment above. |
| Roche | 16 | 550-1 / 4.2 | We recommend removing rankograms and SUCRAs from the list of required outputs. Alternatively, rankograms and SUCRAs should be clearly flagged as optional. These outputs have many limitations and can be misleading. | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | For additional justifications, see: Mbuagbaw et al., "Approaches to Interpreting and Choosing the Best Treatments in Network Meta-Analyses," *Systematic Reviews* 6 (2017): 79, https://doi.org/10.1186/s13643-017-0473-z. | |
| Roche | 17 | 560-4 / 4.3.1 | *"This means that the assumption must be tested"*<br><br>Testing for PH (understood as a formal hypothesis test) has the same drawback as, e.g., testing for heterogeneity: the null hypothesis is actually what you want to show (namely PH).<br><br>Proposal: rather "explore PH" assumption (e.g. through diagnostic plots) rather than use hypothesis test. We suggest to merge / align the text here with the one in Line 577ff. | We refer to our reply to the same comment above. |
| Roche | 17 | 573-6 / 4.3.1 | Current wording:<br>*"In this scenario, there are two alternative approaches that may be undertaken:"*<br><br>Suggested wording:<br>*~~"In this scenario, there are two alternative approaches that may be undertaken:~~*<br>***In this scenario, other effect measures should be considered, e.g.:"***<br><br>Reason for change: | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | The original text reads as if this is a comprehensive list of alternatives to use in case of NPH, though there are other options than the two approaches mentioned.<br><br>We also invite consideration of adding further potential effect measures to the list such as milestone comparison, weighted HR, average HR or HRs based on parametric models. | |
| Roche | 18 | 599 / 4.3.3 | The original text reads as if this is a comprehensive list of alternatives to use in case of NPH, though there are other options than fractional polynomials (FP) or piecewise exponential models. Such alternative approaches should be mentioned too. | We refer to our reply to the same comment above. |
| Roche | 20 | 678-9 / 5.1 | Current wording:<br><br>*"By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable."*<br><br>Suggested wording:<br><br>*"~~By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable.~~"*<br><br>Reason for change:<br><br>Such an absolute statement seems neither fair nor correct. Any method will lead to unreliable results if applied incorrectly. Whilst unanchored ITCs require stronger assumptions than some other | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | evidence synthesis approaches, the technique has proven to be useful in some contexts. | |
| Roche | 22 | 748-9 / 5.3 | *"In this case, standard measures of model fit, such as AIC/BIC, residual deviance, and so on, can be used to select these additional covariates."*<br><br>Such model selection procedures typically again introduce bias for the resulting estimates. We suggest rewording such that these covariates are to be selected based on substantive-matter considerations. | We refer to our reply to the same comment above. |
| Roche | 26 | 922 / 6.1 | *"All commonly encountered sources of evidence outside of RCTs are non-randomised."*<br><br>We see "randomization" vs. "non-randomization" not as a binary concept, but rather a continuum. As an example, we can complement an underpowered RCT with external data using dynamic borrowing.<br><br>While there is increased risk of bias for a trial that deviates from a fully powered RCT we would welcome acknowledgment that in absence of a fully powered RCT there is a range of designs and these designs are at different risks of bias. | We refer to our reply to the same comment above. |
| Roche | 26 | 944-6 / 6.1 | <u>Current wording:</u><br><br>*"The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present."* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Suggested wording:<br>*"~~The requirement of all confounders and effect modifiers being measured is unlikely to be met given that unknown modifiers and confounders are assumed to be always present.~~"*<br><br>Reason for change:<br>Such an absolute conclusion, which is based on an assumption (*"confounders are assumed to be always present"*), does not seem correct. The statement should be removed unless it can be substantiated with sufficient and robust evidence. | |
| Roche | 26 | 946-8 / 6.1 | Current Wording<br>*"These adjustment methods require access to the full IPD information. Aggregated data alone are not sufficient to reliably estimate treatment effects."*<br><br>Suggested Wording<br>*"These adjustment methods require access to the full IPD information**, which may not be feasible in practice. Other methods such as MAIC may be considered in situations where aggregate data can be matched to IPD. Aggregated data when matched (/compared) to other** aggregate data alone are not sufficient to reliably estimate treatment effects."* | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| | | | Reason for Change<br><br>Clarification is required here to distinguish indirect comparisons based on non-randomised evidence where IPD is available for both datasets (e.g., from the clinical trial and RWD) from the methods that evaluate IPD vs aggregate data. | |
| Roche | 28 | 1013-5 / 6.2.2 | Current wording:<br><br>*"Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality."*<br><br>Suggested wording:<br><br>*~~"Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality."~~*<br><br>Reason for change:<br><br>Propensity score based methods do allow for causal inference when applied properly, though they may achieve this using assumptions that are stronger and/or more difficult to meet compared to other approaches. The Practical Guideline D431 should provide recommendations and best practices on how to use such methods. | We refer to our reply to the same comment above. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | HOG answer |
|---|---|---|---|---|
| Mihai Rotaru, EFPIA | 28 | 1013-1015 | Current wording:<br><br>Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality.<br><br>Suggested wording:<br><br>"~~Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality.~~"<br><br>[note: ~~strikethrough~~ indicate suggested deleted text].<br><br>Rationale:<br><br>Propensity score-based methods do allow for causal inference when applied properly, though they may achieve this by using assumptions that are stronger and/or more difficult to meet compared to other approaches. The Practical Guideline D431 should provide recommendations and best practices on how to use such methods. | We refer to our reply to the same comment above. |

## Comments received after the deadline

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | Reply |
|---|---|---|---|---|
| M. Ermisch, GKV-SV | General, 6 and 7 | 126<br><br>Between 128 and 140 | We miss an explicit demand for data economy. The networks should be as small as possible and as large as necessary, and direct evidence (synthesis) should be preferred over indirect (whether Bucher, MIAC or STC) or mixed analyses (NMA or ML-NMR). Only under circumstances with no direct comparisons and/or where simultaneous comparison of several technologies represent the actual PICO, indirect comparisons or NMAs or ML-NMRs should be done. The addition of indirect comparisons just to increase the power of an otherwise sufficient trial set are not acceptable, given the inevitable uncertainties of the added indirect comparisons. (Allowing the addition of actually superfluous trials represent a `researches degree of freedom` issue analogous to the selection of covariates in MAICs etc.)<br><br>This suggestion is not only methodologically motivated but also resource-oriented. Superfluous, over-complex networks would be waste of resources at the HTDs' as well as the MSs' side – besides the potentially opaque results. | This is described in D4.3.2 (Key Points 3, Section 5.2); there is no need to repeat this in D4.3.1. |
| M. Ermisch, GKV-SV | General | | In the case that population-adjusted methods are used, a comparison between the study and target populations is required with respect to population characteristics.<br><br>This aspect should be added at all points in the guideline where population-adjusting methods are mentioned. | See Section 5.6 (Requirements for Reporting) |
| M. Ermisch, GKV-SV | 5 | 81 | "…substituted to another…" should read "…substituted by another…" | This was changed to "… were substituted into another, …" |
| M. Ermisch, GKV-SV | 7 | 168 | We suggest the following addition: "Vice versa, "homogeneity" between trials alone does not imply exchangeability – either because of counterdirectional effects of dissimilar variables not yet identified as | It is already clear that homogeneity between trials is not sufficient for exchangeability. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | Reply |
|---|---|---|---|---|
| | | | effect modifiers, or because of counterdirectional effects of unknown effect modifiers." | |
| M. Ermisch, GKV-SV | 8 | 182 | What are "other methods"? This should be specified in a guideline. | This is just a hint that the use of other methods might be possible in specific situations. |
| M. Ermisch, GKV-SV | 8 | 194 | The meaning of "…especially the study arms in which the comparator is used." is unclear to us. Assuming an RCT if a comparator is present, baseline values of the test-arm are expected to be very similar. | First, we are not dealing only with RCTs; second, later it is said that the values during the study course could be used if the baseline values are not available. |
| M. Ermisch, GKV-SV | 8 | 203 | Because it is correct what is outlined here, and because most of JCAs will deal with new technologies, the a posteriori test for effect modification from the trial(s) under investigation needs to addressed. It will not be possible to identify all relevant potential effect modifiers a priori – due to the novelty of the new technology. | It was added to the guideline that in the context of JCA "a priori" means before the evidence synthesis is performed. |
| M. Ermisch, GKV-SV | 24 | 855 | In unanchored indirect comparisons the necessity of a shifted null-hypothesis should be added, analogous to the RoB statement on p. 21: "To account for the risk of bias (RoB) because of missing or unknown effect modifiers or potential confounders, it is generally required to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified." | This is not required here, because it is mentioned in Section 5.2, which refers to all population-adjusted methods.<br><br>In Section 5.5 we argue against the use of unanchored population-adjusted methods and refer to the methods described Section 6. |
| M. Ermisch, GKV-SV | 26 | 947 | Again, non-randomised synthesis generally requires a shifted null-hypothesis: "Even if full IPD is available, to account for the risk of bias (RoB) because of missing or unknown effect modifiers or potential confounders, it is generally required to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative | In Section 6.1, the use of shifted hypothesis testing was added. |

| Comment from | Page number | Line/ section number | Comment and suggestion for rewording | Reply |
|---|---|---|---|---|
| | | | treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified." | |
| Edwards Lifesciences | General | | This guideline is highly academic however it is not relevant to the JCA for medical devices under the scope of the HTA Regulation.<br><br>**We believe the methodologies should be adaptive and flexible** in incorporating disease and intervention specific considerations for high risk implantable medical devices.<br><br>Also the different guidelines released for public consultation should be consistent and aligned. According to the guidelines D5.2 (JCA report template -in consultation) the JCA report is foreseen to present the "strength and limitation of the available evidence", however the guideline D4.3.1 stipulates that "Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so."<br><br>We would be ready to review this draft guidance once a revised version is available which would address the HTAR framework and once all the methodological draft guidelines are available. | The guideline offers very flexible options to apply the methods in various data situations. |