EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

EUnetHTA 21 wishes to thank the many organizations and individuals who have responded to the public consultation of this practical guideline. We have taken all comments into consideration and provided individual answers to them. Please note that textual/editorial/linguistic suggestions have been taken into account but are not answered here.
Given that many similar comments were made, we will try to address at least some of the main themes in the clarifying text below.

1) We want to remind that this guideline is dedicated to individual studies only (line 90: 'what is considered one data set (one sample of patients)'). Secondary use of these studies for evidence synthesis (for example indirect comparisons) is out of scope of this guideline and is addressed in two others deliverable (D4.3.1 practical guideline and D4.3.2 methodological guideline 'Direct and Indirect comparison'). For example, the use of a single arm trial only (which means without any other data or study) is in the scope of this guideline. But an indirect comparison, or comparison with an external source of data using this same single arm trial is out of scope and is not concerned by this guideline.

2) The HTAR clearly defines JCA as 'the scientific compilation and the description of a comparative analysis of the available clinical evidence on a health technology in comparison with one or more other health technologies or existing procedures' (Article 2). Considering the above, we reaffirm that individual uncontrolled studies (single arm trial, case-series for example) are of limited value in the HTAR context, because they cannot allow a comparative/relative evaluation, which is the cornerstone of HTA, as reminded in the HTAR (Article 2 and 9).

3) We received several comments discussing situation were RCTs may not be feasible, for technical, ethical or any other reasons. Under those situations, the comments suggest that the lack of randomization should be de facto acceptable and ask for the guideline to include this consideration. Firstly, it should be highlighted that there is no consensus in the scientific literature on the acceptance (or not) of such situations, which is MS dependent. Secondly, this guideline is intended for assessors and co-assessors, when assessing the submitted data which answered to the scoping process request. This guideline is focused on certainty of results, and on strengths and limitations of the submitted evidence. The acceptance (or not) for the data submitted (design), and the conclusion on the overall clinical added value to be drawn from it are to be left at MS level. Therefore, this guideline is not intended to provide a definitive list of situations where the lack of randomization would be acceptable at EU level.

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

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Name organisation & abbreviation	Country
European Organisation for Research and Treatment of Cancer (EORTC)	Belgium
International Association of Mutual Benefit Societies (AIM)	Belgium
European Union of General Practitioners/Family Physicians (UEMO)	Belgium
European Confederation of Pharmaceutical Entrepreneurs (EUCOPE)	Belgium
European Federation of Pharmaceutical Industries and Associations (EFPIA)	Belgium
Alliance for Regenerative Medicine (ARM)	Belgium
European Organisation for Research and Treatment of Cancer (EORTC)	Belgium
The European Socienty for Paediartic Oncology (SIOPE)	Belgium
Takeda Pharmaceuticals International AG	Brussels, Switzerland, local operating companies across the European Union
Edwards Lifesciences	Europe
European Federation of Statisticians in the Pharmaceutical Industry (EFSPI) HTA SIG	Europe
MedTech Europe (MTE)	Europe - Belgium
Lymphoma Coalition - Lymphoma Coalition Europe (LCE)	France
EHA	France
EURORDIS	France
Ecker + Ecker GmbH (E+E)	Germany
SKC Beratungsgesellschaft mbH (SKC)	Germany
Verband Forschender Arzneimittelhersteller (vfa) e.V	Germany
GKV-Spitzenverband (GKV-SV)	Germany
Bayer AG & Bayer Vital GmbH	Germany
German Medicines Manufacturer´s Association (BAH)	Germany

AstraZeneca (AZ)	Global (UK based)
F. Hoffmann-La Roche Ltd (Roche)	Switzerland
Medtronic	Switzerland
GSK	UK

Outside the EU

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

Comment from	**Page number**	**Line/ section number**	**Comment and suggestion for rewording**	**Editorial comment?**	**HOG response**
Mihai Rotaru – EFPIA; Anna Vicere - Vaccine Europe		**General**	**Role of non-randomised studies (NRS) and observational studies in JCA** EFPIA & Vaccine Europe believes that the guideline, in its current version, is a missed opportunity to recognize the critical role that NRS and observational studies play in addressing the challenges of evidence		See also main themes at the beginning of the document. We reaffirm that individual non-comparative studies (i.e. a non-

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			generation for some health technologies, where RCTs are not feasible for scientific, technical or ethical considerations. In particular: • For certain highly specialised technologies, such as cell and gene therapies, blinded randomisation and a comparator arm may not be feasible • For ultra-orphan or orphan diseases, as well as in paediatric indications, prevalent or incident cases may not be sufficient to provide patient numbers required for comparative studies • For highly innovative technologies, developed in patient populations with no alternative treatments and a significant risk of mortality, randomisation to an ineffective alternative would pose ethical challenges • Treatments for paediatric populations where there may be ethical considerations in the conduct of clinical trials • For technologies that have early-stage data indicating a potentially transformative impact on patient outcomes, there would be ethical and practical reasons making RCTs unfeasible • For vaccines, where observational studies can provide significant insights on the effectiveness and safety profile relevant for health technology assessment. In case of effectiveness, observation studies allow capturing severe but rare endpoints that clinical trials fail to identify due to inadequate sampling (i.e. invasive meningococcal disease). In the case of safety, observational studies can help identifying very rare but severe adverse events that cannot be captured by RCTs due to their rarity. Regulatory Authorities have long recognised these scientific, technical and ethical considerations in their endorsement and use of NRS and		comparative study 'alone') do not allow to assess a relative effect, and are therefore of limited value in the HTAR context. Using these non-comparative studies for evidence synthesis (for example indirect comparison, or comparison with an external source of data) is out of the scope of this guideline and is addressed in 4.3 'Direct and Indirect comparison'. Comparative evidence (cornerstone of the JCA, Article 2 and 9) are requested by MS through the scoping process, by defining comparator or comparators. The corresponding comparative evidence should be submitted by HTD. This guideline is not intended to define which comparative evidence should be submitted (RCT, comparative cohort study, single-arm trial with an external comparison, indirect comparison, etc.) by HTD. This guideline is intended for assessor and co-assessor, to evaluate and report the validity of clinical studies submitted by HTD to answer to the PICO request. Furthermore, the discussion related on ethical considerations to perform or not a RCT is a matter of national competence

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			observational studies for the purposes of marketing authorisations[1]. EFPIA believes that a similar nuanced approach should be reflected in the D4.6 guideline. Furthermore, EFPIA would like to emphasize that the methods applied need to be relevant for both the initial and any potential updates of JCAs (as per Article 14 of the EU Regulation 2021/2282). Particularly in such updates, NRS and observational studies might play an important role since it will reflect real world use of a given technology. The EU Regulation 2021/2282 acknowledges that for some new health technologies (e.g., orphan medicinal products) some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: "*Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products*". For orphan drugs development, clinical development is shaped by issues such as low disease prevalence, disease severity, small and heterogeneous patient populations, difficulties in patient recruitment, and limited knowledge of the natural history of disease, among others[2]. As a consequence, the design and analysis of clinical trials for these diseases becomes more difficult. Additionally, the EMA's CHMP guideline[3] suggests avoiding unnecessary clinical trials e.g., by extrapolation from a larger source population to a smaller target population, when this is appropriate. NRS and observational studies also play an important role in complementing RCTs, particularly when addressing research questions concerning the impact of novel technologies in clinical practice, or to enhance the generalisability of RCTs when faced with variability in		and is therefore out of scope of this guideline. As the present guideline is about the validity of clinical studies, and internal validity is independent of the medical context, there is no reason to address specific medical circumstances or interventions, such as orphan medicinal products, vaccines and advanced therapy medicinal products. The necessity to use NRS rather than RCT evidence will certainly be higher in these specific situations, but "adapted methodologies" (the EU-HTA-R requirement) do not imply that a study's validity increases.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			standard of care across different Regions and countries. In light of these considerations, EFPIA recommends that the guideline differentiates the issue of risk of bias from the one regarding high variability due to a limited sample size. The guideline should also clarify the role of "underpowered" RCTs that can still generate valuable evidence, particularly when complemented with NRS and/or observational studies[4]. In this context, the role of adaptive trials, with or without the use of surrogate endpoints, should explicitly be recognised as an alternative to conventional randomized controlled trials for rare disease, particularly when targeting rare cancer patients[5]. For paediatric patients, clinical investigations of medicinal products call for targeted generation of evidence, often reflected in extrapolation of data from adult populations as well as in the use of historical or concurrent controls through registry or other data source[6] . Furthermore, the guidance should elaborate on the role of single-arm interventional trials with external (e.g., historical or concurrent) controls. <u>References</u> 1. Burns, L. N. Le Roux, R. Kalesnik-Orszulak, et al, Real-World Evidence for Regulatory Decision-Making: Guidance From Around the World, Clinical Therapeutics, Volume 44, Issue 3, 2022, Pages 420-437 2. Hilgers RD, Roes K, Stallard N. Directions for new developments on statistical design and analysis of small population group trials. Orphanet J Rare Dis. 2016;11(1):78. CHMP. 3. Guideline on clinical trials in small populations. [Online] 2007. [Cited: February 1, 2013] www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			4. Hee, S.W., Willis, A., Tudur Smith, C. et al. 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 12, 44 (2017). https://doi.org/10.1186/s13023-017-0597-1 5. Krendyukov A, Singhvi S, Zabransky M. Value of Adaptive Trials and Surrogate Endpoints for Clinical Decision-Making in Rare Cancers. Front Oncol. 2021;11:636561. Published 2021 Mar 8. doi:10.3389/fonc.2021.636561 6. Reflection paper on the use of extrapolation in the development of medicines for paediatrics, 7 October 2018, EMA/189724/2018		
Mihai Rotaru - EFPIA		**General**	**Target trial framework, causality, observational studies, bias** In absence of RCTs, observational studies can be used to support causal claims. This is particularly important when RCTs cannot be conducted for ethical reasons. The guideline should explicitly mention the target trial framework, when causal inference from observational data emulates a (hypothetical) randomized trial.[1] The target trial is a useful procedure to articulate a causal question, reflecting a hypothetical randomized trial that would answer the question of interest if resource constraints or ethical issues did not preclude conducting it[2]. The process of defining a target trial aids both in the definition of research questions and in the evaluation of various observational data sources and analysis strategies. To address the research question, the target trial is then emulated using available data sources. Description of the target trial for observational studies includes details of the population, experimental intervention, comparator, and outcomes of		See also main themes 1) and 2) at the beginning of the document. As previously answered, only comparative data allow to estimate relative effect, as requested by the HTAR. Uncontrolled individual studies are of limited value in this context. We agree the emulation of a target trial framework can be a tool for helping in reducing biases and confounding inherent to uncontrolled studies such as observational cohorts. We will consider mentioning it in the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			interest. In its current version, the D4.6 guideline is vague is when it comes to observational studies looking at comparative effectiveness. The reference to the target trial framework would help in addressing this gap. The concept of the target trial can be used to avoid some common biases in observational analyses, such as immortal time bias. Bias is the systematic difference between the results of the observational study and the results expected from the target trial. Such bias is distinct from issues of generalisability (applicability or transportability) to types of individuals who were not included in the study. The target trial concept is also central to the Cochrane Collaboration's ROBINS-I tool[3] to assess risk of bias in observational studies. However, it should be recognised that no single tool currently covers all risks of bias common in non-randomised studies[4] and additional risks of bias should be reported where applicable. References 1. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016 Apr 15;183(8):758-64. 2. Hernán MA. Methods of Public Health Research — Strengthening causal inference from observational data. *New England Journal of Medicine* 2021; 385:1345-1348. 3. Sterne JA, Hernán MA, Reeves BC, Savovia J, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions BMJ 2016; 355 D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools; BMJ Open 2021;11		

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mihai Rotaru - EFPIA		**General**	**Risk of Bias (RoB) assessment** EFPIA recommends that adequate training is provided for the assessors, co-assessors and information specialist/statisticians from the assessment team of a given health technology with respect to the use of the several tools to ascertain the risk of bias. This is critical to ensure that clear and consistent decision rules are adopted, so to achieve acceptable reproducibility of the risk of bias assessments. A related recommendation is to use standardized extraction sheets when using the different tools. In addition, all RoB assessments should be described in detail in the JCA report, in order to enable readers to understand the process and the conclusions reached. EFPIA notes that the guideline does not clarify how any potential, identified RoB issue would be addressed. EFPIA recommends that no study is excluded from the JCA, based on the assessment of RoB alone. Furthermore, the D4.6 guideline should clarify whether the sensitivity analyses, described in the context of the D4.5 guideline, would be used to look at different risk of bias categories.		We agree with the need for adequate training for assessors and co-assessors, and we believe that scheduled JCA production/pilot in the EUnetHTA 21 context will provide a great opportunity to do so. Details on data extraction are of interest but these technical considerations are out of scope of this guideline. We confirm that JCA will identify strengths and limitations (including risk of bias) of the evidence allowing to answer to the scoping process request (which contains comparator or comparators). No study will be excluded based on this assessment. Sensitivity analyses is addressed on the D4.5 guideline and is out of scope of this guideline.
Mihai Rotaru - EFPIA		**General**	**Value Judgements** It is important that the guideline does not inadvertently impose what could be a value judgement around validity considerations, including suggesting effect sizes, validity thresholds, and stating that some studies have limited value. Such judgements may mean that relevant evidence to some Member States is excluded from the JCA report. As stated in the Regulation[1], '*It is necessary therefore that Union action is limited to those*		We agree the relevance of effect sizes and validity thresholds should remain at national level. However, JCA have to describe 'the degree of certainty of the relative effects, taking into account the strengths and limitations of the available evidence'

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			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.' References 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).		(Article 9). Therefore, highlighting the fact that individual uncontrolled studies could be considered of limited value (no possible estimation for relative effect) is in line with describing the limitations of this kind of evidence.
Mihai Rotaru - EFPIA		**General**	**Attrition bias, Intention-to-treat (ITT) and ICH E9 (R1) Addendum on Estimands and Sensitivity Analyses** EFPIA recommends that the D4.6 guideline makes appropriate references to the ICH E9 (R1) Addendum on Estimands and Sensitivity Analyses[1]. The Addendum, adopted by ICH in November 2019 and in the course of implementation by Health Authorities, is shaping the design, conduct and analysis of clinical trials (however, its principles are also applicable for single-arm trials and observational studies). Importantly, clinical trials templates are being updated so to reflect the Addendum (see, for example, the standard template protocol developed by TransCelerate BioPharma[2]). The Addendum acknowledges that, in presence of post-randomisation (or "intercurrent") events (e.g., background or concomitant treatments) "[..] the question remains whether estimating an effect in accordance with the ITT principle always represents the treatment effects of greatest relevance to regulatory and clinical decision making". The Addendum then proceeds by introducing the estimand framework, which: "[…] should reflect the clinical question of interest in respect of these intercurrent events". The Addendum then introduces possible strategies		Sensitivity analyses are part of the D4.5 guideline and are therefore out of scope of this guideline. In that context, the estimand framework has been introduced. Sensitivity analyses can be submitted for investigating assumptions regarding an estimand. The scope of the JCA is defined by the answer to the PICO survey, which includes the three main attributes of the estimand (population, treatment, variable). The certainty of results of the evidence submitted by an HTD will be assessed on outcome level based on this assessment scope. It does not prevent the possibility to submit evidence with an analysis strategy (as primary or sensitivity analysis) that corresponds best to the estimand targeted by a given PICO, and it

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			to reflect different questions of interest which may be posed. EFPIA believes that, in line with the principles stated in the Addendum, clinical trials should be designed with the goal to answer *the research question of interest*. As noted in the Addendum: "whilst an inability to derive a reliable estimate might preclude certain choices of strategy, it is important to proceed sequentially from the trial objectives and an understanding of the clinical question of interest, and not for the choice of data collection and method of analysis to determine the estimand" (Addendum, page 11). The D4.6 guideline should recognise that the assessment of the risk of bias for RCTs, however important, needs to be seen in the context of the trial objectives, thus recognising that strategies that depart from the ITT principle may in some cases be more relevant to address the research question of interest. For example, clinical trials in oncology are often characterised, for ethical reasons, by the possibility of participants to switch to alternative treatments to those they were initially randomised to, e.g., upon disease progression[3]. The estimate of treatment effects in these trials, particularly of long-term outcomes such as overall survival, has often been analysed using the intention-to-treat (ITT) approach, comparing patients groups on the basis of the treatment they had been randomised to, irrespective of whether treatment switching occurred and whether any subsequent therapy was received.[4] However, an ITT strategy in this context would generate a clinically meaningful comparison of two treatment arms only if subsequent therapies were already approved and, importantly, reflecting clinical practice in Member States[4]. In these cases, it has often been noted that the ITT principle would provide an underestimate of the "true" survival		does not prevent to assess the evidence submitted with an adequate assessment of the analysis strategy regarding ICEs and missing data.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			benefit associated with the new treatment[5] and be a biased result for the purposes of joint clinical assessment. Alternative strategies may be more pertinent and appropriate to answer the research question of interest in oncology trials where treatment switching takes place. Such strategies would allow to estimate treatment effects in the *hypothetical* scenario where treatment switching had not occurred, in other words, adjusting for cross-over. Methods of analyses in these circumstances would usually consist of randomisation-based (e.g., rank-preserving structural failure time model, iterative parameter estimation algorithm) or observational-based methods (e.g., Inverse Probability of Censoring Weights and Structural Nested Model (SNM) with g-estimation). The choice of the suitable analytical approach would look at issues such as the treatment crossover mechanism, the control group crossover proportion, the treatment effect associated with different patient groups, and data availability when deciding which method to use to address treatment crossover. EFPIA appreciates that strategies for handling intercurrent events need to be based on solid scientific understanding,[6] and that assumptions for primary and sensitivity estimators need to be explicitly stated and justified, whilst avoiding implausible assumptions[7]. However, this should not prevent accepting state of the art analytical methodologies for estimating treatment effects aligned with alternative strategies to the ITT principle (or treatment policy strategy), if these strategies reflect the clinical question of interest. Sensitivity analyses, as highlighted in the D4.5 guideline, will be particularly important in these instances, to assess the robustness of the estimate. Consequently, the assessment of the validity of RCTs requires an in-depth understanding of the trial objectives, trial design, data collection and methods of analysis, which should be adequately reported in the JCA report.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Further, EFPIA believes that, when assessing the validity of clinical trials for the purpose of the JCA, no strategy should, *a priori*, be seen as only acceptable for supplementary analysis for hypothesis generation or sensitivity analysis in special situations. Against this context, the Joint Scientific Consultation (JSC) will be particularly important to reflect the perspectives of different, critical stakeholders regarding the relevant estimand strategies, specifically Regulatory Authorities (such as the European Medicine Agency) and the HTA Coordination Group. References 1. International Council for Harmonisation of Technical Requirements for Human Use (ICH). E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf 2. Clark TP, Kahan BC, Phillips A, et al. Estimands: bringing clarity and focus to research questions in clinical trials. BMJ Open, 2022;12 3. Sullivan, T, R., Latimer, NR, Gray, J., et al, Adjusting for Treatment Switching in Oncology Trials: A Systematic Review and Recommendations for Reporting, Value in Health, Volume 23, Issue 3, 2020, Pages 388-396. 4. Manitz, J, Kan-Dobrosky, N, Buchner, H, et al. Estimands for overall survival in clinical trials with treatment switching in oncology. Pharmaceutical Statistics. 2022; 21(1): 150- 162. https://doi.org/10.1002/pst.2158 5. Latimer N, Lambert P.,Crowther M., Methods for estimating survival benefits in the presence of treatment crossover: a simulation study, University of Sheffield, N_Latimer_detailed.pdf 6. Bornkamp B, Rufibach K, Lin J, et al,. Principal stratum strategy:		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Potential role in drug development. Pharm Stat. 2021 Jul;20(4):737-751. doi: 10.1002/pst.2104. Epub 2021 Feb 23. PMID: 33624407. 1. Mallinckrodt, C.H., Bell, J., Liu, G. et al. Aligning Estimators With Estimands in Clinical Trials: Putting the ICH E9(R1) Guidelines Into Practice. Ther Innov Regul Sci 54, 353–364 (2020). https://doi.org/10.1007/s43441-019-00063-9		
Sebastian Werner vfa		204 - 206	*"To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data whenever useful."* Proposed rewording: To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics~~, including raw data whenever useful~~. Transparency and understandability of the results is important. This can be achieved with aggregated results, along with the underlying documentation in respect of the submitted information, thereby allowing the assessor and co-assessor to verify the accuracy of that information, such as analysis plans, study report results, and descriptive programme code if the analyses and corresponding calculations cannot be described by a specific standard method. Raw data is not required. The provision of raw data has wide implications around privacy and General Data Protection Regulation Compliance, IT-related security, cross border data transfers, and proportionality. The German Ministry of Health decided in 2019 that for reasons of data minimization according to Art. 5 I lit. c of the General Data Protection Regulation (GDPR) and the principle of proportionality, the submission of patient individual data must not be a		We agree to delete this half sentence on raw data.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			general mandatory component of dossiers submitted by HTD in German clinical assessments. The Ministry clarified that individual patient data are generally not necessary for ensuring that the evaluation process is designed appropriately and functionally. At most, the admissibility of a case-by-case request to the HTD could be considered, should this be necessary for the health technology assessment in a specific case. ▪ Decision of the German Ministry of Health on the resolution of the Joint Federal Committee pursuant to Section 91 Social Code Book V, 1. resolution of March 16, 2018, on an amendment to the Rules of Procedure (Amendment of Annex I and II to Chapter 5.) https://www.g-ba.de/downloads/40-268-5737/2018_03_16_2019-02-21_VerfO_Aenderung-Anlage-II_Kapitel-5_konsolidiert_BMG.pdf		
James Ryan AstraZeneca		General	Thank you for the opportunity to respond to the consultation. We have inputted into the EFPIA response to this consultation and want to confirm that we are aligned with their response.		Thank you!
ISPOR		General	Your sections on real-world evidence and non-randomized studies are well-written but in general reflect traditional thinking in these areas. However, recent developments in natural experiments (note the 2021 Nobel Prize in economics) and target trial emulation (note the RCT-DUPLICATE work) have highlighted the potential for valid causal inference with observational data. We also think the value of external validity that real world evidence can contribute to decision-making is understated. While we certainly support the investigation of potential biases via tools like ROBINS-I, we thought your discussion of this area could have been more forward-looking, particularly given a growing need for post-approval evaluations. A fuller explication of this viewpoint can be found in a recent Value in Health Commentary: Berger ML, Crown WC. How Can We Make More Rapid Progress in the Leveraging of Real-World Evidence by Regulatory Decision Makers? *Value in Health*, Volume 25, Issue 2, 167 – 170.		See also main themes at the beginning of the document. We reaffirm that RWD does not define a type of clinical study design and RWE can be produced with varying certainty of results for a given research question. The certainty of the result is therefore mainly determined by the study design. This guideline is intended for assessors and co-assessors, to help them in assessing the submitted evidence. We agree the emulation of a target trial framework can be a tool for helping in

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					reducing biases and confounding inherent to uncontrolled studies such as observational cohorts. We will consider to mention it in the next version of the guideline.
ISPOR		General	We would also like to encourage practices, such as study protocol registration and use of a standard study protocol template, that would improve the reliability and credibility of RWE studies in general. An ISPOR-ISPE special task force report on a standardized RWE study protocol template will be published this fall, while our position on protocol registration can be found here: Orsini LS, Berger M, Crown W, Daniel G, Eichler H-G, Goettsch W, Guerino J, Jonsson P, Lederer NM, Monz B, Mullins D, Schneeweiss S, Van Brunt D, Wang SV, Willke RJ. Improving Transparency to Build Trust in Real-World Secondary Data Studies for Hypothesis Testing—Why, What, and How: Recommendations and a Roadmap from the Real-World Evidence Transparency Initiative. *Value in Health* 2020; 23(9):1128-36 Willke RJ, Wang SV. Registering Study Protocols: Helping RWE Come of Age. *Value & Outcomes Spotlight*. Nov/Dec 2021		Protocol registration or template are out of scope of the HTAR and this guideline.
ISPOR		III. Clinical Study Designs (Lines 235-326)- 311/313	Suggestion to change the term 'intervention' to 'exposure' in the context of observational studies		Thank you! We agree with this suggestion.
ISPOR		IV. Specific strengths, weaknesses,	How should "effect modifiers" be chosen? Clinicians, comparison of outcome variable by intervention in fitted model, literature search, are there preferred variables by disease area or indication?		This guideline is not a methodological guideline and is not intended to go into details on how to perform a proper

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
		and recommendations regarding different designs (Lines 327-449)-372			adjustment for confounding. In the context of a JCA report, more details regarding this issue are dealt within the two guidelines "Direct and indirect comparisons".
ISPOR		V. Particularities (Lines 450-658)- 459-464	Is a master protocol recommended outside of these three subtypes? In previous work, this was implemented to make definitions and descriptions more consistent and comparable across individual clinical study protocols?		It is certainly useful "to make definitions [of endpoints etc.] and descriptions more consistent" (e.g. by agreeing on core outcome sets), but this issue is outside the scope of the present guideline.
ISPOR		V. Particularities (Lines 450-658)- 615	What's the meaning of 'decentralised adjudication'? Should the difference of assessment timelines for certain endpoints (e.g. pfs in oncology study) also be mentioned?		Decentralized adjudication means an adjudication process not performed by a blinded centralized clinical adjudication committee. We will consider a rewording for the next version of the draft. Timing of endpoint assessment is outside the scope of the present guideline.
ISPOR		VI. References (Lines 569-790)- 630	"observational data from routine healthcare practices, data from registries can be considered as RWD" is there a preference for specific databases or vendors by disease area?		This is out of the scope of the guideline.
Denis Lacombe EORTC	10	Line 291	It is confusing to say that cohort studies are always comparative. The majority of cohort studies are actually non comparative. It is unclear how the word 'Comparative' is used in a chapter which otherwise attempts to give definitions on clinical trial methodologies but seems to take another angle which remains unclear.		While "cohorts" can define a longitudinal follow-up of patients irrespective of a comparison or not, "cohort studies" usually means a comparison will be performed between two groups depending

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					on the presence or absence of exposure. We acknowledge this terminology is not universal within the document in the same section, but for consistency we have decided to use this one. We will however consider clarifications for the next version of the draft.
Mihai Rotaru - EFPIA	10	287-292	Current wording "Cohort studies, also known as incidence studies, longitudinal studies, follow-up studies, or prospective studies, are studies following a group of subjects (a cohort) with a common exposure or intervention over time, but without having experienced the outcome of interest at enrolment" Suggested rewording "Cohort studies, ~~also known as incidence studies, longitudinal studies, follow-up studies, or prospective studies, are studies following a group of subjects (a cohort) with a common exposure or intervention over time, but without having experienced the outcome of interest at enrolment~~ **are a type of longitudinal study that follows a group of subjects (a cohort) who share a common exposure or intervention over time. Cohort studies can be prospective or retrospective."** [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Cohort studies can be classified as prospective or retrospective based on when outcomes occurred in relation to the enrolment of the cohort[1]. Incidence studies are a subtype of longitudinal study in which the outcome measure is dichotomous[2]; however, cohort studies can have continuous outcomes.		We disagree with the statement that in a cohort study "subjects share a common exposure or intervention", because from an analytical perspective a cohort design requires that some study participants do not have the exposure of interest. Only by comparing persons with versus persons without exposure (or intervention), an association (or effect) can be estimated. See also Grimes & Schulz 2002, who define: "A cohort study tracks two or more groups forward from exposure to outcome." (https://pubmed.ncbi.nlm.nih.gov/11830217/)

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			References 1. https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_analyticoverview/ep713_analyticoverview3.html 2. Pearce N. (2012) Classification of epidemiological study design. Int J Epidemiol 41; 2 (393-397)		
Mihai Rotaru - EFPIA	10	313	Suggestion to use "exposure" rather than "intervention" in the context of observational studies.		Thank you! We agree with this suggestion.
Denis Lacombe EORTC	10	Line 291	It is confusing to say that cohort studies are always comparative. The majority of cohort studies are actually non comparative. It is unclear how the word 'Comparative' is used in a chapter which otherwise attempts to give definitions on clinical trial methodologies but seems to take another angle which remains unclear.		Duplicated comment. See answer above.
Tanja Podkonjak, Takeda	10	287-292	Current text: Cohort studies, also known as incidence studies, longitudinal studies, follow-up studies, or prospective studies, are studies following a group of subjects (a cohort) with a common exposure or intervention over time, but without having experienced the outcome of interest at enrolment. Proposed text: Cohort studies are a type of longitudinal study that follows a group of subjects (a cohort) who share a common exposure or intervention over time. Cohort studies can be prospective or retrospective. Rationale: Cohort studies can be classified as prospective or retrospective based on when outcomes occurred in relation to the enrolment of the cohort [1]. Incidence studies are a subtype of longitudinal study in which the outcome measure is dichotomous [2]; however, cohort studies can have		Duplicated comment. See answer above.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			continuous outcomes. In addition, the quality of the data is more important instead of whether the study design is prospective or retrospective. [1] https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_analyticoverview/ep713_analyticoverview3.html [2] Pearce N. (2012) Classification of epidemiological study design. Int J Epidemiol 41; 2 (393-397)		
Tanja Podkonjak, Takeda	10	Line 299-301	Current text: Case-control studies are retrospective studies that enroll patients who have experienced a particular outcome of interest ('cases'), compared with patients who have not experienced the outcome of interest but who are representative of the study population on some controlled criterion ('controls'). Proposed text: Case-control studies are studies that enroll patients who have experienced a particular outcome of interest ('cases'), compared with patients who have not experienced the outcome of interest but who are representative of the study population on some controlled criterion ('controls'). They are usually retrospective. Prospective case-control studies are less common. Rationale: Case-control studies can be either retrospective or prospective. [1] A nested case-control study in a prospective cohort can be prospective case-control study. [1] https://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_case-control/ep713_case-control3.html		We agree that case-control studies can also be performed by identifying incident rather than prevalent cases (and controls). Some researchers call this a prospective case-control study. In our view, however, it is easier to label all studies as retrospective, if data collection was essentially retrospective or scientific logic was looking backward in time from outcome to exposure.
~~MTE~~ard	10	Fig 3.1	This traditional classification of clinical studies represented in this figure is antiquated and does not necessarily reflect or consider		The reference provided is mainly not about study designs but discusses

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			contemporary study designs and future use of data and the evolution of data sources. Consider adapting the classification based on more contemporary guidance such as that developed by the NICE decision support unit for observational or non-randomised clinical evidence: TSD17-DSU-Observational-data-FINAL.pdf Please also see the list of references on these alternate guidance documents (such as those from NICE DSU) which will be added support for JCA authors.		different statistical methods that can be used for addressing confounding and biases inherent to non-randomized studies, depending on the data situation. The fact that these methods can be used does not contradict the proposed classification of the guideline. In addition, the guideline recognizes the possibility of the use of such methods (line 368-375). Lastly, the guideline only covers designs relative to individual studies. Other designs, such as evidence synthesis studies, indirect comparisons, or an external comparison between a single-arm trial and an external source of data are covered in the D4.3.1 and D4.3.2 guidelines.
Ermisch – GKV-SV	10	281	A statement should be added here regarding the high risk of result-driven analysis inherent to retrospective study designs.		In line 211, the guideline already contained a clear statement on this issue: "Data-driven statistical tests provide results of low internal validity." In our view, there is no need to restate this for retrospective studies.
Denis Lacombe EORTC	11	Figure 3.1	The table should report all forms of data collections which are described in the document. The forms of data collection described in section 5 are not all reported. If Master protocols, platform trials, basket and umbrella trials could be forms of what already appears in the figure, registries should appear as an entity. Registries could be understood as a form of cohorts but as cohorts are defined as "comparative" in this document (see comment above) it is unlikely that they fall in this category		Master protocols, platform trials, basket, umbrella, RWD, RWE, registries are described in a chapter "particularities" because while they are common terms in the realm of clinical research, they do not constitute study designs in themselves,

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					but describes either an environmental situation of data generation or a way of including participants. Basket trials for example can be either RCTs or single-arm trials. To describe them only as "basket trials" is not operational in terms of assessing the certainty of results.
Denis Lacombe EORTC	11	Chapter 4.1 general	The notion of pragmatic randomised clinical trials developed later (line 586) should be introduced herein. See also comment here under related to pragmatic trials		Pragmatic trial do not describe in itself a particular study design but is an umbrella term pertaining to less stringent eligibility criteria than other RCTs, or less stringent criteria about the intervention protocol, or the use of specific sources for collecting data. The certainty of results provided by these trials can be adequately assessed using the classification provided in the guideline.
Matias Olsen, EUCOPE	11	324-325	Add: Classification and design characteristics for each study submitted as evidence **is considered in the light of the study objective and medical context**.		We do not think this addition is useful.
Matias Olsen, EUCOPE	11	336-346	Randomisation and blinding are insufficient to ensure exchangeability and the absence of confounding biases (in order to conclude on causality of the supplementary causal effect). The population needs to be reasonably homogeneous to ensure confounders are equally distributed among the different treatment arms. Therefore inclusion/non-inclusion criteria should be tight enough to ensure exchangeability, high internal validity, and causal relationship. This will		While we acknowledge there is a balance between internal validity and external validity that needs to be pondered when designing a RCT, we do not think such a general statement is true in all situations and can be added. General statements about the way assessors and co-assessors

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			obviously have a cost on external validity. It is therefore critical that these guidelines acknowledge that the gold standard design is appropriate to maximize internal validity at the cost of external validity. Both cannot be achieved at the highest level within the same clinical study.		should assess internal validity and applicability of the clinical studies submitted as evidence are discussed throughout the guideline.
Matias Olsen, EUCOPE	11	348-350	The guideline state that depending on numerous factors certainty of results can be questioned. For transparency purposes and to guide the applicant in their development, a listing of the most important factors should be developed. This will help the applicant to be able to cross check during the development of their intervention and control the factors that may impact the certainty of results.		Elements that can threaten certainty of results especially when considering RCTs are consensual and can be found in usual scientific literature and are already summarized in Risk of Bias tools such as ROB-2, which is the tool the guideline recommends for assessors and co-assessors.
Mihai Rotaru - EFPIA	11	340-343/4.1 Randomised clinical trials: gold standard	<u>Current wording</u> "Exchangeability (i.e., if patients from one group were substituted to the other, the same treatment effect would be observed). This underlying assumption implies the absence of confounding bias (both on known and unknown confounders and effect modifiers)." <u>Suggested rewording</u> "Exchangeability (i.e., if patients from one group were substituted to the other, the same treatment effect would be observed on average). This underlying assumption implies, **on average,** the absence of confounding bias (both on known and unknown confounders and effect modifiers)." [note: **bold** denotes proposed inclusion] <u>Rationale</u> The statement is phrased in such a way to sound like it applies at the		Randomization protects against confounding because of its theoretical principle. Even if in practice both groups (of a small trial) are unevenly distributed, this should not be considered bias through confounding.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			level of an individual study (which would be inaccurate). Exchangeability is a concept about what happens on average if identical studies were ran many times with different random samples.		
Mihai Rotaru - EFPIA	11	351-360	**Cochrane ROB-2 tool** EFPIA is concerned that the Cochrane RoB 2 tool[1] does not appear to reflect the principles and provisions of the ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis[2]. In its current version, the ROB-2 tool[1] (whether for parallel trials, cluster trials or trials with cross-over) seems to foresee only two specific strategies on the basis of which treatment effects are estimated, that is, the ITT strategy and adherence to the relevant treatment schedule. The tool appears to make no mention of other potential strategies, nor to the potential methods of estimation (that is, the analytical approach) that would inform them. For example, the user of the tool is asked to: "*State your effect of interest - effect of assignment or effect of adherence*". The Addendum, however, recognises that that there may be additional, different strategies (e.g., principal stratum), which may reflect the trial objectives and research question of interest. Furthermore, it is possible that, for the same endpoint, different intercurrent events will be addressed through different strategies; it is not clear how the ROB-2 could reflect these instances of "hybrid/mixed" estimands In addition, some strategies (such as hypothetical strategies and principal stratum strategies) will rely more heavily on statistical modelling (for example, to predict outcomes under a hypothetical condition or using pre-randomisation covariates, respectively). When applying the ROB-2 tool, the appropriateness of the selected choice of data collection and method of analysis need to be considered in the context of the relevant		The scope of the JCA is defined by the answer to the PICO survey, which includes the three main attributes of the estimand (population, treatment, variable). The certainty of results of the evidence submitted by an HTD will be assessed on outcome level based on this assessment scope. It does not prevent the possibility to submit evidence with an analysis strategy (as primary or sensitivity analysis) that corresponds best to the estimand targeted by a given PICO, and it does not prevent to assess the evidence submitted with an adequate assessment of the analysis strategy regarding ICEs and missing data. Moreover, any potential revision of the ROB-2 instrument must be left to those who developed that instrument. For HTA purposes, it is best to select the best-available, validated, and practical instrument, which currently is ROB-2.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			estimand. These considerations highlight the possibility that the user of the tool may conclude an estimate being biased when it would not be, and that different users may reach different conclusions when applying the tool. EFPIA recommends that the ROB-2 tool is thoroughly reviewed and updated so to align it with the principles, provision and even the language of the ICH E9 (R1) addendum, to ensure that it is fit-for-purpose in enabling a transparent and balanced assessment of the risk of bias in RCTs, prior for its use in any potential JCAs. References 1. ROB-2: https://methods.cochrane.org/risk-bias-2 2. International Council for Harmonisation of Technical Requirements for Human Use (ICH). E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf		
Denis Lacombe EORTC	11	Figure 3.1	The table should report all forms of data collections which are described in the document. The forms of data collection described in section 5 are not all reported. If Master protocols, platform trials, basket and umbrella trials could be forms of what already appears in the figure, registries should appear as an entity. Registries could be understood as a form of cohorts but as cohorts are defined as "comparative" in this document (see comment above) it is unlikely that they fall in this category		Duplicated comment. See answer above.
Denis Lacombe EORTC	11	Chapter 4.1 general	The notion of pragmatic randomised clinical trials developed later (line 586) should be introduced herein. See also comment here under related to pragmatic trials		Duplicated comment. See answer above.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Tanja Podkonjak, Takeda	11	Section 4	**General Comment on Section 4:** While a general overview of study designs is helpful to include in any guidance, it should not be implemented as being prescriptive, but rather being considered in the context that the study designs needed for a JCA or HTA should be 'fit for purpose' given the indication and intervention being evaluated. The current guidance could also be viewed as being biased towards limitations of observational studies vs a balanced view recognizing the strengths as well given limitations such as sample size in certain conditions, ethics around blinding, or exposing patients to a 'placebo' or standard of care treatment known to have limited efficacy in situations where patient morbidity/mortality could be impacted.		This guideline is about helping assessors and co-assessors in assessing the certainty of results of evidence submitted by HTD and is not a guideline intended to discuss the context of data generation. In that regard, consensual limitations are recognized. We do not agree the guideline prescribes RCTs is the only design that will be considered in the context of JCA. For example, the guideline recognizes that advanced statistical techniques can be used for addressing biases and confounding inherent to uncontrolled studies.
Hervé Tchala Vignon, Zomahoun/ INESSS	11	343/4.1	In this section, it would be interesting to describe the allocation sequence concealment before the blinding. An appropriate allocation sequence concealment can prevent both selection and confusion biases.		The paragraph is a general statement about the main properties that makes RCTs the most consensual design for allowing counterfactual reasoning. The guideline is not a methodological guideline intended to describe every detail of the conduct of RCTs.
EFPSI	11	351	"To allow proper evaluation by member states, RoB should be assessed using ROB-2 (10)" ROB-2 addresses 'missing outcome data when estimating the effect of assignment to intervention' (Revised Cochrane risk-of-bias tool for randomized trials/RoB 2, section 6.1.1). Since outcomes may have to be considered missing depending on the estimand of interest (for example, the outcome under the hypothetical setting 'had rescue medication not been initiated' will need to be considered missing for patients that did initiate		The scope of the JCA is defined by the answer to the PICO survey, which includes the three main attributes of the estimand (population, treatment, variable). The certainty of results of the evidence submitted by an HTD will be assessed on outcome level based on this assessment scope. It does not prevent the possibility

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			rescue medication, regardless of whether the outcome actually is available) We thus recommend to apply ROB-2 on a per-estimand basis.		to submit evidence with an analysis strategy (as primary or sensitivity analysis) that corresponds best to the estimand targeted by a given PICO, and it does not prevent to assess the evidence submitted with an adequate assessment of the analysis strategy regarding ICEs and missing data. In addition, sensitivity analyses for investigating assumptions regarding an estimand can be submitted and are covered in the D4.5 guideline.
EFPSI	11	322, Figure 3.1	Figure 3.1 is too simplistic. According to ICH E10 there are different types of controls, external control being one of them, but this does not seem to be taken into account here. Proposal: as ICH E10 is clear that external data sets are a type of control arm, the guidance should cover external controls and consider it within this figure.		This guideline covers the validity of individual studies. Studies that result from the comparison between for example one single-arm trial and an external source of data as control are covered as indirect comparisons in the D4.3.1 and D4.3.2 guidelines. In addition, such study could be considered a non-randomized controlled study and is therefore covered within the classification of the guideline.
MTE	11	346-350	**Attrition bias, Intention-to-treat (ITT) and ICH E9 (R1) Addendum on Estimands and Sensitivity Analyses** The D4.6 guideline should recognise that the assessment of the risk of bias for RCTs, however important, should be seen in the context of the relevant research question addressed by a given study, thus recognising that strategies that depart from the ITT principle may in		The scope of the JCA is defined by the answer to the PICO survey. The certainty of results of the evidence submitted by an HTD will be assessed on outcome level in regards to a given PICO question..

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			some cases be more relevant to address the research question of interest. We believe that relevant clinical trials should be designed with the goal to answer *the research question of interest* Consequently, the assessment of the validity of RCTs requires an in-depth understanding of the trial objectives, trial design, data collection and methods of analysis, which should be adequately reported in the JCA report.		
Natacha Bolanos, Lymphoma Coalition	11	338	I am reluctant to describe RCTs as the gold standard because the phrase connotes perfection, and RCTs are far from being the perfect model. They may not provide the answers researchers are looking for, RCTs don't tell critical information as which patients are going to benefit from the treatment neither if the is the totality of evidence outside the trial consistent with the trial result? In addition, in RCT trial participants typically don't represent the real-world population as a whole, so results from RCTs may not apply more generally. This should somehow be underlined in the guideline.		What we meant by gold standard if the fact the characteristics of the design of an RCT are the simplest and most consensual for allowing counterfactual reasoning. We will consider a better way of conveying this idea in the next version of the draft.
François Houÿez (Eurordis)	11	329-332	"*The JCA will report the certainty of results of the relative effectiveness of the treatment(s) of interest, taking into account the strengths and limitations of the available evidence [Article 9(1)]. As previously described, the certainty of results is determined by internal validity, applicability, and statistical precision.*" For pharmaceuticals, the certainty of results (internal validity, applicability and statistical precision) were already determined by the European Medicines Agency to a large extent. Re-assessing this certainty		The HTAR requests that JCA contains an assessment of the certainty of results of the relative effectiveness of the treatment(s) of interest.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			for HTA raises a risk of re-assessing what other European experts have done already. The guidance should explain how this risk will be limited (collaboration between EMA and EUnetHTA). EU funds should not be spent on duplicating work already done. This guidance of course fully applies to clinical studies that would not have been assessed by the EMA already. For other technologies, medical devices that fall under the scope of the HTAR in particular, EUnetHTA and evaluators of the clinical studies for the authorisation of the devices in question should also collaborate to limit the risk of duplicating their respective assessment of the results' certainty.		
François Houÿez (Eurordis)	11	338-339	"*RCTs are the gold standard for evaluating causal relationships between interventions and outcomes because randomisation eliminates much of the bias inherent to other designs*" Bayesian approaches could also be added to this paragraph (they are mentioned in D4.5). Results from RCTs usually ignore prior knowledge or prior data gained outside of the RCT. When patient populations are small, ignoring data stemming from previous clinical studies is not intellectually satisfying. On the contrary, Bayesian statistical methods applied to RCT can provide information that takes into account other data not strictly limited to the results of the RCT (this can also apply to other types of clinical trials).		This guideline covers general principles of design for assessing certainty of results. It does not prevent the use of a frequentist nor Bayesian framework for performing any kind of statistical analysis.
Ermisch – GKV-SV	11	323	In figure 3.1, the branch of observational studies (see also comment on line 241) lacks a differentiation with regards to the availability of a control group.		We do not agree with this statement. Control is a characteristic of interventional studies, where the assignment to different treatment regimens is controlled by the researcher. Observational studies, which are by definition non-interventional

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					studies, cannot be controlled. They can be comparative though.
Ermisch – GKV-SV	11	251	For HTA, RoB-2 provides no relevant benefits in comparison to RoB. Its more algorithmic approach leads to more complexity and less flexibility. Its focus are study publications, while HTA should primarily take into account study reports containing more and more detailed information than publications. For study reports of RCTs, RoB-2 is less usable.		Unfortunately, the commentator's statement is not supported by any literature. According to a recent study, using RoB-2 is indeed "complex", but achieves moderate to high reliability (https://pubmed.ncbi.nlm.nih.gov/34537386/)'
Karen Facey	11	Figure 3.1	This classification of clinical studies differentiating interventional and observational studies isn't very helpful. It would be better to focus on interventional studies and expand the RCTs section to outline the range of novel designs now being used (see comments on Table 3.1).	X	Some of the particularities are described in a specific chapter. We do not think their characteristics would be better qualified by expanding the proposed classification in terms of assessing the certainty of results.
Karen Facey	11	Line 338	(It is welcomed that RCTs are still named as the gold standard for evaluating causal relationships.)		Thank you!
GSK	11	336	Should this be merged with previous lines 333-335? Sentence says "therefore" and should explain how this conclusion follows from the previous lines in a self-contained way (without the cited literature references).		We agree with this comment and change the wording.
Jasmine Toomey PHMR	11	Line 332, section 4	This section could benefit from a reference to the different tools used to evaluate RoB. They are outlined further down in the section but a brief introduction or reference to them would be beneficial.		We believe that no specific overview on RoB tools is necessary, here.
Mihai Rotaru - EFPIA	11-14	Section 4 General	**General Comment on Section 4:** While a general overview of different types of study design is helpful to include in any guidance, it should not be prescriptive, but instead its implementation should consider the context for the research question.		Duplicated comment. See answer above.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Therefore, the study design of the data informing a JCA or HTA should be 'fit for purpose' given the indication and intervention being evaluated. As written, the current guidance may be viewed as being biased towards limitations of observational studies vs providing a balanced view recognizing both the strengths and limitations of observational studies such as sample size in certain conditions, ethics of blinding, or exposing patients to a 'placebo' or standard of care treatment known to have limited efficacy in situations where patient morbidity/mortality could be impacted.		
MTE	11-14	Section 4 General	**General Comment on Section 4:** While a general overview of different types of study design is helpful to include in any guidance, it should not be prescriptive, but instead its implementation should consider the context for the research question. Therefore the study design of the data informing a JCA or HTA should be 'fit for purpose' given the Medical Device, indication and intervention being evaluated. As indicated before, the current guidance may be viewed as being biased towards the limitations of observational studies vs providing a balanced view recognising both the strengths and limitations of observational studies, such as sample size for certain MedTech, practicalities of blinding, or exposing patients to a standard of care treatment known to have limited efficacy. A balanced view is also needed on the limitations (as well as the strenght) of RCT to reflect the effectiveness effects/		Duplicated comment. See answer above.
Denis Lacombe EORTC	12	Chapter 4.2 general	The value of treatment allocation by choice rather than by chance is quite debatable: should it really be an option described in such paper as it may stimulate ill-conceived trials		While we agree, the purpose of this guideline is not to prescribed what studies should be performed but to guide assessors and co-assessors when assessing certainty of results.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Matias Olsen, EUCOPE	12	376-379	It is unclear if there will be deliberative process for the ROBIN-I assessment or no appraisal will be done. Further clarification is needed.		Assessment will be done as requested within the guideline (box after line 379). Appraisal is a concern at national level, and is therefore out of scope of this guideline.
Matias Olsen, EUCOPE	12	381-387	In some circumstances, such as areas of high unmet need and where the treatment effect is of a large magnitude, uncontrolled studies are useful to estimate the relative effect size. For OMPs, ATMPs and paediatric populations, it may not be feasible or ethical to undertake an RCT, and single arm studies and observational studies play an important role. The guidelines should consider how to assess the validity of single arm trials and eventually the usefulness based on the validity of conducting an indirect treatment comparison using an external arm without judging on the indirect comparison. This is a typical example where the segmentation of topics and the lack of consolidation of the draft guidelines is preventative of a complete view on the evidence and a comprehensive, clear, and fair view of the relative effectiveness.		There are controversies and debates in the scientific literature about situations where randomization should be de facto declared impossible and dismissed. For example, meta-epidemiological studies show that RCTs are regularly performed in rare diseases, even against placebo, provided the concept of clinical equipoise is correctly introduced to patients (see: https://pubmed.ncbi.nlm.nih.gov/26677491/). Therefore, we will not propose situations where randomization can be de facto dismissed. Nonetheless, difficulty in performing randomization is acknowledged at the beginning of the 4.4 section about cohort studies. This guideline is about the certainty of results coming from individual studies. A single-arm trial study as an only source of evidence cannot, by design, allow relative effectiveness assessment and JCA are an assessment of comparative effectiveness. We agree data from single-arm trials can be used in conjunction with an external

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					source of data for an indirect comparison. This is covered in the D4.3.1 and D4.3.2 guidelines. We will nonetheless consider for the next version of the draft if these elements need a better balance and will recall the case of external and indirect comparisons is dealt in another guideline. Moreover, as the present guideline is about the validity of clinical studies, and internal validity is independent of the medical context, there is no reason to address specific medical circumstances or interventions, such as orphan medicinal products or advanced therapy medicinal products.
Mihai Rotaru - EFPIA	12	386-387 and the Practical Guideline summary box on Page 13	Current wording "[…] In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness". Suggested rewording ~~"[…] In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness".~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale EFPIA strongly disagrees with this statement and recommends that it is removed from the guideline. In fact, HTA agencies across Europe and beyond have used such trials with state of the art methodology for the purposes of HTA and therefore the statement is not true.		There exists controversies and debates in the scientific literature about situations where randomization should be de facto declared impossible and dismissed. For example, meta-epidemiological studies show that RCTs are regularly performed in rare diseases, even against placebo, provided the concept of clinical equipoise is correctly introduced to patients. Therefore, we will not propose situations where randomization can be de facto dismissed. Nonetheless, this difficulty in performing randomization is

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Some health technologies (including orphan drugs, ATMPs, oncology drugs) will be considered for JCA on the basis of regulatory approval based on single-arm, uncontrolled clinical trials. In these instances, such studies are informative for the purpose of the JCA, particularly as part of indirect-treatment comparisons and network-meta-analyses. Uncontrolled clinical trials (i.e., single arm trials (SAT)) are valuable in new drug development on rare disease and highly targeted patient populations. In such cases, RCTs could be either unethical or unpractical. In addition, SATs can also be used for earlier phase trials to understand whether patient would benefit from the new treatment. Therefore, SATs are useful in under-recognized patient populations, despite the absence of internal controls.[1] EFPIA believes a more balanced approach to this topic is necessary, given the proposed Phase I and II JCA process, and 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. As such, EFPIA wishes to highlight this in the context of the EU HTA regulation, which states that, *'Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products (L 458/5; Section 24).'* [2] In addition, EFPIA believe the statement is subjective and constitutes a value judgement by the authors of this guide. An assessment on the importance of NRS for REA remains the responsibility of each EU MS to consider as part of its national HTA process. This is stated in the EU HTA		acknowledged at the beginning of the 4.4 section about cohort studies. This guideline is about the certainty of results coming from individual studies. A single-arm trial study as an only source of evidence cannot, by design, allow relative effectiveness assessment and JCA are an assessment of comparative effectiveness. We agree data from single-arm trials can be used in conjunction with an external source of data for an indirect comparison. This is covered in the D4.3.1 and D4.3.2 guidelines. We will nonetheless consider for the next version of the draft if these elements need a better balance and will recall the case of external and indirect comparisons are dealt in another guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			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.*' <u>References</u> 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 2. 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).		
Mihai Rotaru – EFPIA	12	365-367/4.2 Nonrandomised controlled trials	<u>Current wording</u> "However, such non-random allocation breaks the underlying assumption of exchangeability and, therefore, is likely lead to confounding bias. Thus, the estimated association between intervention and outcome is likely to be biased and will differ from its true causal effect." <u>Suggested rewording</u> "However, such non-random allocation breaks the underlying assumption of exchangeability, **making the trial results susceptible** to confounding bias. ~~Thus, the estimated association between intervention and outcome is likely to be biased and will differ from its true causal effect~~ **When confounding bias is present, the estimated association between intervention and outcome can differ from the true causal effect, on average."**		We do not think changing this statement adds value to what is said. We agree that little is known about the frequency or size of bias in NRS. In our view, however, it is sufficient to know that RCT are protected against confounding, while NRS are not (https://pubmed.ncbi.nlm.nih.gov/34800676/). In addition, several empirical studies have shown that treatment effects as estimated in NRS are frequently different from RCT effect estimates (https://pubmed.ncbi.nlm.nih.gov/26858277/,

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			[note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The word "likely" implies something is known about the frequency of bias in nonrandomised studies. However, not only is the frequency of bias in such studies not known, but one generally cannot confirm the presence or absence of bias in an individual study.		https://pubmed.ncbi.nlm.nih.gov/9794851/).
Paolo Morgese – ARM	12	380-394	For reasons related to the rarity of target diseases, the transformative impact on high unmet need conditions, the limitations in randomisation and the significant magnitude of effect, single-arm trials and non-RCT clinical evidence are often the only ethically viable clinical trial option when it comes to ATMPs. Statements like - "uncontrolled trials do not allow relative effectiveness assessment", "uncontrolled clinical trials are of very limited value for estimating treatment effectiveness", "given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials" – basically imply that JCA of many ATMPs cannot be performed. This is of major concern, because it would discriminate against therapies that have transformative potential in high unmet need indications, resulting at best in significantly delayed access. Furthermore, experience with HTA of ATMPs at member state level shows that robust methods to make indirect comparisons exist and can be used. It would be missing a major opportunity, if EU JCA were not to adopt and develop methods successfully used by national HTAs to assess ATMPs. Specifically, 1) blinded randomisation is not always possible for ATMPs, given the highly specialised nature of the technologies; 2) the prevalent and incident population sizes for orphan/ultra-orphan diseases which are often the target of ATMP development may not provide sufficient patient numbers for comparative studies; 3) in areas of high unmet need with no		There exists controversies and debates in the scientific literature about situations where randomization should be de facto declared impossible and dismissed. For example, meta-epidemiological studies show that RCTs are regularly performed in rare diseases, even against placebo, provided the concept of clinical equipoise is correctly introduced to patients. Therefore, we will not propose situations where randomization can be de facto dismissed. Nonetheless, this difficulty in performing randomization is acknowledged at the beginning of the 4.4 section about cohort studies. The term "blinded randomisation" should not be used, as it blurs the difference between randomisation and blinding. If

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			currently available treatment options and significant mortality risk (again, often the case for ATMPs), randomisation to an ineffective alternative is not always possible from an ethical perspective; and 4) for ATMPs with truly transformational potential, early signals of a major patient outcome benefit often means that trials are unblinded early, with patients allowed to cross-over to the ATMP arm, leading to randomised data that is perceived as immature. For all of these reasons specific to ATMPs, non-randomised evidence must be considered in the context of EU JCA, particularly given that ATMPs are amongst the first wave of technologies to be subject to the new regulation in 2025.		blinding is not possible, randomisation still should be adequately concealed (https://pubmed.ncbi.nlm.nih.gov/11867132/). This guideline is about the certainty of results coming from individual studies. A single-arm trial study as an only source of evidence cannot, by design, allow relative effectiveness assessment and JCA are an assessment of comparative effectiveness. Thus, statements such as "uncontrolled trials do not allow relative effectiveness assessment" are only factual. We agree data from single-arm trials can be used in conjunction with an external source of data for an indirect comparison. This is covered in the D4.3.1 and D4.3.2 guidelines. The statement "given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials" does not mean an single-arm trial if provided as the only evidence will not be assessed, it means

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					that as a single study only, their assessment does not require the use of a standardized tool (moreover as such consensual and highly validated tool is currently not available). We will nonetheless consider for the next version of the draft if these elements need a better balance and will recall the case of external and indirect comparisons are dealt in another guideline.
Denis Lacombe EORTC	12	Chapter 4.2 general	The value of treatment allocation by choice rather than by chance is quite debatable: should it really be an option described in such paper as it may stimulate ill-conceived trials		Duplicated comment. See answer above.
Tanja Podkonjak, Takeda	12	386-387	Current text: In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness. Proposed text: ~~In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness.~~ Rationale: While single arm trials may not be considered as rigorous as a randomized control trial, under certain conditions, for e.g. late line oncology, efficacy benefit from single arm studies should not be discounted as in this type of setting, it could be considered highly unethical for a HTD to conduct randomized trials. Furthermore, when evaluated in the context of an unanchored indirect comparison, single arm trials can be quite valuable in estimating relative effectiveness in the context of HTA. This context should also be added to the first point in to		Duplicated comment. See answer above.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the "Practical Guideline" box.		
EFPSI	12	382	Current wording: *Unlike comparative clinical trials, uncontrolled trials **do not allow** relative effectiveness assessment.* Proposed wording: "Unlike comparative clinical trials, uncontrolled trials **generally** do not allow relative effectiveness assessment" This language should be reformulated to allow for flexibility and allow for use of external controls where feasible. Also designs such as the threshold crossing design provides options for establishing a counterfactual in uncontrolled settings (DOI: 10.1002/cpt.515).		Use of external controls are out of scope of this guideline (see D4.3 'Direct and indirect comparison').
EFPSI	12	385	Current wording: *"patients in a single-arm trial can receive a treatment in a more-standardised manner and with a more-rigorous follow-up compared with those from a case-series."* Proposed wording: We recommend removing the analogy made with case-series. Rational: We consider the statement that single-arm trials "can be considered mostly akin to case-series" highly misleading. Ultimately, HTA will work with such evidence only when EMA has deemed the evidence sufficient to conclude a positive benefit-risk profile.		For the comparative perspective (evaluation of relative effect) of JCA, individual single-arm trials are analogous to case-series.
EFPSI	12	386-7	Current wording:		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness. Proposed wording: "Methods for comparative effectiveness assessment using single-arm trial data include the use of external controls [1], and Population Adjusted Indirect Comparisons [2-3]. Such methods often rely on strong assumptions and must be used with caution." [1] Dony Patel et al., "Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials," Value in Health 24, no. 8 (August 1, 2021): 1118–25, https://doi.org/10.1016/j.jval.2021.01.015. [2] David M. Phillippo et al., "Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal," Medical Decision Making 38, no. 2 (February 1, 2018): 200–211, https://doi.org/10.1177/0272989X17725740. [3] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio, "Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study," Research Synthesis Methods 12, no. 6 (2021): 750–75, https://doi.org/10.1002/jrsm.1511. Rational: We recall that the HTA step follows a positive Marketing Authorization. This means that ultimately, HTAs will work with single-arm trials only in cases where the EMA has deemed the evidence sufficient to conclude a positive benefit-risk. Therefore, we consider the wording too strong and judgemental. Instead, the guidance should point to suitable methods that can be used in such cases (e.g. the use of external control groups, or		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Population Adjusted Indirect Comparisons). It should be acknowledged that under certain circumstances (ethical considerations, very limited patient sample size) a controlled clinical trial might not be feasible. In those cases, as long as the HTD has provided a reasonable rationale for the conduct of an uncontrolled clinical trial and why a controlled clinical trial is not feasible, the evidence should still be considered for estimating relative treatment effectiveness. Of course this should take into account the weaknesses and limitations that might impact the certainty of results and thereby the results should be interpreted with caution.		
MTE	12	361	As per the previous comment, the distinction between non-randomised controlled/uncontrolled studies and other observational evidence particularly certain cohort study designs is nuanced and often similar in terms of RoB and methods to allow for causality. Again, more contemporary classification systems should be investigated.		Already addressed issue
MTE	12	386-387	Current wording "[…] In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness". Suggested rewording ~~"[…] In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness".~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We strongly disagree with this statement and recommend that it is removed from the guideline. Some Medical Devices will be considered for JCA on the basis of single-arm, uncontrolled clinical trials. In these instances, such studies are informative for the purpose of the JCA, particularly as part of indirect-treatment comparisons and network-meta-analyses. Uncontrolled clinical trials (i.e., single arm trials (SAT)) are valuable in MedTech		
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ingo Hantke SKC Beratungs-gesellschaft mbH	12	386 - 387	Comment: The statement that "uncontrolled clinical trials are of very limited value for estimating treatment effectiveness" is too generalized. It does not consider the difficulties and hurdles in conducting trials in (ultra) orphan diseases, especially with a high life-threatening burden of disease and/or a pediatric population. The statement should be put into context or needs to be complemented. Suggestion for rewording: "*Uncontrolled clinical trials can contribute to the overall evidence in selective occasions, such as a rare disease setting, ethical hurdles and/or dramatic treatment effects.*"		Already addressed issue
Sebastian Werner vfa	12 13	386 388 - 394	*"In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness."* *"Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials. Some tools have been developed in the past (41–44), but RoB of uncontrolled studies appears to be affected by only a few specific aspects of internal validity, such as the consecutiveness of recruitment, the prespecification of sample size and analyses, and the blinded assessment of outcomes. Nevertheless, RoB of an uncontrolled study is very unlikely to be changed*		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			by formal RoB assessment; thus, this work appears dispensable." Please remove ~~In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness.~~ ~~Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials.~~ ~~Despite some tools have been developed in the past (41–44) to assess the RoB in single-arm trials, but RoB of uncontrolled studies appears to be affected by only a few specific aspects of internal validity, such as the consecutiveness of recruitment, the prespecification of sample size and analyses, and the blinded assessment of outcomes. Nevertheless, RoB of an uncontrolled study is very unlikely to be changed by formal RoB assessment; thus, this work appears dispensable.~~ The best available evidence should always be used. Notwithstanding challenges in conducting a RoB assessment for uncontrolled trials, their role in the joint clinical assessment can still be critical, particularly for rare diseases, ATMPs and oncology drugs. In these instances, such studies are informative for the purpose of the joint clinical assessment, particularly as part of indirect-treatment comparisons and network-meta-analyses. In addition, the guidance says topics such as historical controls are out of scope. However, it is clear that historical controls, as well as other related approaches such as external controls, are highly relevant to the section on uncontrolled trials. Existing standard of care is also mentioned in Section 5.1.3 in the context of relative effectiveness assessments for umbrella trials.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			The vfa recommends forming elaborated guidance on the assessment of single-arm interventional studies using external ("historic") controls. The guidance should propose formal rules for assessing Risk of Bias of single-arm studies and of external controls, including factors that influence the degree of uncertainty. The guidance should consider that established confounder adjustment methods such as Propensity Score Matching, or Matching Adjusted Indirect Comparisons can increase the certainty of results of the analyses. The joint clinical assessment shall be based on the best available evidence.		
Ermisch – GKV-SV	12	Sect. 4.3	The discussion of uncontrolled "single arm" trials seems contradictory. If they have no role in effectiveness assessments (as the first sentence suggests) then this should be clearly said so and maintained. Consequently, further considerations on validity assessments would not be required. It they are expected to have a role, this role should be described, regarding criteria for their consideration in – likely very special – circumstances. If so, the plausible quality criteria (consecutiveness of recruitment, prespecification of sample size and analyses, blinded assessment of outcomes) will be useful und should not be forgone.		Single arm trial could have a role as evidence synthesis (indirect comparison) but is of limited value as individual study.
	12	354	"Given that ROB-2 assumes that overall RoB is performed at the outcome level, RoB should be performed for every outcome required in the assessment scope 355 (i.e., 'O', from PICO)." In German early benefit assessment ROB is not limited to outcome level, but also has to be performed on study level. Are these analysis not required in the JCA?		ROB-2 includes some items which are scored at the study level, while other items are scored for specific outcomes. This standardized approach, which is already used in many HTA agencies (including IQWiG in Germany), will also be used in EU-HTA.
	12	386	"In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness." Especially in cases of low prevalence or ethical reasons, such as in rare or		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			ultra rare diseases, 1-arm studies are often the best avialable data. For that these data must be given due recognition in the JCA.		
Karen Facey	12	Line 361 Section 4.2	As per section 3.1.2 this is a rather academic presentation about critical assessment of trials that use different forms of treatment allocation that are not random. It does not cover the trials that will be submitted for oncology, ATMPs, OMPs and high risk devices. I suggest it is deleted.	X	According to other commentators' view and our own expectations, observational studies are likely to expect in some fields of drugs.
Roche	12	366-367 / 4.2	Current wording: *"Thus, the estimated association between intervention and outcome is likely to be biased and will differ from its true causal effect"* The last two decades have shown remarkable progress in the analysis of observational data. Proper causal inference from such data is possible, though requires stronger assumptions and more complex analytical tools. We recommend reformulating the text as follows. Suggested wording: *"Proper causal effect estimates from non-randomised data are possible, though they require more sophisticated analysis methods and may rely on untestable assumptions."*	X	We think that the original formulation is fine. Whether the analysis of observational data has improved over the last two decades remains doubtful. Even with modern methods, such as propensity-score matching, observational data often gave misleading results (https://pubmed.ncbi.nlm.nih.gov/26858277/).
Roche	12	380 / 4.3	Uncontrolled trials such as single arm trials should be taken into consideration as a valid option when RCTs are not feasible. Therefore further considerations for performing a valid single arm trial would be beneficial in this guidance. In particular, we feel that the specific reference to quantitative bias assessment (QBA) should be made in this guidance [1]. QBA can be used to assess how strong a confounding effect an unknown confounder would need to have to eliminate the estimated treatment effect. It is useful in scenarios where a synthetic control arm has been developed using RWD		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			and subsequently matched to a single arm trial to act as an external control group to the clinical study. Using such methodology has recently been adopted by NICE into their RWE framework (and specifically included in guidance on minimising bias). As this guidance document relates heavily to performing valid clinical studies and controlling for bias and confounding, we feel it appropriate that guidance for uncontrolled trials is expanded to include such detail. [1] Leahy et al., "Unmeasured Confounding in Nonrandomized Studies: Quantitative Bias Analysis in Health Technology Assessment," *Journal of Comparative Effectiveness Research*, June 9, 2022, https://doi.org/10.2217/cer-2022-0029.		
Roche	12	381 / 4.3	Current wording: *"Unlike comparative clinical trials, uncontrolled trials do not allow relative effectiveness assessment (i.e., supplementary effect over comparator treatment effect)."* The statement provides a very closed interpretation of the outcomes and usefulness of uncontrolled studies. However, within many circumstances, this is the only mechanism in which the HTD can collect clinical information to inform decision making. Therefore, we propose that this language is removed or revised to incorporate a statement around the use of external control arms for comparative effectiveness. Suggested wording: ~~*"Unlike comparative clinical trials, uncontrolled trials do not allow relative effectiveness assessment (i.e., supplementary effect over comparator treatment effect)."*~~	X	Already addressed issue

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Roche	12	383-4 / 4.3	Current wording: *"In terms of strengths and weaknesses, they can be considered mostly akin to case-series."* We consider the statement that single-arm trials *"can be considered mostly akin to case-series"* highly misleading. Ultimately, HTA will work with such evidence only when EMA has deemed the evidence sufficient to conclude a positive benefit-risk profile. Therefore, we recommend removing the analogy made with case-series. Suggested wording: *"~~In terms of strengths and weaknesses, they can be considered mostly akin to case-series.~~"*	X	Already addressed issue
Roche	12	386-7 / 4.3	Current wording: *"In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness."* We recall that the HTA step follows a positive Marketing Authorization. This means that ultimately, HTAs will work with single-arm trials only in cases where the EMA has deemed the evidence sufficient to conclude a positive benefit-risk. Therefore, we consider the wording too strong and judgemental. Instead, the guidance should point to suitable methods that can be used in such cases (e.g. the use of external control groups, or Population Adjusted Indirect Comparisons). Suggested wording: *"Methods for comparative effectiveness assessment using single-arm trial data include the use of external controls [1], and Population Adjusted Indirect Comparisons [2-3]. Such methods often rely on strong assumptions and must be used with caution."*	X	Already addressed issue

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			[1] Dony Patel et al., "Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials," Value in Health 24, no. 8 (August 1, 2021): 1118–25, https://doi.org/10.1016/j.jval.2021.01.015. [2] David M. Phillippo et al., "Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal," Medical Decision Making 38, no. 2 (February 1, 2018): 200–211, https://doi.org/10.1177/0272989X17725740. [3] Antonio Remiro-Azócar, Anna Heath, and Gianluca Baio, "Methods for Population Adjustment with Limited Access to Individual Patient Data: A Review and Simulation Study," Research Synthesis Methods 12, no. 6 (2021): 750–75, https://doi.org/10.1002/jrsm.1511.		
Silke Walleser Autiero Medtronic	12	361	As per the previous comment, the distinction between non-randomised controlled/uncontrolled studies and other observational evidence particularly certain cohort study designs is nuanced and often similar in terms of RoB and methods to allow for causality. Again, more contemporary classification systems should be investigated.		Already addressed issue
GSK	12	372	How should "effect modifiers" be chosen? Clinicians, comparison of outcome variable by intervention in fitted model, literature search…? Are there preferred variables by disease area or indication?		This guideline is not a methodological guideline and is not intended to go into details on how to perform a proper adjustment for confounding.
Bayer	12/ 13	380-394	With the statement "In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness" it is deemed inadequate that evidence coming from uncontrolled trials is of limited value for performing relative effectiveness assessment. While single arm designs have many limitations, they are of value to		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			HTA. They have been used to obtain marketing authorization for specific indications where RCTs are not feasible, such as very rare diseases or in oncology or haematology or ATMP relevant indications, where assigning patients to a supposedly inferior control arm would be unethical. In addition, single arm trials could be pooled with external control arms or could be used to inform indirect treatment comparisons. Therefore, uncontrolled clinical trials/single arm trials should be accepted as the only source of evidence, especially in combination with further data from real world sources.		
Karen Facey	12-13	Line 380 Section 4.3	It is stated that uncontrolled clinical trials "do not allow"/"are of very limited value" for relative effectiveness assessment. As a result it is "deemed unnecessary to propose any formal rules for assessing risk of bias" in such studies and "this work appears dispensable". This is summarised in the practical guideline box on p13. As outlined above, given the initial focus of the HTAR is on products that may only have single arm trials, more consideration needs to be given as to how such studies will be evaluated on their own (e.g. needing to be of very high quality) as well as the to demonstrate effectiveness (augmenting the short cross referencing to the comparators guideline).	X	Already addressed issue
Denis Lacombe EORTC	13	Chapter 4.4 general	Cohorts defined as comparative and used to replace an interventional study may be misleading. The community should be encouraged to develop appropriate methodology for interventional studies using adequate randomisation on relaxed eligibility criteria (possibly pragmatic but robust trials). There is concern here that blurred methodology is being stimulated		We clearly define RCT as gold standard for allowing causal interpretation of the treatment effect estimation (4.1), and precise that 'Cohort studies can be used when allocation of an intervention in a controlled manner is deemed unethical or unfeasible' (4.4). It is not a call for blurred methodology.
Matias Olsen, EUCOPE	13	394-395	Practical Guideline Replace		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			"Evidence coming from uncontrolled trials is of very limited value for performing relative effectiveness assessment. Although the (partial) use of some tools for RoB assessment is possible, the overall conclusion on the (very limited) internal validity of uncontrolled studies is very unlikely changed by RoB assessment. Therefore, RoB assessment is not required.". With: "Evidence coming from uncontrolled trials ~~is of very limited value~~ **may be valuable** for performing relative effectiveness assessment. Although the (partial) use of some tools for RoB assessment is possible, the overall conclusion on the ~~(very limited)~~ internal validity of uncontrolled studies ~~is very unlikely~~ **may be** changed by RoB assessment **and consolidation of the dimension of certainty and contextualisation**. ~~Therefore RoB assessment is not required~~		
Matias Olsen, EUCOPE	13	413-414	Exposure may be free from recall biases if collected in databases that are reliable. Case control studies may be powerful for assessing rare events and causation may be respected. The appreciation has to be specific.		The actual formulation 'are likely' already take into account the fact that this is not a generality
Mihai Rotaru - EFPIA	13	372	Nonrandomised controlled trials, on which regulatory approval is based, will be part of the JCA, and have value in JCA when combined with appropriate comparator data and analytics. Whilst we acknowledge the challenges they present, the assessor and co-assessor may benefit from the guideline outlining general principles on how to identify effect modifiers and confounders (including considerations of both clinical and statistical principles) to ensure that state of the art methodology can be applied to reduce any potential bias and deliver on the aims of the Regulation.		Already addressed issue

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mihai Rotaru - EFPIA	13	388-394	Current wording "Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials. Some tools have been developed in the past (41–44), but RoB of uncontrolled studies appears to be affected by only a few specific aspects of internal validity, such as the consecutiveness of recruitment, the prespecification of sample size and analyses, and the blinded assessment of outcomes. Nevertheless, RoB of an uncontrolled study is very unlikely to be changed by formal RoB assessment; thus, this work appears dispensable." Suggested rewording ~~"Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA, it is deemed unnecessary to propose any formal rules for assessing RoB of single-arm trials.~~ Despite some tools have been developed in the past (41–44) to assess the RoB in single-arm trials, ~~but~~ RoB of uncontrolled studies appears to be affected by only a few specific aspects of internal validity, such as the consecutiveness of recruitment, the prespecification of sample size and analyses, and the blinded assessment of outcomes. Nevertheless, RoB of an uncontrolled study is very unlikely to be changed by formal RoB assessment; thus, this work appears dispensable." [note: ~~strikethrough~~ denotes proposed deletion] Rationale EFPIA strongly disagrees with this statement and recommends that it is removed from the guideline. Notwithstanding challenges in conducting a RoB assessment for uncontrolled trials, their role in the JCA can still be critical, particularly for rare diseases, ATMPs and oncology drugs. In these instances, such studies are highly informative for the purpose of		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the JCA, particularly as part of indirect-treatment comparisons and network-meta-analyses. EFPIA believes a more balanced approach to this topic is necessary, given the proposed Phase I and II JCA process, and 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. As such, EFPIA wishes to highlight this in the context of the EU HTA regulation, which states that, *'Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products (L 458/5; Section 24).'* [1] In addition, EFPIA believe the opening statement "Given the lower importance of uncontrolled trials for relative effectiveness assessment and HTA," is subjective and constitutes a value judgement by the authors of this guideline. An assessment on the importance of NRS for REA should remain 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.'* <u>References</u> 1. Regulation (EU) 2021/2282 of the European Parliament and of		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the Council of 15 December 2021 on health technology assessment and amending Directive 2011/24/EU		
Mihai Rotaru – EFPIA	13	402-403/4.4. Cohort studies	Current wording "Given that the intervention is not randomised between patients, the underlying assumption of exchangeability cannot hold, which is very likely to lead to confounding bias." Suggested rewording "Given that the intervention is not randomised between patients, the underlying assumption of exchangeability cannot hold, ~~which is very likely to lead~~ making the study susceptible to confounding bias." [note: ~~strikethrough~~ denotes proposed deletion] Rationale The phrase "very likely" implies something is known about the frequency of confounding bias in cohort studies, when in fact it is not possible to confirm the presence of absence of confounding bias.		We think that the original wording is fine.
Mihai Rotaru – EFPIA	13	413-414/4.5 Case-control studies	Current wording "Moreover, case-control studies are also likely to lead to a measurement bias (e.g., recall bias), because exposure is measured after the onset of the disease or outcome." Suggested rewording "Moreover, case-control studies are ~~likely to lead~~ susceptible to measurement bias (e.g., recall bias), because exposure is measured after the onset of the disease or outcome" [note: ~~strikethrough~~ denotes proposed deletion]		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Rationale The word "likely" implies the frequency of measurement bias is known.		
Denis Lacombe EORTC	13	Chapter 4.4 general	Cohorts defined as comparative and used to replace an interventional study may be misleading. The community should be encouraged to develop appropriate methodology for interventional studies using adequate randomisation on relaxed eligibility criteria (possibly pragmatic but robust trials). There is concern here that blurred methodology is being stimulated		Already addressed issue
Hervé Tchala Vignon, Zomahoun/ INESSS	13	395/4.4	As mentioned in another comment above, it could be confusing to name the exposure as an intervention in the observational studies (here, cohort studies). This potential confusion can lead to an inappropriate choice of the risk of bias assessment tool. For example, observational studies like cohort studies should be assessed with ROBINS-E, not with ROBINS-I. Please, clarify it and report the appropriate tool if judged relevant. ***Relevant reference:*** *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.*		Already addressed issue
MTE	13	388-395	The wording and tone of this section on single-arm studies reads very negatively. All evidence included for review should be critically appraised on their individual merits and how they ultimately address the research questions posed by the JCA. Negative dismissal on		Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			basis of study design before careful examination of the conduct of the study and how it is reported etc. is not considered a best practice approach to HTA.		
MTE	13	402-405	The statement here that exchangeability does not hold and therefore any causal effect should be interpreted with caution or will be inherently biased is not reflective of all cohort studies. This conclusion has large implications for limiting the use of RWD.		Already addressed issue
Ermisch – GKV-SV	13	396	See comment on line 271 – delete "deemed"		As previously discussed, there is no consensus on that.
Ermisch – GKV-SV	13	397-398	Cohort studies do not allow for larger sample size or longer follow-up due to design advantages. They have been described to be administratively cheaper, which could (if the same amount of money is invested – result in larger sample size or longer follow-up. However, this is not always the case. The sentence should be changed accordingly, e.g.: "…Compared to interventional studies, they can be administratively cheaper and, thus, allow for larger sample sizes and longer follow-up,…"		Economic or administrative concerns are out of scope of this guideline.
Ermisch – GKV-SV	13	407	Delete the parenthesis. Rare diseases do not equal rare outcomes.		We will modify it for the next version of the draft.
Ermisch – GKV-SV	13	410	Change the sentence to: „Thus, they are at high risk of especially selection and recall bias." Recall bias is as prominent in case-control studies as selection bias, due to the retrospective design and other biases must be considered.		We will modify the next version of the draft.
Karen Facey	13 14	Line 395, Line 406 Line 425 Line 436	Similar to the previous comment about the section on non-randomized controlled trials, the sections on cohort studies, case controlled trials, cross sectional studies, case series and case reports is rather academic and as such studies are unlikely to be the key evidence to demonstrate treatment effectiveness, they could be summarised in a few sentences or deleted.		We prefer to keep them as originally written.
Roche	13	390 / 4.3	Please include a link to the following reference: "threshold-crossing", https://pubmed.ncbi.nlm.nih.gov/27650716/	X	No need to further develop this point.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Silke Walleser Autiero Medtronic	13	388-395	The wording and tone of this section on single-arm studies reads very negatively. All evidence included for review should be critically appraised on their individual merits and how they ultimately address the research questions posed by the JCA. Negative dismissal on basis of study design before careful examination of the conduct of the study and how it is reported etc. is not considered a best practice approach to HTA.		Already addressed issue.
Silke Walleser Autiero Medtronic	13	402-405	The statement here that exchangeability does not hold and therefore any causal effect should be interpreted with caution or will be inherently biased is not reflective of all cohort studies. This conclusion has large implications for limiting the use of RWD.		Already addressed issue.
Matias Olsen, EUCOPE	14	433-434	Replace: "Evidence coming from cross-sectional studies is of very limited value for performing relative effectiveness assessment.". With: "Evidence coming from cross-sectional studies ~~is of very limited value for performing relative effectiveness assessment.~~ **must be discussed in light of the study objective**".		We reaffirm the original formulation.
Matias Olsen, EUCOPE	14	446-449	It is considered in this guideline that case-series and case report studies have very limited validity and therefore should not be analysed for a risk of bias (RoB). Under specific circumstances, when the effect size is outstanding and refer to a patient relevant endpoint such as mortality or a severe morbidity, evidence coming from case series should be considered. Furthermore, case series need to be discussed specifically in the medical context and RoB		See main themes discussed at the beginning of the document. We reaffirm that these studies are of limited value and should not be analysed for RoB.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			should be assessed.		
Mihai Rotaru – EFPIA	14	450	Please add "fit for purpose" design to the scientific question in this introduction.		The words "fit for purpose" are too broad and address the feasibility of the study, which is out of the scope of the present guideline.
Mihai Rotaru – EFPIA	14	462-3	Current wording "Usually, the concept of a master protocol encompasses three subtypes…" Suggested rewording "Usually, the concept of a master protocol encompasses three subtypes, **however others may arise in the future which are deemed relevant for this approach**". [note: **bold** denotes proposed inclusion] Rationale EFPIA recommends to add a statement to acknowledge that trial methodology keeps advancing and improving and new designs/methods may be available in the future. It needs to be acknowledged that this guidance only highlights three particularities, but that others may arise in the future which would need to be considered.		As this sentence starts with the word "usually", there is enough room for future (even unusual) developments in trial methodology. No change required.
EFPSI	14	441-443	Current wording: *"Any effect estimate generated from a study lacking a control group is only a pre–post change, thus the interpretation of such change as a causal effect requires the very unlikely assumption that no change would have occurred without the intervention."* It should be noted that a pre-post change may hold relevance for		In terms of methodology, we think our proposition is factual and do not think it adds value to the document to discuss ambiguous contextual interpretations.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			quantifying the range of effects, if accompanied by contextual considerations. For example, a pre-post change may arguably be so dramatic that it cannot meaningfully be ascribed to regression to the mean or trial effects.		
EFPSI	14	456	Please add a statement to acknowledge that trial methodology keeps advancing and improving and new designs / methods may be available in the future. It needs to be acknowledged that this guidance only highlights three particularities; however, others may arise in the future which would need to be considered.		Already addressed issue.
MTE	14	446-449	As per previous comment on single arm studies, the wording of the BOX contents reads as a negative dismissal of case-series and case reports. Value of these study designs is based on how well they answer the research questions of the JCA which is context specific.		See main themes discussed at the beginning of the document.
MTE	14	462-464	Please indicate where platform trials, basket and umbrella trials fit (even conceptually) within Fig 3.1 or provide a new classification diagram to improve understanding of these trials.		Already addressed issue
Ermisch – GKV-SV	14	427	Please add: "Cross-sectional studies are primarily used for the determination of prevalence and have no role in estimation of effectiveness of health technologies."		Limited value is already stated in the box.
Roche	14	450 / 5	Please add "fit for purpose" design to the scientific question in this introduction.	X	Duplicated comment.
Roche	14	456 / 5	Please add a statement to acknowledge that trial methodology keeps advancing and improving and new designs / methods may be available in the future. It needs to be acknowledged that this guidance only highlights three particularities; however, others may arise in the future which would need to be considered.		Duplicated comment.
Roche	14	463 / 5	Current wording: *"Usually, the concept of a master protocol encompasses three subtypes"*	X	Already addressed issue

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We propose expanding the statement as methods can evolve and other alternatives for master protocol trials may be added. Suggested wording: *"Usually, the concept of a master protocol encompasses three subtypes, however others may arise in the future which are deemed relevant for this approach."*		
Silke Walleser Autiero Medtronic	14	446-449	As per previous comment on single arm studies, the wording of the BOX contents reads as a negative dismissal of case-series and case reports. Value of these study designs is based on how well they answer the research questions of the JCA which is context specific.		See main themes discussed at the beginning of the document.
Silke Walleser Autiero Medtronic	14	462-464	Please indicate where platform trials, basket and umbrella trials fit (even conceptually) within Fig 3.1 or provide a new classification diagram to improve understanding of these trials.		Duplicated comment.
GSK	14	459-464	Is a master protocol recommended outside of these three subtypes? In previous work, this was implemented to make definitions and descriptions more consistent and comparable across individual clinical study protocols.		We agree standardization of definitions and descriptions are advisable, but these issues are outside of the scope of this guideline.
Mihai Rotaru – EFPIA	15	5.1.1 Platform trials	Current wording "Platform trials are mainly phase 3 RCTs …"; "Therefore, platform trials can provide the same certainty of results as well-performed RCTs providing they are conducted in conformity with the same methodological principles." Comment Inconsistent language is used throughout Section 5.1.1 when describing platform trials. The first example phrase above makes it clear that some platform trials can be considered RCTs. The second example makes it sound as if platform trials are separate and different from RCTs. The		We will consider if this issue needs rewording for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			guideline should use consistent, aligned language.		
Mihai Rotaru – EFPIA	15	465	There is a mixing of concepts in section 5.1.1 between master protocols and adaptive trials, e.g., seamless designs with master protocols. These are different concepts and address different scientific questions. This section of the guideline should be revisited to clarify these differences.		We agree there is some kind of mixing, but these concepts are not mutually exclusive, and furthermore, some of them such as adaptive trials are umbrella terms which encompasses various situations. But, we will consider if this section needs rewording for the next version of the draft.
Mihai Rotaru – EFPIA	15	473-475	Current wording "Platform trials are mainly phase 3 RCTs (i.e., a confirmatory assessment of effectiveness), but they sometimes start as phase 2 trials (i.e., an exploratory assessment of effectiveness, which can be uncontrolled), and the switch from phase 2 to phase 3 is conducted under the same master protocol (this is sometimes called a 'seamless' design) (47). In that case, the most promising interventions based on the results of the phase 2 trial are retained for the phase 3 trial. Therefore, the follow-up of some patients from a phase 2 trial can be extended to the phase 3 trial (providing they meet the phase 3 eligibility criteria)." Suggested rewording "Platform trials can be either Phase 2 or 3 trials in which the majority of patients were assigned by randomization. Compared to umbrella or basket studies, Platform studies are more often Phase 3 studies (47). Additionally, platform studies can also switch from Phase 2 to phase 3 during the same master protocol. However, distinctions should be made regarding whether the switch is between operational and inferential. Operational seamless designs (only patients from Phase 3 contributes to the Phase 3 inference) usually do not pose methodological issues whereas, inferential seamless design (patients from Phase 2 are used for		We will consider if this section needs rewording for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the Phase 3 inference) may require appropriate statistical methods to ensure validity of the inference." Rationale The statement on platform trials is incorrect. In our experience, platform trials are primarily non-pivotal. While platform trials can be Phase 3 studies, it is not correct to say they are mainly Phase 3 trials (Park et al. 2019[1] found <50% were Phase 3), although (as proposed in the revised text) they are more likely to be Phase 3 relative to umbrella/basket trials. Additionally, the description of the seamless studies lacked clarity around the design of the trial. We have therefore provided additional language to highlight the differences in validity between the type of seamless design. References: 1. Park, J.J.H., Siden, E., Zoratti, M.J. et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials 20, 572 (2019). https://doi.org/10.1186/s13063-019-3664-1		
Mihai Rotaru – EFPIA	15	480	EFPIA invites a discussion on the use of master protocols for exploratory and confirmatory purposes. "Adaptivity" in a statistical sense is not an issue in exploratory master protocols (where no type I error control is needed), but it becomes much more relevant in confirmatory master protocols.		We do not think this discussion between exploratory and confirmatory purposes is necessary within this guideline.
Mihai Rotaru – EFPIA	15	480-485	Current wording "Methodologically, the main strength of platform trials is their adaptive nature. " Suggested rewording "Methodologically, the main strength of platform trials is their ~~adaptive~~		We will consider this change in wording for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			flexible nature" [note: ~~strikethrough~~ denotes proposed deletion] Rationale In its current version, the statement is confusing two terms, flexibility to add new arms and take arms out vs adaptations that are pre-specified in the protocol.		
Mihai Rotaru – EFPIA	15	484-487	Current wording "Thus, platform trials offer the potential to generate comparative evidence for multiple treatments that are simultaneously in clinical development and could reduce the need to use indirect comparison methods, such as network meta-analyses, for assessing the relative effectiveness of multiple interventions." Suggested rewording "Thus, platform trials offer the potential to generate comparative evidence for multiple treatments that are simultaneously in clinical development ~~and could reduce the need to use indirect comparison methods, such as network meta-analyses, for assessing the relative effectiveness of multiple interventions."~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale The statement is subjective, and the need of an indirect-treatment comparison or network-meta analysis is based more on the selection of the PICO than purely on methodological considerations. Platform trials are usually not designed to compare between active arms, albeit they may have a common control arm.		We agree to delete this statement in the next version of the draft.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mihai Rotaru – EFPIA	15	507 – Practical Guidance	Current wording "If the platform trial starts as a phase 2 trial, then the rules to select interventions that are going to phase 3" Suggested rewording "If the platform trial starts as a phase 2 trial, then **please consider** the rules to select interventions that are going to phase 3". [note: **bold** denotes proposed inclusion] Rationale Please modify the statement as suggested, as it contains an incomplete sentence.		We will consider a more appropriate wording for the next version of the draft.
Mihai Rotaru – EFPIA	15	507 – Practical Guidance	Current wording "Design considerations when adding new intervention(s) (criteria, process, or timing) to the trial." Suggested rewording "Design considerations when **conducting an open label study or** adding new intervention(s) (criteria, process, or timing) to the trial." [note: **bold** denotes proposed inclusion] Rationale Please add in a note to the bullet on study design to call out the open-label tendencies of these types of trials (as highlighted in the previous paragraph – "Third, although blinding of patients and investigators is possible, it requires the use of multiple dummies, which can be difficult to achieve when there are multiple treatments with different pharmaceutical		The statement is a requirement for the proper reporting of design considerations precisely when adding new interventions. It is not about blinding which is already addresses at the beginning of the practical guidance (same principle of reporting than section 4.2).

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			formulations that are assessed simultaneously. Thus numerous platform trials are conducted in an open manner.")		
Mihai Rotaru – EFPIA	15	510	Current wording "… a specific mechanism of action" Suggested rewording "…a specific **target within the** mechanism of action" [note: **bold** denotes proposed inclusion] Rationale The proposed language more accurately describes the aim of the basket study. Please use consistent terminology and descriptors of basket trials. The terminology used currently varies from MOA to processes to specific risk factor to molecular entities. We would suggest stating it as a specific target throughout and use (e.g., as a way to embellish what this could entail).		We do not think this precision is useful. Moreover, we do not agree the definition of basket trials should be restricted to molecular target only.
Mihai Rotaru - EFPIA	15	512-13	Current wording "Therefore, the targeted intervention is supposed to produce a beneficial effect for all patients because it targets a common process". Suggested rewording "Therefore, the targeted intervention **can** ~~is supposed to~~ produce a beneficial effect for all patients because ~~it targets a common process~~ **the intervention has a common molecular target that is independent of a specific disease.**" [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively]		It is correct to state that basket trials are designed on the assumption that the intervention is supposed to produce an effect on all patients. Moreover, we do not agree the definition of basket trials should be restricted to molecular target only.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Rationale Proposed alternative language better reflect this critical aspect of health technologies assessed in basket trials. Please use consistent terminology and descriptors of basket trials. The terminology used currently varies from MOA to processes to specific risk factor to molecular entities. We would suggest stating it as a specific target throughout and use (e.g. as a way to embellish what this could entail).		
MTE	15	473-479	Again, the terms used here, i.e., phase 2, 3 trials, seamless design, adaptive trial need to be included in the classification of trial designs Fig 3.1 before they can be introduced.		Already addressed issue
Ermisch – GKV-SV	15	487	Please add the clause: "… or ease their applicability." In cases, where direct comparisons of different arms of a platform trial cannot be performed due to methodological issues, the uniform design inherent to platform trials still provides advantages regarding the possibility and validity of indirect comparisons.		As discussed above, this statement will be deleted for the next version of the draft.
Ermisch – GKV-SV	15	507 (box)	The sentence „If the platform trial starts a phase 2 trial, then the rules to select interventions that are going to phase 3" is a stub and should be revised. As the classification of "phase" is not applied to medical devices the limitation to drug trial should be made explicit.		We will consider if a more appropriate wording is necessary for the next version of the draft.
Roche	15	465 / 5.1.1	There is a mixing of concepts in section 5.1.1 between master protocols and adaptive trials. e.g. seamless designs with master protocols. They are quite different sometimes and also answer different scientific questions. Proposal: Please revisit this section to clarify these differences.		Duplicated comment.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Roche	15	473-475 / 5.1.1	Current wording: *"Platform trials are mainly phase 3 RCTs (i.e., a confirmatory assessment of effectiveness),"* The language as currently stated is misleading. While platform trials can be Phase 3 studies, it is not fair to say they are mainly phase 3 trials. Please modify the language around the current statement on platform studies. Suggested wording: *"Compared to umbrella or basket studies, Platform studies are more often phase 3 studies (response based on article 47, Park et al 2019)."*	X	Already addressed issue
Roche	15	480-485 / 5.1.1	Current wording: *"Methodologically, the main strength of platform trials is their adaptive nature."* The current statement is confusing two terms: flexibility to add new arms in and take arms out vs adaptations that are pre-specified in the protocol. We suggest modifying the current language to accurately reflect the intention of the statement. Suggested wording: *"Methodologically, the main strength of platform trials is their flexible nature."*	X	Already addressed issue
Roche	15	485-486 / 5.1.1	Current wording: *"Thus, platform trials offer the potential to generate comparative evidence for multiple treatments that are simultaneously in clinical development and could reduce the need to use indirect comparison methods, such as network meta-analyses, for assessing the relative effectiveness of multiple interventions."*	X	Already addressed issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			This statement around use of ITCs should be removed as these studies are usually not designed to compare between active arms, although they may have a common control. Additionally, the need for an ITC should result from the agreed upon PICO. The ITC statement here is therefore, unrelated to the use of such methods. Suggested wording: *"Thus, platform trials offer the potential to generate comparative evidence for multiple treatments that are simultaneously in clinical development ~~and could reduce the need to use indirect comparison methods, such as network meta-analyses, for assessing the relative effectiveness of multiple interventions.~~"*		
Roche	15	507 - Practical Guideline / 5.1.1	Current wording: *"If the platform trial starts as a phase 2 trial, then the rules to select interventions that are going to phase 3."* This is an incomplete sentence. Please modify the statement. Suggested wording: *"If the platform trial starts as a phase 2 trial, then* please consider *the requirements to select interventions that are going to phase 3."*	X	We will consider a more appropriate wording for the next version of the draft.
Roche	15	507 - Practical Guideline / 5.1.1	Please add in a bullet about the implications of conducting an open-label study. This is currently addressed within the paragraph but left out of the summary guidance table.		The first sentence of the guidance already addresses what assessor do for reporting common elements of designs such as blinding (by using the same methodology described in section 4.2).
Silke Walleser Autiero Medtronic	15	473-479	Again, the terms used here, i.e., phase 2, 3 trials, seamless design, adaptive trial need to be included in the classification of trial designs Fig 3.1 before they can be introduced.		Already addressed issue.
Mihai Rotaru	16	515-517	Current wording		In all fairness, one of the first trial that

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
- EFPIA			"Basket trials are mainly used in oncology for assessing the effectiveness of interventions designed to target specific molecular alterations, but other medical areas can be concerned by the use of such trials" Suggested rewording "Basket trials ~~are mainly~~ **have historically been** used in oncology for assessing the effectiveness of interventions designed to target specific molecular alterations, ~~but other areas can be concerned by the use of such trials~~ **(i.e. biomarkers), however, this may be expanded to other medical areas of study. "** [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The proposed language more accurately describes the current setting of use and the future options where basket studies may be applicable.		can be considered a basket trial was not conducted in oncology, so "historically" is not strictly accurate. We will add that there are "currently" mainly used in oncology. We do not agree to add the word "biomarkers" as the word describes a type of outcomes and we do not want to convey the idea that because basket trials in oncology assess treatment that target specific molecular alterations, the effectiveness should be assessed via biomarkers.
Mihai Rotaru - EFPIA	16	519	Current wording "a specific risk factor" Suggested rewording: "targeting a specific **mechanism of action**~~risk factor~~" [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Please use consistent terminology and descriptors of basket trials. The terminology used currently varies from MOA to processes to specific risk		We will correct it for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			factor to molecular entities. We would suggest stating it as a specific target throughout and use (e.g. as a way to embellish what this could entail). A trial evaluating risk factors is very different than one assessing a specific target.		
Mihai Rotaru - EFPIA	16	521	Please consider including some wording around the expansion of basket trials to broader populations through the use of RWD. The use of different RWD sources allow the generalizability/representativeness of the clinical trial population to be assessed qualitatively. Similarly, the pooling or linking of separate datasets allows to deliver larger and more representative or complete datasets.		RWD are already covered in a separate section. We do not think there is a need for such expansion within the guideline.
Mihai Rotaru - EFPIA	16	524	Current wording "therefore, do not provide a higher certainty of results compared with single-arm trials" Suggested rewording "therefore, ~~do not~~ provide a ~~higher~~ similar certainty of results ~~compared with single-arm trials~~ **as other uncontrolled trials mentioned previously**". [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The term "higher" denotes a judgement on the level of evidence and information submitted. The statement should be more general in regard to the limitations of the type of evidence submitted.		In our view, the reformulated sentence does not constitute an improvement.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mihai Rotaru - EFPIA	16	536	Current wording "Third, the specific effect of the targeted intervention in a specific 'cohort' (e.g., patients with breast cancer only) can suffer from a lack of statistical precision because it can be expected that some cohorts will have a low number of patients given that the occurrence of the targeted risk factor can be rare" Suggested rewording ~~"Third, the specific effect of the targeted intervention in a specific 'cohort' (e.g., patients with breast cancer only) can suffer from a lack of statistical precision because it can be expected that some cohorts will have a low number of patients given that the occurrence of the targeted risk factor can be rare"~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale Given the research question defined within a basket study, we believe that the third point raised around the statistical limitations should be removed from the paragraph or should be rephrased as a reason to prohibit the analysis of subgroup/cohorts to be conducted in these types of studies. These studies should not be divided by disease type for the analysis of subgroups/cohorts; therefore, this statement is contradictory to the objective and design of the study.		We do not agree with this proposal as the fact that the treatment will benefit all patients in the same way is, before the conduct of the trial, still mainly an assumption based on a pathophysiological rationale only, and it should therefore be assessed experimentally based on the results of the basket trial.
Mihai Rotaru	16	543	Current wording		A documentation only is insufficient.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
- EFPIA			"…must be known and must be of an acceptable level" Suggested wording "must be **documented** ~~known and must be of an acceptable level~~" [note: ~~strikethrough~~ denotes proposed deletion, **bold** proposed additiona] Rationale The proposed reference (51) only discusses that these factors need to be documented. Acceptable levels may infer a value judgement and should be excluded.		Because, if the performance of the test is poor, it should not be used, while indeed documented.
Mihai Rotaru - EFPIA	16	544	Current wording "Moreover, the test must be available for all potentially eligible patients" Suggested rewording ~~Moreover, the test must be available for all potentially eligible patients~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale There is so much variable access to NGS/advanced diagnostic testing in clinical practice, within and between countries. Therefore, a statement mandating availability does not address the barriers that exist to improve availability and access. The lack of funded testing should not prohibit the ability to move forward with scientific advances in targeted therapies.		We will remove this sentence for the next version of the draft.
Mihai Rotaru - EFPIA	16	544 - Practical	Current wording		We do not think we need to change the text as the text describes general issues

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
		Guideline	"If the eligibility of patients within the basket trial relies on the results of a companion test, its performance, availability, and methods used for detection (e.g., on which tumor sample the test is performed)". Recommendation Please ensure that the statements included within the practical guideline match the text provided within the prior paragraph and within the umbrella trial section. The following text was included but appears to have no mention of testing methods. Please update for consistency in messaging. "Finally, eligibility criteria often rely on the screening of a specific molecular alteration or biomarker. Therefore, inclusion within a basket trial often relies on the results of a companion test and, therefore, the performance of the test (sensitivity, specificity, predictive values, or probability reports, calibration, and discriminatory capacity for biomarkers measured on a continuum) must be known ~~and must be of an acceptable level~~ (51). "		regarding basket trials while the guidance specifies what assessors and co-assessors should report.
Mihai Rotaru - EFPIA	16	544 - Practical Guideline	Current wording "If an interaction test for homogeneity of effect was performed, its method and result". Recommendation The statement is incomplete. please modify the language so that it is clear what is intended to be asked.		We will consider a more appropriate wording within the next version of the draft.
EFPSI	16	536	Given the research question defined within a basket study, we believe that		Already addresses issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the third point raised around the statistical limitations should be removed from the paragraph or should be rephrased as a reason to prohibit the analysis of subgroup/cohorts to be conducted in these types of studies. These studies should not be divided by disease type for the analysis of subgroups/cohorts therefore, this statement is contradictory to the objective and design of the study.		
Sebastian Werner vfa	16	523 - 525	*"Although basket trials can be RCTs, most are currently uncontrolled trials and, therefore, do not provide a higher certainty of results compared with single-arm trials (47)."* Comment: The best available evidence should always be used.		The purpose of this guideline is to factually describe limitations inherent to certain study designs independent of the medical context.
Ermisch – GKV-SV	16	Sec. 5.1.2	It is important to maintain this precise description of methodological issues, basket trials bring with them.		Thank you.
Roche	16	509-517 / 5.1.2	This paragraph uses a mix of different terms to identify the type of target that is being measured within a basket trial (e.g. MOA, risk factor, molecular entities). To clarify the intention, we propose that the paragraph is rewritten to consistently use the same terminology and descriptors of basket trials. We propose using "specific target" to identify the intended goal of the study.		We will consider the most adequate terminology for the next version of the draft.
Roche	16	510 / 5.1.2	Current wording: *"a specific mechanism of action"* Propose update for consistency in terminology used. Suggested wording: *"a specific* target within the *mechanism of action"*	X	Duplicated comment.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Roche	16	512-513 / 5.1.2	Current wording: *"Therefore, the targeted intervention is supposed to produce a beneficial effect for all patients because it targets a common process."* Propose update for consistency in terminology used. Suggested wording: *"Therefore, the targeted intervention can ~~is supposed to~~ produce a beneficial effect for all patients because the intervention has a common molecular target that is independent of a specific disease."*	X	Duplicated comment.
Roche	16	515-517 / 5.1.2	Current wording: *"Basket trials are mainly used in oncology for assessing the effectiveness of interventions designed to target specific molecular alterations, but other medical areas can be concerned by the use of such trials"* Propose update for consistency in terminology used. Suggested wording: *"Basket trials have* historically been *used mostly in oncology for assessing the effectiveness of interventions designed to target specific molecular alterations (i.e. biomarkers),* however, this may be expanded to other medical areas of study.*"*	X	Duplicated comment.
Roche	16	519 / 5.1.2	Current wording: *"a specific risk factor"* Propose update for consistency in terminology used. Suggested wording: *"a specific target"*	X	Duplicated comment.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Roche	16	524 / 5.1.2	Current wording: *"therefore, do not provide a higher certainty of results compared with single-arm trials"* Suggested wording: *"therefore, ~~do not~~ provide a ~~higher~~ similar certainty of results ~~compared with single-arm trials~~ as other uncontrolled trials mentioned previously."*	X	Duplicated comment.
Roche	16	536-538 / 5.1.2	Current wording: *"Third, the specific effect of the targeted intervention in a specific 'cohort' (e.g., patients with breast cancer only) can suffer from a lack of statistical precision because it can be expected that some cohorts will have a low number of patients given that the occurrence of the targeted risk factor can be rare."* The use of the term "cohort" within this statement goes against the principles in which the basket trial is defined. The basket is defined by the specific molecular target and designed to measure efficacy outcomes by the defined basket. Splitting out cohorts by histology (e.g. breast, lung) goes against the principles in which the study was designed and therefore, should not be considered appropriate. Proposal: Analyses conducted on these cohorts should be acknowledged to be inappropriate and text should be modified to reflect the intention of the designed measurement of the study.		We do not agree with this proposal as the fact that the treatment will benefit all patients in the same way is, before the conduct of the trial, still mainly an assumption based on a pathophysiological rationale only, and it should therefore be assessed experimentally based on the results of the basket trial.
Roche	16	537-539 / 5.1.2	This statement needs to be accompanied by an understanding that the estimation of effectiveness within each cohort of a basket trial can be challenging in practice, for example due to very small cohort sizes. Such cohort specific estimates are typically not part of the primary study objective and should, therefore, be interpreted with utmost caution.		We do not agree with this proposal as the fact that the treatment will benefit all patients in the same way is, before the conduct of the trial, still mainly an assumption based on a pathophysiological rationale only, and it should therefore be

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					assessed experimentally based on the results of the basket trial. This statement is here precisely to indicate the analysis of the plausibility of the assumption of homogeneous effect based on the data of the trial can be challenging because of small sample size.
Roche	16	543 / 5.1.2	Further guidance should be included as to the level of acceptability required from a specific test. Please provide clarity on what is meant by it "must be of an acceptable level" . Additionally, the requirements of companion diagnostics should be included within the discussions during the JSC. The discussion during this engagement with the HTD should provide a baseline understanding on diagnostic requirements.		The guideline addresses the general limitations of basket trials, and one of the limitations is the performance of the companion test. But "acceptable level" can be subject to appraisal, and this is why the practical guidance suggests an adequate reporting within the JCA. We agree this can be discussed during JSCs (which are not meant to be "an engagement" but are consultations).
Roche	16	544 - Practical Guideline / 5.1.2	We propose the removal of availability as a standard requirement regarding the important properties to evaluate within a diagnostic test. With the intention of understanding the value of the medicinal product, availability does not contribute to the certainty of the test nor the value of the product. Likewise, availability is not further defined regarding the parameters in question. Proposal: remove availability from the text and list of information required on a companion test.		Already addressed issue.
Bayer	16/ 17	544 (box below)	"Results of effectiveness within each 'cohort' of patients with appropriate statistical estimates". We completely disagree to claim separate analysis for each patient group within a basket trial as, obvoulsy, this contradicts		Already addressed issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the objective, research question and design of basket trials. We therefore suggest to avoid sub-group analyses within basket trials.		
Mihai Rotaru - EFPIA	17	544 - Practical Guideline	Current wording: "Results of effectiveness within each 'cohort' of patients with appropriate statistical estimates." Suggested rewording Results of effectiveness within each 'cohort' of patients with appropriate statistical estimates. **The estimation of effectiveness within each cohort of a basket trial can be challenging in practice, for example due to very small cohort sizes. Such cohort specific estimates are typically not part of the primary study objective and should, therefore, be interpreted with utmost caution. Best practices based on the scientific literature on subgroup analysis [1] and on Bayesian Hierarchal modelling [2-4] should be considered to inform these analyses.** [note: **bold** denotes proposed inclusion] Rationale The current statement needs to be accompanied by an understanding that the estimation of effectiveness within each cohort of a basket trial can be challenging in practice, for example due to very small cohort sizes. The recommended additional wording is important to properly contextualise and clarify the preceding statement, with the inclusion of references and caveats. References 1. Xin Sun et al., "How to Use a Subgroup Analysis: Users' Guide to the Medical Literature," JAMA 311, no. 4 (January 22, 2014): 405–11, https://doi.org/10.1001/jama.2013.285063.		Already addresses issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			2. Scott M Berry et al., "Bayesian Hierarchical Modeling of Patient Subpopulations: Efficient Designs of Phase II Oncology Clinical Trials," Clinical Trials 10, no. 5 (October 1, 2013): 720–34, https://doi.org/10.1177/1740774513497539. 3. Yiyi Chu and Ying Yuan, "A Bayesian Basket Trial Design Using a Calibrated Bayesian Hierarchical Model," Clinical Trials 15, no. 2 (April 1, 2018): 149–58, https://doi.org/10.1177/1740774518755122. 4. Laurie J. Hannigan et al., "Improving the Estimation of Subgroup Effects for Clinical Trial Participants with Multimorbidity by Incorporating Drug Class-Level Information in Bayesian Hierarchical Models: A Simulation Study," Medical Decision Making, August 18, 2021, 0272989X211029556, https://doi.org/10.1177/027		
Mihai Rotaru - EFPIA	17	555	Current wording "[…] development of stratified medicine" Suggested rewording "development of ~~stratified medicine~~ **personalized medicine"** [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Please consider changing the terminology to encompass the broader term of personalized medicine.		In our view, the term "personalized medicine" is partly misleading, because it suggests the use of participative, patient-centered approaches. The World Health Organization (WHO) has argued that "stratification" more accurately "reflects the realistic effects of medicines at population level, while the term 'personalized medicine' reflects the possibly overambitious promise of individualized unique drug targeting and development" (cited after: https://pubmed.ncbi.nlm.nih.gov/31708685/).
Mihai Rotaru	17	559	Current wording		In our view, the reformulated sentence

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
- EFPIA			[…] therefore, do not provide a higher certainty of results compared with single-arm trials Suggested rewording […] therefore, ~~do not~~ provide a ~~higher~~ similar certainty of results ~~compared with single-arm trials~~ **as other uncontrolled trials mentioned previously**. [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Suggested rewording adds necessary context and clarity to the statement.		does not constitute an improvement.
Mihai Rotaru - EFPIA	17	563-566	The information provided within this paragraph should match the text included within the previous Basket trial paragraph as the reasons and context for knowing the information on the test is the same in both circumstances.		We will consider the most appropriate wording for the next version of the draft.
Mihai Rotaru - EFPIA	17	572-573	Current wording "Therefore, the term excludes data collected explicitly for research purposes." Suggested rewording ~~"Therefore, the term excludes data collected explicitly for research~~		We will consider the most appropriate wording for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			~~purposes.~~" [note: ~~strikethrough~~ denotes proposed deletion] Rationale The text indicates that "data collected explicitly for research purposes" are excluded from the definition of RWD. While the principle is endorsed, a registry contains data from clinical practice and, as stated in Line 631-632 (Section 5.3), "it could be advocated that some registries are organised systems that are explicitly devoted to research purposes." As such it is suggested that this sentence be deleted. If the sentence remains, we propose **"**Therefore, the term excludes data collected explicitly for **experimental intervention** research purposes" [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Clarification is needed as to whether non-interventional observational cohort studies fall under the RWD. All RWD meets the criteria for observational data as defined: "In observational studies, there is no "forced" change in routine care and neither is the usual decision for intervention affected by an observational study." (line 259).		
Mihai Rotaru - EFPIA	17	573-575	Current wording "In relation to the concept of RWD, real-world evidence (RWE) is a term defining clinical evidence of a health technology derived from the analysis of RWD for a given research question. RWD can be used to generate RWE for different purposes: generating hypotheses ..."		We will include this modification.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Suggested rewording "In relation to the concept of RWD, real-world evidence (RWE) is a term defining clinical evidence of a health technology derived from the analysis of RWD for a given research question. RWD can be used to generate RWE for different purposes~~:~~ **for example** generating hypotheses ..." [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The text should allow for other examples of the use of RWE rather than restrict to specific uses only.		
Marko Ocokoljic (SIOPE)	17	575, 576	*"RWD can be used to generate RWE for different purposes: generating hypotheses for testing in future RCTs* ***or generating external control data to single arm trial****, assessing trial feasibility..."*		We will include a "for example", but the text is not mean to be an exhaustive description.
Tanja Podkonjak, Takeda	17	Line 568-580	"Real-world data (RWD) is an umbrella term encompassing the use of various types of data that share the common property they have been generated in the context of routine healthcare [e.g., electronic health records, medical claims and billing data, administrative healthcare databases, patient-generated data (including in-home-use settings) and data produced from various sources (such as electronic devices) that can inform on health status] (52–54). Therefore, the term excludes data collected explicitly for **experimental intervention** research purposes. In relation to the concept of RWD, real-world evidence (RWE) is a term defining clinical evidence of a health technology derived from the analysis of RWD for a given research question." [note: **bold** denotes proposed inclusion]		Duplicated comment.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Rationale: Clarification is needed as to whether non-interventional observational cohort studies falls under the RWD. All RWD meets the criteria for observational data as defined: 'In observational studies, there is no "forced" change in routine care and neither is the usual decision for intervention affected by an observational study.' (line 259) Data used can be drawn from pre-existing longitudinal cohort studies by other investigators collected only for research purposes similar to those of a registry.		
Hervé Tchala Vignon, Zomahoun/ INESSS	17	567/5.2	The use of Real-world data (RWD) for the trials could facilitate the identification of trial participants and insure a good statistical precision of the effects to be estimated. However, a specific attention must be paid to the phenomen of potential statistical overpower. Therefore, it would be interesting to have a point on that in the box on practical guideline. **Relevant references** *Case LD, Ambrosius WT. Power and sample size. Methods Mol Biol. 2007;404:377-408. doi: 10.1007/978-1-59745-530-5_19. PMID: 18450060.* *Gamad N, Shafiq N, Malhotra S. Meta-analysis of cardiovascular superiority trials published in the New England Journal of Medicine to elucidate the concept of superiority margin. Postgrad Med J. 2021 Apr;97(1146):227-233. doi: 10.1136/postgradmedj-2019-136569. Epub 2020 Mar 10. PMID: 32156742.*		Statistical overpowering of a study is mainly an ethical issue and therefore out of the scope of the present guideline.
Hervé Tchala Vignon, Zomahoun/ INESSS	17	567/5.2	The recommendations on good practices for real-world data studies of treatment and/or comparative effectiveness can be helpful to reinforcing the content of this section and the box associated. Please, see more details in the following reference (Figure 1):		We will consider adding this reference.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Berger ML, Sox H, Willke RJ, Brixner DL, Eichler HG, Goettsch W, Madigan D, Makady A, Schneeweiss S, Tarricone R, Wang SV, Watkins J, Daniel Mullins C. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiol Drug Saf. 2017 Sep;26(9):1033-1039. doi: 10.1002/pds.4297.		
EFPSI	17	559	Current wording: *"....therefore, do not provide a higher certainty of results compared with single-arm trials"* Proposed wording: ....therefore, provide a similar certainty of results as other uncontrolled trials mentioned previously.		Already addressed issue.
MTE	17	581-583	The use of RWD taken in this guideline is very narrow and seems to only reference 'within study' uses. Routine health data not in association or part of a trial is required in HTA to describe current epidemiological estimates and trends, prescribing patterns or usual care pathways etc. RWD can also be used for both treatment and control arms in study designs not mentioned in this guidance e.g., Propensity score matched trials etc.		As described by the HTAR, the main purpose of HTA is relative effectiveness assessment which is why we have focused the paragraph on these types of use. The guideline is not meant to be a comprehensive description of all uses of RWD.
Natacha Bolanos, Lymphoma Coalition	17	567	While there is great promise in using RWD and RWE, there are some major challenges that must be overcome: potential bias of data sources, incomplete data sets and lack of harmonization of data between RWE data sources, access to such data, and lack of standards in assessing the value of RWE data.		We agree.
Ermisch – GKV-SV	17	Sec. 5.2	It is important to note that RWD is not a design but rather a data source for studies. The simple fact that RWD is generated in the context of routine healthcare does not provide any genuine benefit compared to other data generated in routine practice. Neither does the often excessive		A dedicated section on RWD was required through the project plan.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			amount of data that often results in RWD being described as big data. Validity and usefulness of RWD for research question needs to investigated and proven, e.g. with regards to biases and confounders. Therefore, we suggest to remove section 5.2 and include the necessary considerations regarding in the discussion on observational designs (e.g. in the section 4.4 on cohort studies, where routine healthcare data is already mentioned in the context of cohort studies).		
Karen Facey	17	Line 581	This sentence about RWD repeats poorly what's been stated in the previous paragraph and could be deleted.	X	We will consider if this sentence needs deletion.
Karen Facey	17	Line 582-591	These sentences are more about the different forms of RWD within a trial and it's not clear they are relevant in this guideline.		We think they are relevant as the primary purpose of HTA is relative effectiveness assessment and these kinds of designs can achieve such goal and should be mentioned.
Karen Facey	17	Lines 591-596	These sentences indicate that RWD can be used for indirect comparisons, but says this is outside the scope of this guideline. As above, consideration of how RWE studies may be used to demonstrate treatment effectiveness is needed. Reference could be made to a range of papers from learned organisations such as FDA (RWE program), HAS or the NICE RWE Framework that provide the kind of guidance I would have expected in this section.		This guideline is not meant to detail how RWE needs to be implemented but is meant to provide general considerations for assessors and co-assessors. The main point if the fact RWE are not a type of designs in themselves.
Roche	17	544 - Practical Guideline / 5.1.2	Current wording: *"If an interaction test for homogeneity of effect was performed, its method and result."* This is an incomplete statement. Proposal: Please modify this language so that it is clear what is intended to be asked.	X	Duplicated comment.
Roche	17	544 - Practical	Current wording:		Already addressed issue.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
		Guideline / 5.1.2	*"Results of effectiveness within each 'cohort' of patients with appropriate statistical estimates."* This statement needs to be accompanied by an understanding that the estimation of effectiveness within each cohort of a basket trial can be challenging in practice, for example due to very small cohort sizes. Such cohort specific estimates are typically not part of the primary study objective and should, therefore, be interpreted with utmost caution. The Practical Guideline D4.6 should call out such caveats and seek to establish best practices based on the scientific literature. The literature on subgroup analysis [1] and on Bayesian Hierarchical modeling [2-4] seem particularly important. Please modify the statement to include these references and caveats. [1] Sun et al., "How to Use a Subgroup Analysis: Users' Guide to the Medical Literature," JAMA 311, no. 4 (January 22, 2014): 405–11, https://doi.org/10.1001/jama.2013.285063. [2] Berry et al., "Bayesian Hierarchical Modeling of Patient Subpopulations: Efficient Designs of Phase II Oncology Clinical Trials," Clinical Trials 10, no. 5 (October 1, 2013): 720–34, https://doi.org/10.1177/1740774513497539. [3] Chu and Yuan, "A Bayesian Basket Trial Design Using a Calibrated Bayesian Hierarchical Model," Clinical Trials 15, no. 2 (April 1, 2018): 149–58, https://doi.org/10.1177/1740774518755122. [4] Hannigan et al., "Improving the Estimation of Subgroup Effects for Clinical Trial Participants with Multimorbidity by Incorporating Drug Class-Level Information in Bayesian Hierarchical Models: A Simulation Study,"		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Medical Decision Making, August 18, 2021, 0272989X211029556, https://doi.org/10.1177/027		
Roche	17	559 / 5.1.3	Current wording: *"Therefore, do not provide a higher certainty of results compared with single-arm trials"* Suggested wording: *"therefore, ~~do not~~ provide a ~~higher~~ similar certainty of results ~~compared with single-arm trials~~ as other uncontrolled trials mentioned previously.* "	X	Duplicated comment.
Roche	17	559-563 / 5.1.3	Please add a link to the subgroup analysis sections in the related Guideline D4.5, currently under public consultation, as it is relevant to this statement.	X	We are not sure the statement in this guideline which is about the issue of a common control is really related to the reporting of the results of subgroup analyses.
Roche	17	563-566 / 5.1.3	The information provided within this paragraph should match the text included within the previous Basket trial paragraph as the reasons and context for knowing the information on the test is the same in both circumstances.		Duplicated comment.
Roche	17	574 / 5.1.3	Current wording: *"Real-world evidence is a term defining clinical evidence of a health technology...."* RWE does not necessarily need to be evidence derived from a health technology. It could for example also be used when describing the epidemiology of a disease. Suggested wording: *"Real-world evidence is a term defining clinical evidence of a* disease or *health technology...."*	X	We will consider if this statement needs rewording for the next version of the draft.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Roche	17	575-580 / 5.2	Add here "for example" or some other indiciation to suggest that this is not a full list of what RWE can be used for (too many uses to list succinctly and no need to be exhaustive in this document)	X	Already addressed issue.
Silke Walleser Autiero Medtronic	17	581-583	The use of RWD taken in this guideline is very narrow and seems to only reference 'within study' uses. Routine health data not in association or part of a trial is required in HTA to describe current epidemiological estimates and trends, prescribing patterns or usual care pathways etc. RWD can also be used for both treatment and control arms in study designs not mentioned in this guidance e.g., Propensity score matched trials etc.		Duplicated comment.
Edwards Lifesciences	17-18	567-623/ Section 5.2 Real-world data and real-world evidence	In order to ensure the **predictability** of the process and its outcomes for the innovative technology developers, and on the other hand make sure that the JCA accelerates the access for the patients and is linked and secures **reimbursement/coverage decisions, more precise information on the RWD sources to be used should be provided.** Separately, many of the data sources are not accessible to the HTD, what are the measures foreseen by the guideline, to **facilitate and help the HTD access to the relevant data sources** in order to meet this draft guidance's expectations?		This guideline is not intended to facilitate or help the HTD access to data nor to list RWD sources to be used but is intended for assessor and co-assessor.
Matias Olsen, EUCOPE	17-19	567-624	Since it is not possible to exhaustively list every aspect of bias or study design choice in this guidance document as it relates to RWD, and there are many nuances to data quality and validity of RWD that need to be thought through, we would encourage a deeper dive into the nuances of RWD+RWE, referring to the existing literature when providing detailed guidance, including the GPP (Good Pharmacoepidemiology practices).		The primary purpose of HTA according to the HTAR is relative effectiveness assessment. We think we have described in general that RWD+RWE are not a type of designs in themselves and therefore the validity of such studies should mostly by assessed using the principles described in the previous sections.
Denis	18	Lines 586-	Pragmatic trials are not limited to trials performed under an existing RWD		Our statement does not imply pragmatic

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Lacombe EORTC		587 and general	collection. TWICs being referred here or cohort based clinical trials are only a form of pragmatic trials. Very pragmatic clinical trials can be created de-novo for specific questions and should therefore be as well integral part of chapter 4.1		trials are limited to existing RWD.
Matias Olsen, EUCOPE	18	586-587	"Pragmatic trials" are trials in which an intervention is allocated for the purpose of the trial and usually randomly allocated. It can't be part of RWD. From randomized double-blind studies with very tight inclusion/non-inclusion criteria to the pragmatic trial there is a continuum. The border between randomized double-blind trials and pragmatic trials is large. The misconception of a pragmatic trial and the lack of consideration of pragmatic trials in this guideline is an issue and should be addressed. Pragmatic trials are considered as the best way to increase external validity while minimizing the risk of biases as treatment is allocated randomly to minimize biases and maintain some element of internal validity. The proposed guidelines remain very generic and do not bring any additional information and in fact revert to the study design and RoB as relevant to the study design. Clearer rules for accepting, analysing and interpreting RWE studies should be created at the EU level and not be left for interpretation at the MS level.		What we meant here is the fact that a source of RWD can be used to collect information for the conduct of a RCT, and this way of collecting data are "sometimes considered as part of pragmatic trials" (which does not imply they encompass all kinds of pragmatic trials). We agree there is a continuum between RCTs with tight inclusion/exclusions and pragmatic trial. The assessment of acceptability of evidence is already addressed earlier within the guideline. Moreover, we reaffirm that from the perspective of HTA, RWE are not a type of design in themselves.
Mihai Rotaru - EFPIA	18	615	<u>Current wording</u> Finally, particular attention to the assessment of endpoints is required because there is a risk of unblinded and/or decentralised adjudication of endpoint processes (60). <u>Suggested rewording</u>		We will consider a more appropriate wording for the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Finally, particular attention to the assessment of endpoints and how those endpoints were adjucated on **(e.g., investigator vs central review, differences between sites)** (60) **as well as timing of assessments**" [note: **bold** denotes proposed inclusion] Rationale: It is not clear from the current wording what this means. Proposed wording gives more clarity. The timing of assessments may also be important to consider if they differ between sites or across studies.		
Denis Lacombe EORTC	18	Lines 586-587 and general	Pragmatic trials are not limited to trials performed under an existing RWD collection. TWICs being referred here or cohort based clinical trials are only a form of pragmatic trials. Very pragmatic clinical trials can be created de-novo for specific questions and should therefore be as well integral part of chapter 4.1		Duplicated comment.
Marko Ocokoljic (SIOPE)	18	593, 594, 595, 596	*"Although this is out of the scope of this Guideline, they can be used as sources of data for indirect comparisons (see the EUnetHTA 21 Methodological Guideline Direct and Indirect Comparisons), or as additional historical data borrowing for enriching data of a control group in an already existing clinical trial (e.g., when the trial* ***concerns rare diseases or paediatrics****)."*		The assessment of the validity of clinical study is independent of the medical context.
Marko Ocokoljic (SIOPE)	18	597, 598	*"The use of RWD in generating evidence can be useful in multiple ways. First, their use can enhance the recruitment of patients in clinical trials, especially for rare diseases* ***and paediatrics****."*		We do not think specifying medical context will benefit the document.
MTE	18	584-596	Again, the terms used here, 'pragmatic trial' and 'contactless trial' trial need to be included in the classification of trial designs Fig 3.1 or		Already addressed issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			another classification figure to improve understanding and clarity.		
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ingo Hantke SKC Beratungs-gesellschaft mbH	18	597	Comment: RWD and RWE are of substantial value for HTA assessments, e.g., in terms of providing evidence on treatment effects and epidemiological data. Their specific acknowledgement is seen highly positive.		Thank you.
Ermisch – GKV-SV	18	623 (box)	Please add the clause: "; its value as a primary source for performing relative assessment may be limited" To the seconds sentence to reflect the potential weaknesses of RWD and RWE described earlier.		This statement seems too vague and unnecessary.
Karen Facey	18	Line 597	This paragraph about use of RWD to improve clinical trials is more about improving efficiency of clinical trials, not assessment of validity of a clinical trial. It seems out of place and should be deleted. And as a result the 4th point in the practical guideline box could be deleted.		We think this is one of the general strengths of the use of RWD and it can be cited.
Karen Facey	18	Below line 623	This box seems to be written differently to that in other sections. Are the first 4 points "requirement for JCA reporting", if so the last 2 points should be moved up to be above the "specific points of attentions".		We will check the consistency of the box for the next version of the draft.
Silke Walleser Autiero Medtronic	18	584-596	Again, the terms used here, 'pragmatic trial' and 'contactless trial' trial need to be included in the classification of trial designs Fig 3.1 or another classification figure to improve understanding and clarity.		Duplicated comment.
GSK	18	615	What's the definition of 'decentralised adjudication'? Should the difference of assessment timelines for certain endpoints be addressed too? (e.g. progression free survival in oncology studies).		We will clarify this for the next version of the draft.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Denis Lacombe EORTC	19	Chapter 5.3 General	EUnetHTA should take this opportunity to clarify what a registry is/is not. It is not clear what could the difference be between any prospective collection of RWD and a registry. It could be of value to indicate that a registry should be exhaustive of all patients bearing the intended disease in a given area to avoid any bias. To some extent, it is what seems to be defined in line 650 referring to national or regional level population based registry.		As the guideline's focus is on clinical studies, there is no benefit in formulating a definition what a registry is.
Mihai Rotaru - EFPIA	19	628-630/5.3	Current wording "Data collected within the registry can then be used to conduct registry-based studies. Given that they are often a collection of observational data from routine healthcare practices, data from registries can be considered as RWD (54), but it could be advocated that some registries are organised systems that are explicitly devoted to research purposes. Nevertheless, registry data can be used in the same way (e.g., as the sole source of data or as a primary source of data) and for as many purposes as RWD" Recommendation Some clarity should be added to describe when registries are considered as RWD. For example, the use of existing registry data for other research questions not included in the original protocol should be considered RWD. The use of registry data, when considered RWD, should also be included in the Section 5.2 (Real-world data and real-world evidence)		There is no need to better differentiate between registries and RWD, as the validity of a clinical study depends mostly on other aspects of research.
Mihai Rotaru - EFPIA	19	642-644/5.3 Registries	Current wording "Therefore, some registry-based studies can have the ability to produce true statistics in a population of interest and not estimates; therefore, external validity of a registry-based study can be better than of a study conducted on a sample of patients only (providing the target population of the corresponding clinical study is the same as the population covered by the registry)"		We agree with this suggestion.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Suggested rewording "Therefore, some registry-based studies can have the ability to produce ~~true statistics in a population of interest and not estimates; therefore, external validity of a registry-based study can be better than of a study conducted on a sample of patients only (providing the target population of the corresponding clinical study is the same as the population covered by the registry)~~ **the true parameter value in the population of interest rather than an estimate (provided the population covered by the registry is the same as the target population).”** [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The phrase "true statistics" is an oxymoron. A statistic is calculated from a sample to estimate a population-level parameter of interest. If a population is sampled exhaustively or nearly exhaustively, the resulting quantity that is calculated would not be a statistic – it would be the true value of the parameter. Similarly, if a population is sampled exhaustively or nearly so, then the concept of external validity is not relevant. External validity refers to how well the results from a sample are expected to generalize to the whole population. But if the sample is the whole population, obviously the results generalize (trivially so).		
Denis Lacombe EORTC	19	Chapter 5.3 General	EUnetHTA should take this opportunity to clarify what a registry is/is not. It is not clear what could the difference be between any prospective collection of RWD and a registry. It could be of value to indicate that a registry should be exhaustive of all patients bearing the intended disease in a given area to avoid any bias. To some extent, it is what seems to be defined in line 650 referring to national or regional level population based		Duplicated comment.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			registry.		
Marko Ocokoljic (SIOPE)	19	607	*"…conduct of a given clinical study (58).* ***In addition, the situation of single arm trials and the need for external control through RWD is likely to be more frequent in paediatrics and rare or ultra-rare diseases. Hence, building academic capability to generate robust RWD that would comply with regulatory and HTA requirements could be a potential way to facilitate collection of robust RWD.*** *A related issue can be the use of certain variables from databases…"*		Building academic capability is not the topic of the present guideline.
Tanja Podkonjak, Takeda	19	Line 637	Current text: A particular point that can sometimes apply is the fact some registries aim toward exhaustivity. Proposed text: A particular point that can sometimes apply is the fact some registries aim toward applicability. Rationale: Exhaustivity is referred to 'conditional exchangeability', i.e. all relevant confounders and effect modifiers must be known and adequately measured within the trial (line 368-375, 622-623). It belongs to 'internal validity'. However, this paragraph discussed external validity 'therefore, external validity of a registry-based study can be better than of a study conducted on a sample of patients only' (line 642-643),		We will consider a more appropriate wording for the next version of the draft.
Tanja Podkonjak, Takeda	19	628-630/5.3	Current wording: "Data collected within the registry can then be used to conduct registry-based studies. Given that they are often a collection of observational data from routine healthcare practices, data from registries can be considered		There is no need to better differentiate between registries and RWD, as the validity of a clinical study depends mostly on other aspects of research.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			as RWD (54), but it could be advocated that some registries are organised systems that are explicitly devoted to research purposes. Nevertheless, registry data can be used in the same way (e.g., as the sole source of data or as a primary source of data) and for as many purposes as RWD" Takeda requests clarity be added to describe when registries are considered as RWD. For example: The use of existing registry data for other research questions not included in the original protocol should be considered RWD. The use of registry data, when it is considered RWD, should also be included in the section 5.2 Real-world data and real-world evidence.		
Edwards Lifesciences	19	624-658/ Section 5.3 Registries	In order to ensure the **predictability** of the process and its outcomes for the innovative technology developers, and on the other hand make sure that the JCA accelerates the access for the patients and is linked and secures **reimbursement/coverage decisions, more precise information on the registry data sources to be used should be provided.** Separately, many of the data sources are not accessible to the HTD, what are the measures foreseen by the guideline, to **facilitate and help the HTD access to the relevant data sources** in order to meet this draft guidance's expectations?		Data sources accessibility for HTD is of interest but out of scope of this guideline
Natacha Bolanos, Lymphoma Coalition	19	624	The biggest limitation is the lack of established methodology and protocol for designing registries. Poor conceptualization and incorrect design can lead to misinterpretation of data or ineffective data. If the definitions for data collection are not well defined at the time of building the registry, it may lead to highly suggestive, but incorrect data. The data collected by registries is not validated as compared to traditional RCTs. A lot of data is sourced from other administrative databases and records. Many a times there are gaps in the databases resulting in difficulties in processing the		How to run registries is outside the scope of the present guideline. Some of the cited articles may contain information on this.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			data. Due to lack of stringent validation techniques doubts may be raised regarding its value. Registries are not only a tool for data collection, they should be for action.		
Karen Facey	19	Below line 658	As above this box seems to have a different form of presentation of points. Are all the first elements "requirements for JCA reporting" too?		
GSK	19	630	"observational data from routine healthcare practices, data from registries can be considered as RWD". Is there a preference for specific databases or vendors by disease area?		This comment addresses details that go beyond the scope of the guideline.
GSK	4	75	Please note that the choice of abbreviation for Health Technology Assessment Regulation is not consistent between this document ("HTAR") and guidance document 5.1 ("HTA R"). Suggest a consistent approach.	X	The consistency of wording will be checked at a later stage.
Mihai Rotaru - EFPIA	5	81-85	Current wording In 2020, the European Network of Health Technology Assessment (EUnetHTA) Executive Board concluded that GRADE (1) (or any other system for rating the overall quality of evidence and developing healthcare recommendations) can only partially be applied within EUnetHTA because overall conclusions or recommendations might interfere with the independent contextualisation and decision-making at the national level (2). Suggested additional text In 2020, the European Network of Health Technology Assessment (EUnetHTA) Executive Board concluded that GRADE (1) (or any other system for rating the overall quality of evidence and developing healthcare recommendations) can only partially be applied within EUnetHTA because overall conclusions or recommendations might interfere with the independent contextualisation and decision-making at the national level (2). **In addition, it is important that no critical or negative wording is included in the JCA report with respect to the assessment of the validity of clinical studies. The "Subdeliverable - 'negative list'**		This comment addresses linguistic details that go beyond the scope of the guideline. Nevertheless, prejudging vocabulary should have no place in a JCA.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Task Group for Common Phrases and GRADE[1]" provides examples which clarifies the language that should be avoided in the JCA report". [note: **bold** text denotes proposed inclusion] Rationale EFPIA believes the guideline should further stress the importance that appropriate wording and language is used throughout the JCA report, particularly when assessing the validity of the relevant clinical studies. Value judgements should not form part of the JCA report, and this subdeliverable helps "to avoid the use of sentences/words in an assessment report which may unintentionally imply or predetermine reimbursement decisions in some jurisdictions (creation of a negative list of phrases)" i.e., value judgements. References 1. Subdeliverable - 'negative list' Task Group for Common Phrases and GRADE, available at: https://www.eunethta.eu/wp-content/uploads/2021/05/EUnetHTA-Negative-list-of-phrases.pdf		
Mihai Rotaru - EFPIA	5	89-90	Current wording "This Practical Guideline is dedicated to the definition, classification, and certainty of results..." Suggested rewording "This Practical Guideline is dedicated to the definition, classification, and **assessment of the** certainty of results..." [note: **bold** denotes proposed inclusion].		We thank the commentator for bringing this slipping of words to our attention. The text will be changed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Rationale Additional wording suggested to clarify statement.		
Mihai Rotaru - EFPIA	5	90	Current wording: "... statistical analysis of what is considered one data set (one sample of patients)." Suggested rewording: "… statistical analysis of ~~what is considered one data set (one sample of patients)~~ **data considered as originating from or part of a single study (i.e., one sample of patients)**." [note: **bold** and strikethrough denotes proposed inclusion and deletion, respectively]. Rationale: This phrasing is not precise. Data set can be construed as the compilation of data and could originate from multiple sources. As per Oxford Dictionary: a data set is "a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer."		The proposed wording appears useful and will be incorporated into the guideline.
Mihai Rotaru - EFPIA	5	99	Current wording "The way in which the validity of clinical studies will be assessed for drawing conclusions…" Suggested rewording "~~The way in which the validity of clinical studies will be assessed~~ **The**		The proposed wording does not lead to a complete sentence. Although validity is not binary, there is no need to rephrase this sentence.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			degree to which clinical study results are valid will be assessed and interpreted for drawing conclusions ..." [note: **bold** and strikethrough denotes proposed inclusion and deletion, respectively]. <u>Rationale</u> Validity is not binary but, rather, continuous, since we are sampling and estimating. The suggested change will contribute to add consistency throughout the document.		
	5	90-93	**Current wording**: "Studies that consist in evidence synthesis by pooling the results of multiple already-analysed data sets from multiple samples of patients [e.g., pairwise meta-analysis, indirect comparison, or interventional studies using external control (including historical control)] are not included in this Guideline." The text suggests that interventional studies that borrows from external controls in a pre-defined manner are invariably considered evidence synthesis. Since such innovative trial design is increasingly being considered by regulators, it should at least be acknowledged, along the lines of ICH E10, that "[...] some approaches to design and conduct of externally controlled trials could lead them to be more persuasive and potentially less biased."		We reaffirm that such studies are out of scope of this guideline and won't change the wording.
Karen Facey	5	Line 79	A better first sentence is needed in the problem statement to set the scene for the rationale of the guideline in terms of why HTAs critically assesses clinical studies and to limit the considerations in this guideline to evaluation of relative effectiveness (as in some sections it strays off into wider considerations that don't seem so relevant), e.g something like	X	In our view, it is easier to mention "clinical study results" without differentiating between "key clinical studies in a technology development programme and other sources (such as published literature)". The validity of a clinical study

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Assessment of the relative effectiveness of a health technology is one of the key purposes of an HTA. This requires submissions of evidence from the key clinical studies in a technology development programme and other sources (such as published literature), which must be critically appraised according to the PICOs specified for the HTA.		does essentially not depend on the type of study sponsor or the format of reporting study results. No change required.
Roche	5	85-86 / 1.1	Current wording *"not only to guide the development of Joint Clinical Assessments (JCAs) at the European level"* Suggested rewording: *"not only to guide the development of Joint Clinical Assessments (JCAs) and* Joint Scientific Consultation (JSC) *at the European level"* We propose adding in this statement to highlight the importance in both the JCA and JSC in the context of agreement in the evidence needs and decision making. These parameters are also relevant in the early scientific advice discussions.		The primary aim of the present guideline is the assessment of completed clinical studies. Although this may lead also to recommendations how to design future studies, we feel that the JSC process is better not mentioned here.
Roche	5	90-95 / 1.2	This guideline should cover the potential use or utility of external or synthetic controls as a valid way to provide comparative evidence by Real World Data for non-randomized, non-controlled trials such as single-arm studies		This issue is dealt with in the guideline on indirect comparisons.
Jasmine Toomey PHMR	5	Line 82, section 1.1	Clarity required for "GRADE" as it is not clear what this methodology is.		Unfortunately, "GRADE" was missing in the list of abbreviations. This will be corrected.
Matias Olsen, EUCOPE	6	125-127	The assessment of the additional benefit of a drug has been segmented over a series of guidelines, although the topics are interrelated. This lack		As this sentence is mainly a repetition of the HTA Regulation and its description of

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			of consolidation hinders an appreciation of the relative effectiveness, which is by definition a global assessment of one intervention over one or more reference intervention(s).		'the relative effects of the health technology', we see no need to expand this definition.
Matias Olsen, EUCOPE	6	128-134	The three dimensions/concepts for certainty of effectiveness are defined as internal validity, applicability and statistical precision. They are considered as independent and must be assessed independently, without influencing each other. In theory, studies (or a series of studies) can have a good internal and external validity. For a single study this might mean a randomised study, but with limited inc/exc criteria and assessments coupled with a treatment policy type estimand. However, in real life, there is a fine balance and trade-off between the internal validity and the applicability. Applicability should be restricted to external validity and potentially generalisability, which are overlapping concepts. The highest the internal validity, the lowest the external validity and the lowest the generalisability. Increasing the internal validity can reduce the external validity and vice versa. This is in part why all regulators aim to maximize the internal validity of clinical studies. To achieve a high internal validity, the trial should control all potential confounders that may affect the outcome of the trial beyond the intervention(s) tested. The precis-2 tool for designing trials that are fit for purpose (BMJ 2015 May 8;350H2147 by Loudon K and al.) shows well that one needs a very low score to answer the question "does it work?", whereas one needs a very high score over 9 independent dimensions to answer the question "does it work under real conditions?". Obviously, one cannot have a high and a low score at the same time. For example, in clinical practice, tricyclics are used by GPs at suboptimal		We disagree with the general assumption that internal and external validity are conflicting aspects of a clinical trial, which cannot be achieved both at the same time. The PRECIS-2 tool is not useful in EU-HTA, because applicability has to be assessed in each member state separately. The example of a very large effect illustrates nicely that the overall certainty of a positive effect can be increased. However, a very large effect size does not increase the internal validity of a study. Thus, the three dimensions are correctly described to be (largely) independent from each other. The comment on using "the full set of evidence" is correct, but the present guideline's focus is only on individual studies.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			dose (below SPCs recommendations) because of the high risk of ventricular arrhythmia, whereas SSRIs are used at full therapeutic dose and their dose is increased in case of none response, as they are perceived as safe. GPs prescribe about 75% of antidepressants. Comparing tricyclics to SSRIs using tricyclics at sub-therapeutic dosage, to reflect the clinical practice, in a study increases external validity while reducing internal validity. Using tricyclic at the optimal dose recommended by SPCs increases external validity while reducing internal validity. The same applies for piroxicam, which is prescribed in patients contraindicated to non-steroidal anti-inflammatory drugs, because of its low digestive haemorrhage risk. Therefore, the risk of haemorrhage in clinical practice is higher with piroxicam because of misuse. A high internal validity study will show a lower rate of haemorrhage, while a high extern validity study will show the opposite. So the three concepts cannot be stated as independent, but the three concepts should be considered first separately and then jointly. The interdependence between internal validity and statistical precision is illustrated by the following example. Assuming a very large clinical effect size, with a perfect statistical precision, and some limitation in the internal validity, it is likely that the very large effect size will overcome the limitation in internal validity. We can illustrate with an extreme example. Assume all patients in the reference trial died while none died in the new active intervention, with a p-value of 0.001 and a power of 90%. However, the randomisation was not centralized but done at centre level and the study not blinded. Such study carries some internal validity limitation and a high statistical precision. The high statistical precision overcome the internal validity limitation and the overall clinical study certainty will be considered as high.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			The appraisal should not be done by considering individual studies in isolation, but by considering the full set of evidence that is submitted. For example, it could well be that limitation of one study is addressed by another study. So, looking at individual studies independently without consolidation of the full set of evidence submitted may lead to an inappropriate assessment.		
Matias Olsen, EUCOPE	6	137-140	It may not be ethical or appropriate to eliminate evidence when one of the three concepts does not reach a minimum level. Increasing the expectation of certainty of results beyond EMA accepted thresholds would only delay patients access while high unmet needs exist. It is important to recognise that interventions in areas of high unmet need, e.g. therapies for rare diseases or paediatric populations and, ATMPs are approved with single-arm trials. Under this new rule, single-arm trials would be systematically eliminated while historical/external arm comparison may be available and bring highly relevant information to support HTA decisions. As the three concepts related to certainty are inter-related and dependent, no threshold should be used to eliminate the evidence.		EU-HTA is a process that is independent from EMA assessments and EMA methods. Patient access to medicines will not be decided within the EU-HTA process but in each member state. The issue of using single-arm studies for indirect comparisons is being addressed in another EUnetHTA guideline.
Matias Olsen, EUCOPE	6	140-143	The current guideline specifies that the medical context should not be considered when assessing the relative effectiveness (for example rarity or impossibility of blinding). This clearly implies that products based on single-arm studies will be concluded as not assessable, or assessments will conclude that there is no evidence of any relative effectiveness benefit over current available SoC. It is important to note that a vast majority of rare disease interventions and ATMPs are approved with single-arm trials and most of them recommended by HTA agencies. As per the current guidance, although it		The issue of using single-arm studies for indirect comparisons is being addressed in another EUnetHTA guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			may be inappropriate or not feasible to undertake a randomised control trial, single-arm trials would be systematically eliminated. This will lead to relevant evidence not being considered in the evaluation process. For example, a recent gene therapy targeting a rare fatal condition in babies was approved with a single-arm trial and a very small sample size and was recognized as having a very high relative additional benefit by almost all HTA bodies. Under the proposed guidelines, such a therapy would not be reviewed because of the lack of consideration of the medical context and the low level of internal validity. Ignoring the medical context makes the assessment purely theoretical and disconnected from the reality. The specificity or rarity of a medical condition may make it impossible to achieve a predefined level of certainty. Historical (external) controls may be available and bring highly informative information to support decision. EMA acknowledges this medical context through approval under exceptional circumstances, early access, etc. The EU HTA methodology needs to be consistent with the current regulatory methodologies. Remove: "~~Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence.~~"		

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Matias Olsen, EUCOPE	6	144-146	The statement is correct, but the conclusion is incorrect. If international standards of evidence-based medicine set the highest grade when internal validity is high, they also recommend that all evidence is considered to draw a conclusion and a recommendation. Therefore, according to the international standards of evidence-based medicine, all evidence must be considered. As already mentioned above, under this new rule, single-arm trials would be systematically eliminated, even if historical (external) arm comparisons may be available and may bring highly informative information to support decision. It is important to note that a vast majority of rare disease interventions or paediatric populations and ATMPs target conditions with high unmet needs are approved with single-arm trials. Most of them are recommended by HTA agencies, often with a high recognition of relative additional benefit over the reference intervention. Therefore, a more pragmatic approach is required. Achieving a high internal validity can negatively impact external validity and therefore the trade-off between the two concepts must be considered. In the same way as high statistical precision and large effect size may mitigate concerns about internal validity, the medical context may make it impossible to achieve a predefined level of certainty. The European Commission routinely approves certain products when it is obvious that they deliver a high added value in a context of high unmet need, despite limitations in trial evidence. Therefore, a pragmatic approach needs to be applied to areas of high unmet need, e.g. OMPs, ATMPs, and for paediatric populations, where it is not always feasible or ethical to undertake a RCT.		The issue of using single-arm studies for indirect comparisons is being addressed in another EUnetHTA guideline. While using lower levels of evidence may be helpful in some situations, assessing ALL evidence (including case reports, animal experiments, cell and other bench research) would neither be practical nor useful.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			The analysis of certainty should consider the interdependence of the three concepts as well as the medical context, the feasibility and the magnitude of the effect size. Remove: "Following international standards of evidence-based medicine, the internal validity of a study has a paramount role in determining the overall certainty of the study results ~~(i.e. if study results have a low level of internal validity, the levels of statistical precision and external validity are irrelevant)~~ (5,6)"		
Matias Olsen, EUCOPE	6	147 – 150	The level of quality of a study depends very much on the appropriateness of the design to address the research question. For example, to assess the potential excess mortality associated to an antipsychotic that prolong the QT interval, a retrospective study using a very large database will provide more precision and faster results than a prospective cohort study, that may not be able to provide a relevant precision within a reasonable timeline. So, the appropriateness of the study design should always be assessed in relation to the study objective and not be considered in isolation, with prior opinion on what design is better or worse.		The guideline's focus does not include the question, which study designs have higher or lower practicability.
Mihai Rotaru - EFPIA	6	131-132/2	Current wording "[…] and statistical precision (i.e., the extent to which study results are free from random errors resulting from sampling hazards)" Suggested rewording "and statistical precision (i.e., the extent to which study results are free from random errors resulting from sampling hazards). **Random error decreases with increasing sample size, while systematic error (bias) does not. Systematic error indicates the extent to which a study is free from**		While the suggested sentences appear correct, we see no need to add them. The guideline should remain readable and general information must be left to epidemiology or statistics textbooks.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			bias, that is, how far the expected value of the estimate is from the true value of the parameter being estimated. Random error is a measure of how similar the multiple estimates are to each other (statistical precision), so the uncertainty associated with study is due to random sampling variability" [note: **bold** text denotes proposed inclusion] Rationale The suggested text adds further clarity on the differences between random and systematic errors.		
Mihai Rotaru - EFPIA	6	128-143	**Assessment of certainty of results (internal validity, external validity, and statistical precision).** EFPIA wishes to stress that any description of the certainty of the results of a given study in the context of the JCA, including the reporting of such assessment in the JCA report, must not lead to a value judgment of the relevant technology, nor should it create obstacles for the proper consideration of the submitted evidence at Member State level; furthermore, the assessment of the certainty of the results should not lead to a de-valuation or exclusion of the relevant studies. In addition, it is important that no critical or negative wording is included in the JCA report. The "Subdeliverable - 'negative list' Task Group for Common Phrases and GRADE" is a still relevant source of examples which clarifies the language that should be avoided in the JCA report[1]. This recommendation reflects both the letter and the spirit of the EU HTA Regulation[2], specifically:		The guideline authors are well aware of the EU HTA regulation. If deemed necessary, additional (linguistic) guidance will be developed and implemented in order to avoid value judgment in the JCA report.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Article 9 (14) "*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*" L 458/5 (28) "*The joint clinical assessment report should be factual and should not contain any value judgement, ranking of health outcomes, conclusions on the overall benefit or clinical added value of the assessed health technology*." Article 9 L 458/16 "*Joint clinical assessment shall result in a joint clinical assessment report that shall be accompanied by a summary report. Those reports shall not contain any value judgement or conclusions on the overall clinical added value of the assessed health technology and shall be limited to a description of the scientific analysis*" References 2. EUnetHTA Subdeliverable 'negative list' Task Group for Common Phrases and GRADE, 2019, available at: EUnetHTA-Negative-list-of-phrases.pdf, accessed on 25 June 2022 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).		
Mihai Rotaru - EFPIA	6	137-140	Current wording "Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results."		Assessing ALL evidence on a regular basis would neither be practical nor useful. The current guideline does not per se exclude observational studies.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Suggested rewording "~~Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher certainty results.~~ **In a situation where higher and lower-level evidence is available to address the PICO questions, preference will be given to the higher-level evidence for a JCA given it has a higher level of internal validity, applicability, or statistical precision. "** [note: **bold** and strikethrough denotes proposed inclusion and deletion, respectively]. Rationale EFPIA believes that all evidence should be considered. Scientific and clinical rationale should drive the assessment and inclusion of available and generated evidence, taking into consideration acceptance by regulators, ethics committees and investigators in situations where RCTs (which is identified as the gold standard) are not suitable or feasible. Recital (35) of the Regulation states: "For medicinal products, directly comparative clinical studies which are randomised, blinded and include a control group, the methodology of which conforms to international standards of evidence-based medicine, should be preferentially considered when conducting a joint clinical assessment. **That approach should however not *per se* exclude observational studies, including those based on real world data, when such studies are accessible.**"		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mihai Rotaru - EFPIA	6	147-150	**Hierarchy of evidence** EFPIA wishes to point out that traditional evidence pyramids that were first developed in the 1990s are now considered too simplistic[1], and that most versions represent purely a hierarchy of internal validity (risk of bias), thus neglecting critical additional criteria, such as external validity. A fit-for-purpose, valid, transparent approach, balancing internal and external validity of each study, is what matters most for the purpose of the JCA. It has already been noted that "no system of evidence assessment currently exists that is generally accepted and universally applicable to all systematic reviews"[2]. Furthermore, with specific reference to the classification of non-randomized studies (NRS) with regard to their risk of bias, the study design alone can no longer provide sufficient orientation, even if the basis distinction between comparative and non-comparative studies seems meaningful. Other factors, including the control for potential confounders, will also be important[2]. As a consequence, to give due consideration to the studies submitted for the purpose of the JCA, EFPIA recommends that no conclusive judgement on the quality of the studies is inferred from the hierarchy of evidence alone, but that a detailed and holistic assessment of the totality of the evidence is instead conducted, taking account also of the context, and reflected in the JCA report. In addition, given the Regulation's initial scope to review all new ATMPs (followed by orphan medicinal products) seeking EMA approval, it is absolutely vital that clinical evidence across the 'evidence hierarchy' is not just considered for the purpose of the JCA, but that HTA assessors adopt a flexible approach to assessing internal and external validity and statistical		We agree with the commentator's statement that "no conclusive judgement on the quality of the studies [should be] inferred from the hierarchy of evidence alone". We also support the view that "moving away from the reference to classical hierarchy of evidence" is scientifically justified. The current guideline therefore states that "classification of study design alone ... is insufficient", but has "much practical value". Thus, EFPIA's worries are unfounded. The evidence pyramid is used only as a very basic orientation.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			precision. As many ATMPs will be developed with these considerations in mind, a nuanced approach will be required to ensure EU patients can access new innovative technologies. It is only by transparency in the particulars of a study that we can judge study design validity; this includes objectives, required strength of evidence, and trade-offs made. This is a multi- dimensional matrix / spectrum (not a pyramid). Transparency in rationale, objectives and the study particulars are much more important. Therefore, the field is moving away from the reference to classical hierarchy of evidence[3], overall and at least to a reference of classification in study design, as described in the text (e.g., "case-control studies (level 4), retrospective (or lower-quality) cohort studies (level 3)"). EFPIA believes that a classification based of study design alone is insufficient for a full assessment of internal validity; rather, the totality of evidence should be considered instead. As clearly noted by Rawlins (2008): "The notion that evidence can be reliably or usefully placed in 'hierarchies' is illusory. Rather, decision makers need to exercise judgement about whether (and when) evidence gathered from experimental or observational sources is fit for purpose".[4] <u>References</u> 1. Murad MH, Asi N, Alsawas M, et al, New evidence pyramid, BMJ Evidence-Based Medicine 2016;21:125-127. 2. IQWiG General Methods guide, Version 6.1 of 24 January 2022, accessed 25 June 2022 3. Petticrew M, Roberts H. Evidence, hierarchies, and typologies: horses for courses. J Epidemiol Community Health. 2003		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Jul;57(7):527-9. doi: 10.1136/jech.57.7.527. PMID: 12821702; PMCID: PMC1732497. 4. Rawlins M. De Testimonio: on the evidence for decisions about the use of therapeutic interventions. Clin Med (Lond). 2008 Dec;8(6):579-88. doi: 10.7861/clinmedicine.8-6-579. PMID: 19149278; PMCID: PMC4954394		
Mihai Rotaru - EFPIA	6	131-132/2 General Considerations	Current wording […] "statistical precision (i.e., the extent to which study results are free from random errors resulting from sampling hazards)" Suggested rewording "[…] statistical precision (i.e., ~~the extent to which~~ **the uncertainty associated with** study results ~~are free from random errors resulting from sampling hazards~~ **due to random sampling variability**)" [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The existing statement is an inaccurate layman's translation of statistical precision, as no study can be completely free from random errors. In addition, sampling "hazards" is a misleading choice of words.		We thank the commentator for this useful suggestion. Text is changed.
Mihai Rotaru - EFPIA	6	141-143	**General considerations, PICO question** Current wording Furthermore, the certainty of the results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding		The current guideline is about how to assess the validity of a clinical study, not about when to use a certain study design for HTA purposes.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence. Suggested rewording ~~Furthermore, the certainty of the results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence.~~ **The rareness of a disease and the impossibility of blinding are independent considerations from the certainty of results. Furthermore, in conditions with small and very small populations, less conventional and/or less commonly seen methodological approaches may be acceptable if they help to improve the interpretability of the study results. This reflects the need for efficient trial designs for small populations, fostering continued development of new treatments for rare diseases, particularly when traditional randomised control trial designs are not possible due to limited number of available patients and/or ethical considerations."** [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale The text is unwarranted and unnecessarily provocative. HTDs do not make excuses, they simply report the genuine realities and practicalities of conducting randomized, controlled trials for interventions for rare diseases. Ignoring the context of the disease and the ethical, feasibility and operational challenges of running clinical trials, when undertaking assessments, is in itself methodologically inappropriate; furthermore, this understanding can help guide the most appropriate state of the art methodology to reduce uncertainty and bias when estimating the		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			treatment effect of a technology. The EU Regulation[1] 2021/2282 acknowledges that for some new health technologies (e.g., orphan medicinal products) some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: "*Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products*". References 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		
Mihai Rotaru - EFPIA	6	150-151	Current wording "Classification of study design alone (see Section 3) is insufficient for a full assessment of internal validity (8,9) […]" Suggested rewording "Classification of study design alone (see Section 3) is insufficient for a full assessment of internal validity (8,~~9~~) […]" [note: ~~strikethrough~~ denotes proposed deletion] Rationale The reference (9) for this statement should be reassessed, since reference #9 does not address assessment of internal validity with regard to classification (as neither internal, nor validity appear in the document).		In this reference (citation No. 9, https://pubmed.ncbi.nlm.nih.gov/28215660/), Figure 1 clearly shows the transition from evidence pyramid to a more holistic approach in the assessment of evidence.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Paolo Morgese – ARM	6	141-143	Current wording It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence. ARM recommend deletion of this from the document. Rationale: ARM believe this statement is highly inappropriate when considered in the context of the proposed JCA Phase I/II assessments, which includes ATMPs, which are largely studied in rare disease in areas of high unmet need, leading to smaller and sometimes non-randomised clinical trials. The statement suggests that manufacturers are deliberately choosing to ignore the resultant uncertainty associated with the evidence base, which is not the case. In fact, the EU Regulation 2021/2282 acknowledges that for some new health technologies (e.g., ATMPs) some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: "*Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products*".		We do not see how this sentence could suggest that manufacturers deliberately ignore scientific principle of clinical trials. The use of new methodologies (mainly indirect comparisons) is outside the scope of this guideline.
Tanja Podkonjak, Takeda	6	137-140	Current text: Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results. Suggestion to rephrase or remove this sentence as it implies that only higher-certainty results for a given PICO question should be addressed in		Assessing ALL evidence on a regular basis would be neither practical nor useful (see previous comments and replies).

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the JCA dossier, thereby excluding potentially relevant evidence for the EU level HTA or a national decision maker. For a comprehensive HTA assessment, the totality of evidence should be considered and therefore no evidence should be excluded as suggested above.		
Tanja Podkonjak, Takeda	6	140 - 142	Current text: Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence. Proposed text: Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as reason to not discuss the limitations and ~~an excuse to ignore or to euphemise~~ resulting uncertainties in the clinical evidence. Rationale: The current text is subjective and not constructive. We recommend the guidance reframe the sentence in neutral language to focus on promoting discussion and transparency of the uncertainties of data.		We will delete the "euphemise" statement. It should nevertheless be noted that this statement primarily addresses those who write the JCA, not those who prepare a JCA submission.
Tanja Podkonjak, Takeda	6	144-153	The guideline around the classical hierarchy of evidence, as currently presented, may not be representative of the circumstances and evidence based across all conditions. A flexible approach based on the therapeutical area and circumstances should be considered given that in certain disease areas for e.g. rare disease and oncology, study designs other than randomized clinical trials may be more appropriate to assess the relative effectiveness of the technology under evaluation. Scientific and clinical rationale should drive the assessment of available and generated evidence, taking into consideration acceptance by regulators,		The current guideline is about how to assess the validity of a clinical study, not about when to use a certain study design for HTA purposes.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			ethics committees and investigators in situations where RCTs (which is identified as the gold standard) are not suitable/feasible, not simply a generic hierarchy of evidence.		
EFPSI	6	133-134	Current wording: "These three concepts assess three different dimensions of the certainty of results, which, for example, means that shortcomings in internal validity cannot be remedied by higher statistical precision". This is an important point. We note however that the overall exposition in this guidance heavily promotes internal validity/randomized controlled studies.		The guideline is indeed intended to promote internal validity and RCT evidence.
EFPSI	6	137-140	Current wording: "[...] there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results." The text suggests that evidence submitted by the HTD can be omitted from the JCA at the assessor's discretion - as such, the decision/proposal to exclude evidence from assessment is a value assessment that should be documented and then left to the member countries. We recommend the JCA to consider the totality of evidence submitted by the HTD and let the member countries make a judgment on the certainty of the results. As such, we suggest to clarify this sentence as above.		Assessing ALL evidence on a regular basis would be neither practical nor useful (see previous comments and replies).
EFPSI	6	140-141	Current wording: "Furthermore, the certainty of results is independent of the medical context of the PICO question." Whilst this might be true in a purely mathematical sense, the (medical) context should always be considered in an analysis. Otherwise, it is not possible to put the results into a proper perspective.		Already addressed.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We recommend to remove this sentence.		
EFPSI	6	141-143	Current wording: "It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence." The guideline should recognise that the prevalence of a disease could constrain the design, conduct and analysis of trials for small populations. In these cases, novel approaches could be considered in situations when it is difficult to recruit a large number of patients. This in turn presents a challenge of developing new methodology for trials in small populations. The lessons and recommendations from EU-funded programmes such as Asterix, IDeAl and InSPiRe on methods for clinical trials in the small population setting are particularly helpful and should play a role when assessing the clinical studies for the purpose of JCA for orphan drugs. We suggest to either remove this sentence or rephrase “The rareness of a disease or the impossibility of blinding should not be an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence. However we recognize that trials in small population pose additional challenges.”		The current guideline is about how to assess the validity of a clinical study, not about when and under what specific circumstances to use a certain study design for HTA purposes.
EFPSI	6	156	The example of ROB ‘publication bias in a systematic review or meta-analysis' may not be appropriate as meta-analyses were not in scope for this guideline.		As this sentence specifically refers to the meta-level, mentioning publication bias is fully correct.
MTE	6	128-143	**Assessment of certainty of results (internal validity, external validity, and statistical precision).** We wish to stress that any description of the certainty of the results of a given study in the context of the JCA, including the reporting of such assessment in the JCA report, must not lead to a value judgment of the relevant technology, nor should it create obstacles for the proper consideration of the submitted evidence at Member State level;		Duplicated comment. See above.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			furthermore, the assessment of the certainty of the results should not lead to a de-valuation or exclusion of the relevant studies. In addition, it is important that no critical or negative wording is included in the JCA report. This recommendation reflects both the letter and the spirit of the EU HTA Regulation, specifically: Article 9 (14) "*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*" L 458/5 (28) "*The joint clinical assessment report should be factual and should not contain any value judgement, ranking of health outcomes, conclusions on the overall benefit or clinical added value of the assessed health technology*." Article 9 L 458/16 "*Joint clinical assessment shall result in a joint clinical assessment report that shall be accompanied by a summary report. Those reports shall not contain any value judgement or conclusions on the overall clinical added value of the assessed health technology and shall be limited to a description of the scientific analysis*"		
MTE	6	137-140	All evidence should be considered, based on the context. Scientific and clinical rationale should drive the assessment and inclusion of available and generated evidence, taking into consideration		Assessing ALL evidence on a regular basis would be neither practical nor useful (see previous comments and replies).

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			acceptance by regulators, ethics committees and investigators.		
MTE	6	141-143	Evidence is by its very nature is contextual. This final sentence implies that all evidence can be assessed the same without taking into account the reason for its PICO. E.g. a rare disease that affects several hundred people worldwide will most certainly have a different evidence base than a condition or disease whose population base is in the several millions; similarly, it is unfeasible to expect the blinding of all parties involved in a study involving a surgical procedure. While an understanding that RoB will be inherently higher in different study designs, the context relating to why these study designs were chosen should not be separate from the overall assessment.		We see no necessity to change the current text.
MTE	6	147-150	**Hierarchy of evidence** A classification based on study design alone (see Section 3) is insufficient for a full assessment of internal validity; rather, the totality of evidence should be considered instead. We propose giving due consideration to the studies submitted for the purpose of the JCA and recommend that no conclusive judgement on the quality of the studies are inferred from the hierarchy of evidence alone, but that a detailed and holistic assessment of the totality of the evidence is instead conducted and reflected in the JCA report. In addition, given the Regulation's scope to review MedTech, it is important that clinical evidence across the 'evidence hierarchy' is not just considered for the purpose of the JCA, but that HTA assessors adopt a flexible approach to assessing internal and external validity		A "flexible approach" which pays similar attention to internal validity, applicability, and statistical precision, does not appear useful, as internal validity is the most important aspect. While using lower levels of evidence may be helpful in some situations, assessing ALL evidence (including case reports, animal experiments, cell and other bench research) would neither be practical nor useful.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			and statistical precision. As many Medical Devices will be developed with these considerations in mind, a flexible approach will be required to ensure EU patients can access the best new innovative Medical Technologies.		
MTE	6	162-163	A list of the RoB tools to be used in the JCA needs to be more explicit here or linked in an Appendix. It is not adequate to state "see other Guidance documents". Please state which documents there are.		We will consider referencing previous EUnetHTA guidelines on the validity of randomized and observational studies (see here: https://www.eunethta.eu/methodology-guidelines/)
Dr. Thomas Ecker, Ecker + Ecker GmbH	6	137-140	**Statement in guideline:** "Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results." **Comment:** The term `minimum level of internal validity, applicability, and statistical precision´ is unspecific and therefore unclear. A concrete threshold for the `minimum level´ should be given, since otherwise evidence could be systematically excluded. We would like to point out that single arm studies as well can contribute valuable evidence.		It is not possible to specify what "minimum level" of internal validity, applicability, or statistical precision is to be applied, because this is highly context-dependent.
Dr. Thomas Ecker, Ecker +	6	162/163	The reference is unclear. Documents should always be specified in a clear and unambiguous way.		The "different tools" for RoB assessment, which are described here only vaguely, are specifically explained in sections 4.1

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Ecker GmbH					and 4.2. No change required.
Sebastian Werner vfa	6	141 - 143	*"Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence."* Proposed wording: Please remove. ~~*"Furthermore, the certainty of results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence."*~~ The vfa does not agree with these claims. The vfa is convinced that the formal assessment of the degree of certainty should consider the medical context and must be adapted to consider the specificities of special therapeutic situations. The vfa also believes that these claims made in the guidance violate the intentions of the European HTA Regulation. The EU Regulation 2021/2282 acknowledges that for some new health technologies some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: *"Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products"*. That indicates that methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to consider the medical context. In special therapeutic situations, such as rare diseases, high medical need areas, targeted patient groups, or paediatric patients, randomised		In our view, the position of the vfa appears scientifically untenable, because the certainty of the results of a clinical study is independent from medical context. All instruments for RoB assessment and all statistical principles fail to pay any attention to the medical context of the study. Just imagine reading a clinical study on disease X without knowing whether disease X is rare or does affect children. After assessing the validity of the clinical study results, this judgment remains constant and does not change when learning that disease X is indeed very rare or leads to the death of children. We nevertheless agree that "the uncertainties of the clinical evidence in special situations should be treated more lenient", but this takes place after the assessment of clinical study validity. In fact, such deliberations belong to the responsibility of each EU member state.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			controlled trails are often not ethical or feasible. Therefore, it is not appropriate to evaluate these situations according to the same standards applied in regular therapeutic situations in which randomized controlled trials are possible. The uncertainties of the clinical evidence in special situations should be treated more lenient than in regular situations where RCTs are possible. For that, the methods must be adapted to consider the specificities of the medical context, especially with respect to rare diseases, high medical need areas, targeted patient groups, or paediatric patients. The vfa recommends considering the medical context in the formal assessment of the degree of certainty. The methods must be adapted to consider the specificities of the therapeutic situations. Uncertainties of the clinical evidence in special situations should be treated more lenient than in regular situations. For instance, in the case of rare diseases, the probability of observed results should be meaningfully defined, considering the small number of patients by increasing the necessary p-value.		
Sebastian Werner vfa	6	144 - 145	*"the overall certainty of the study results (i.e., if study results have a low level of internal validity, the levels of statistical precision and external validity are irrelevant) (5,6)."* Comment: Applicability (transferability or external validity) should be generally assumed as long as no evidence-based medical reasons are known that argue against transferability.		In principle, we agree with this comment, but the current guideline is about internal validity, not about applicability and how to assess it.
Ermisch – GKV-SV	6	120	HTA assess the added benefit of a new technology with regards to patient-relevant outcomes. As such, the comparator needed is standard of care and several outcomes might be relevant. Please adjust line 120 to reflect this by changing it to:		We will delete the "placebo" as an example comparator. Nevertheless, in a study comparing new drug X + standard of care versus standard of care, it still is

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			…to a comparator (standard of care) on outcomes of interest.		best to compare versus standard of care + placebo.
Ermisch – GKV-SV	6	137-140	We strongly support the notion that it should be possible "to not assess the evidence that ranges below a minimum level of internal validity." Clarification is needed at which point the decision on the threshold of internal validity below which results will not be considered in the assessment will be made. This might also depend on the criteria for assessment of internal validity (e.g. if they are based on study design characteristics).		It is not possible to specify what "minimum level" of internal validity, applicability, or statistical precision is to be applied, because this is highly context-dependent.
Ermisch – GKV-SV	6	141-143	It is important to note that rarity does not justify ignoring uncertainties in the context of HTA. Handling these uncertainties is part of the mandate of the decisions that will be made based on HTA. Thus, HTA must clearly outline all strengths and weaknesses of a given data set.		We thank the commentator for supporting our point of view.
BAH	6	136	"Although HTA usually requires a high target certainty of results, it is necessary to assess all available data, as submitted by the Health Technology Developer (HTD). Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results." What might be a justification to not assess the evidence? And what is meant by minimum level of internal validity? Please add definitions.		It is not possible to specify what "minimum level" of internal validity, applicability, or statistical precision is to be applied, because this is highly context-dependent.
Karen Facey	6	Line 123 and Box above 128	These sections indicate that results must be reported with their "certainty". In statistics and HTA, it is well understood that "uncertainty" must be clearly communicated, but certainty is not a generally accepted term and most would argue is not measurable. The text should be altered throughout to refer to uncertainty, instead of certainty.	X	The word "certainty" was taken from the HTA Regulation (Article 9).
Roche	6	128-133 / 2	Please add a statement around the use and relevance of external validity in discussions around generalizability.		The current guideline is about internal

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					validity, not about applicability (or external validity) and how to assess it.
Roche	6	137 / 2	We recommend that no conclusive judgment on the quality of the studies be inferred from the hierarchy of evidence alone, but that a detailed and holistic assessment of the totality of clinical evidence submitted is conducted and reflected in the JCA report. It is absolutely vital that HTA assessors adopt a flexible approach to assessing internal and external validity and statistical precision and that certain study designs or types are not excluded from review due to their inherent limitations. The guideline must allow for context specific considerations.		A "flexible approach" which pays similar attention to internal validity, applicability, and statistical precision, does not appear useful, as internal validity is the most important aspect. While using lower levels of evidence may be helpful in some situations, assessing ALL evidence would neither be practical nor useful.
Roche	6	141-142 / 2	Current wording *"Furthermore, the certainty of the results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence."* Suggested rewording *"~~Furthermore, the certainty of the results is independent of the medical context of the PICO question. It is methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence.~~ The rareness of a disease and the impossibility of blinding are independent considerations from the certainty of results. Furthermore, in conditions with small and very small populations, alternative methodological approaches (e.g. NRS, observational studies) may be acceptable if they are more appropriate given the population in question and help to improve the interpretability of the study results."*		See replies to previous comments.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We suggest the modification in order to account for the fact that the prevalence of a disease could constrain the design, conduct and analysis of trials for small populations. In these cases, novel approaches could be considered in situations when it is difficult to recruit a large number of patients. This in turn presents a challenge of developing new methodology for trials in small populations. The lessons and recommendations from EU-funded programmes such as Asterix, IDeAl and InSPiRe on methods for clinical trials in the small population setting are particularly helpful, and should play a role when assessing the clinical studies for the purpose of JCA for products that cover these populations.		
Roche	6	154 / 2	RoB discussions should be based on a clear understanding of the quantity the HTD is trying to estimate in an unbiased way. It is therefore recommended that the guideline uses the estimand concept put forward in the ICH E9(R1) addendum for that purpose. Estimands defined for HTA purposes must not necessarily be the same as, e.g., the primary estimand used for regulatory approval.		A statement on the estimand concept will be added to the guideline. However, we have to rely on those RoB instruments that are currently available.
Silke Walleser Autiero Medtronic	6	136-140	The sentences here are contradictory and not adequate for a thorough and high-quality evaluation. An explicit statement needs to be made here that all evidence relating to the PICO and research question(s) is to be identified and assessed on its individual merits. No evidence should be dismissed and not assessed. The resulting assessment if done correctly will allow the end user of the report to easily identify the included evidence that is consistent with either high internal validity, high applicability and/or statistical precision. There should not be exclusion of any piece of evidence that fits the PICO criteria before a formal assessment is conducted.		While using lower levels of evidence may be helpful in some situations, assessing ALL evidence would neither be practical nor useful.
Silke Walleser Autiero Medtronic	6	141-143	Evidence is by its very nature is contextual. This final sentence implies that all evidence can be assessed the same without taking		Please see the reply to vfa above. The certainty of the results of a clinical study is independent of medical context.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			into account the reason for its PICO. E.g. a rare disease that affects several hundred people worldwide will most certainly have a different evidence base than a condition or disease whose population base is in the several millions; similarly, it is unfeasible to expect the blinding of all parties involved in a study involving a surgical procedure. While an understanding that RoB will be inherently higher in different study designs, the context relating to why these study designs were chosen should not be separate from the overall assessment.		
Silke Walleser Autiero Medtronic	6	162-163	A list of the RoB tools to be used in the JCA needs to be more explicit here or linked in an Appendix. It is not adequate to state "see other Guidance documents". Please state which documents these are.		We will consider referencing previous EUnetHTA guidelines on the validity of randomized and observational studies. These describe the RoB tools in detail (see here: https://www.eunethta.eu/methodology-guidelines/)
Roche	6, 12, 14	141 / 2; 380 / 4.2; 425 / 4.6; 436 / 4.7	We suggest that the guideline should consider the uncertainty of evidence for all study designs, specifically where RCTs are not appropriate as uncontrolled clinical trials i.e. single arm trials, cross-sectional studies and case-series also mentioned in the guideline without any specification to assess RoB (the current guideline even recommending not to assess RoB in these cases).		We cannot see any advantage in describing uncertainty, when certainty was already assessed. Guidelines for assessing indirect comparisons (based on single-arm trials) are currently under preparation.
Matias Olsen, EUCOPE	7	168-171	There is a practical trade-off between internal and external validity. To achieve a high internal validity, the external validity would be reduced/sacrificed. The effect size observed in a study with a high internal validity is unlikely to be the same in the target population in clinical practice, under normal conditions of use. Therefore, it cannot be required to achieve within the same study high		We disagree with the commentator's assumption that internal and external validity have a reciprocal influence on each other. Pragmatic RCT are one example, where both aims are achieved.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			internal and external validity.		
Matias Olsen, EUCOPE	7	172-180	The concept of applicability appears to mix several different things especially the external validity, the generalizability, the heterogeneity of the effect size and the endpoints' validity, which are different concepts even though they may sometimes be overlapping or impacting each other. By definition, the internal validity requires inclusion/ exclusion criteria that make patients interchangeable. If a surrogate endpoint is used, this does not affect the generalizability *per se*, nor the external validity as the same effect size would be expected in the general population for this surrogate end point. Surrogacy is part of the outcome and should not be considered in the applicability. Using estimand to assess the applicability is critical and currently missing in the guideline.		Use of surrogate endpoints means that the study PICO question does not fully match the PICO question of interest. This leads to "indirectness" (in GRADE terminology) and impairs the applicability of evidence. We agree that the generalizability or external validity are not affected. This is one of the reasons why the term applicability was selected to describe the overarching concept. A statement on the estimand concept will be added to the guideline. However, we have to rely on those RoB instruments that are currently available and validated.
Matias Olsen, EUCOPE	7	191-195	The JCA needs to provide a final assessment for the agreed PICOs and discuss variability when applicable. This cannot be left open for the member states to decide. It has to be clearly discussed as part of this guideline within the JCA. It is not analysed in this document how the multiple PICOs will impact the statistical precision from the perspective of multiple analysis and related alpha risk inflation.		Statistical precision and correction for multiplicity issues are not within this guideline's scope.
Matias Olsen, EUCOPE	7	198 – 234	Different aspects of applicability (external validity and generalisability) should be addressed in the JCA. However, it must be acknowledged that increasing internal validity will reduce external validity. It is challenging in clinical trials to have a high internal validity and at the same time generate evidence relevant to all patient profiles, all settings, all conditions of use of reference intervention etc. The PRECIS tool describes		We disagree with the commentator's assumption that internal and external validity have a reciprocal influence on each other. Pragmatic RCT are one example, where both aims are achieved.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			9 dimensions to assess the ability of a trial to produce results applicable to clinical routine condition of use of a clinical trial.		
Matias Olsen, EUCOPE	7	200-202	Analysing variability among different populations subgroups, for example age subgroups, requires a post-hoc analysis. It is currently unclear when to do this analysis and how? What is the impact of the variability on the statistical precision? By definition, a post-hoc analysis has a low statistical precision while a high statistical precision is required. EUCOPE recommends clearer recommendations on post hoc analyses in the joint clinical assessments.		Statistical precision is not within this guideline's scope
Mihai Rotaru - EFPIA	7	168-171/2 General Considerations	Current wording "In statistical terms, the applicability of clinical evidence threatens the overall certainty of results if, because of relevant effect modification, the effect in the population of interest is probably different from the effects in the clinical studies." Suggested rewording "In statistical terms, the applicability of clinical evidence threatens the overall certainty of results if, ~~because of relevant effect modification, the effect in the population of interest is probably different from the effects in the clinical studies~~ **the effect in the population of interest is different from the effects in the clinical studies, because of relevant effect modifiers such as patient characteristics**. **Applicability should be assumed unless there are evidence-based medical reasons not to do so.**" [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Relevant effect modification presumably refers to an effect modifier as described in D4.5 Applicability of Evidence (line 354). In addition, EFPIA		The proposed sentence on applicability will not be used, because the guideline's focus does not include this issue.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			suggests omitting the word "probably," as the likelihood of low applicability is not relevant to its definition.		
Mihai Rotaru - EFPIA	7	175-176	Current wording […](iii) the study outcomes (e.g., surrogate outcomes) fail to offer information about the outcomes of interest Suggested rewording […](iii) the study outcomes ~~(e.g., surrogate outcomes)~~ fail to offer information about the outcomes of interest [note: strikethrough denotes proposed deletion]. Rationale The applicability of an outcome, including surrogate outcomes, is directly linked to, and a reflection of, the relevant PICOs, and should not be pre-judged on whether a given outcome is a surrogate. Should the PICO identify a given outcome, then this would be applicable for JCA, even when such an outcome was a surrogate one. Therefore, the wording in the brackets should be removed, as it would unnecessarily bias the acceptability and validity of surrogate outcomes in the context of the JCA.		The guideline clearly states that "a final judgment on applicability can only be made at the national (or even regional) level by each member state itself". If problems with surrogate outcomes are included in the applicability domain (which is scientific standard), it also will be each member state's decision whether to accept the surrogate outcome. Thus, the current guideline does not "bias the acceptability of surrogate outcomes".
Mihai Rotaru - EFPIA	7	179-180	Current wording "For the applicability of clinically relevant evidence, effect modification has to be taken into account (15). For example, if the relative effectiveness of a drug was shown to vary substantially with age, the		Since the current guideline's focus does not include subgroup analyses, there is no need to specify that such analyses have to

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			application of overall study results would be questionable. Instead, the subgroup results for the corresponding age groups or other analytical techniques could supplement information on relative effectiveness." Suggested rewording "For the applicability of clinically relevant evidence, effect modification has to be taken into account (15). For example, if the relative effectiveness of a drug was shown to vary substantially with age, **and the credibility of the subgroup results was shown using best practices for subgroup analyses, i.e., taking into account the likelihood that the differences in effects can be explained by chance,** the application of overall study results would be questionable. Instead, the subgroup results for the corresponding age groups or other analytical techniques could supplement information on relative effectiveness." [note: **bold** denotes proposed inclusion] Rationale If the design of a study did not incorporate a power calculation for subgroups of interest, then the sample size in those subgroups may not be sufficient to provide conclusive evidence on the relative effectiveness in the subgroups. This issue is particularly relevant for the D4.6 guidance as it only pertains to data from a single study. A priori, the effect estimate from the overall study population is the best estimate within each subgroup/subpopulation. Accordingly, priority should be given to higher certainty results from overall study population unless major heterogeneity can plausibly be demonstrated based on established best practice.		follow standard scientific principles.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			However, in alignment with comments submitted on the D4.5 guideline, in order to avoid wrong conclusions and data dredging due to multiple analyses that can lead to deviating results, the scope of the JCA should be based as much as possible on the prespecified analyses in the trial and should be limited to analyses that are really necessary for the JCA. As a consequence, subgroup analyses should not be a regular requirement in the JCA.		
Mihai Rotaru - EFPIA	7	204-206/2	**Requests for raw data** Current wording "To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data whenever useful." Suggested rewording "To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics. ~~including raw data whenever useful."~~ [note: ~~strikethrough~~ denotes proposed deletion] Rationale EFPIA agrees that transparency and understandability of the results are important for the conduct of the JCA. However, EFPIA also believes that pre-specified analyses, statistical analysis plans, descriptive statistics and results, along with additional detailed statistic tables (similar to those submitted to the EMA), as well as underlying documentation (such as code for non-standard specific methods and discussed in Art 9(3d)) will be sufficient to allow the assessor and co-assessor to verify the accuracy of the information in a transparent manner. Raw data is therefore not		The clause on raw data will be deleted.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			required to achieve the aims of the JCA. At Member State level, raw data/individual patient data are also generally considered not necessary for the conduct of the HTA[1]. In addition, a discussion around provision of raw data (whose definition is ambiguous) is beyond the scope of this consultation, with wider implications around privacy and GDPR, IT security, cross border data transfers, patient consent, international regulations (e.g. China's Regulation on Management of Human Genetic Resources) and proportionality to achieve the aim of the Regulation, as well circumstances for its request and criteria of how it would be analysed and used. References 1. Decision of the German Ministry of Health on the resolution of the Joint Federal Committee pursuant to Section 91 Social Code Book V, 1. resolution of March 16, 2018, on an amendment to the Rules of Procedure (Amendment of Annex I and II to Chapter 5.) https://www.g-ba.de/downloads/40-268-5737/2018_03_16_2019-02-21_VerfO_Aenderung-Anlage-II_Kapitel-5_konsolidiert_BMG.pdf		
Paolo Morgese – ARM	7	198-199	"Different aspects of applicability (primarily any PICO mismatch between assessment scope and clinical study) should be addressed in a JCA, but the final judgment on the applicability of study results must be left to the discretion of each member state". The procedural recommendation looks unnecessary, not consistent with the regulation (QUOTE ARTICLES) and highly inefficient. What would be the point of undergoing a cumbersome coordinated scoping and assessment process if its outcome can be easily dismissed by member		Applicability inevitably lies in each member state's responsibility. For example, if a new drug requires regular monitoring by lab testing or the availability of certain emergency services, not every country might have these preconditions at the local level. Therefore, a study (and its results on severe adverse

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			states on grounds of applicability of clinical evidence?		events) may fit the healthcare setting in a given member state well or not.
Marko Ocokoljic (SIOPE)	7	204, 205, 206	*"To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data and* ***Real World Data*** *whenever useful."*		The clause on raw data will be deleted.
Tanja Podkonjak, Takeda	7	204-206	Current text: To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data whenever useful. Proposed text: To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics.~~, including raw data whenever useful.~~ Rationale: We are fully committed to providing the appropriate information to the JCA assessors in a transparent, proactive and timely manner. In line with the expectations of the EMA and current HTA practices, the submission dossier and supporting materials should not require the HTD to submit individual patient level data (i.e. IPD) or raw data due to patient confidentiality. This extends to raw patient data but also patients listing or individual scans which are sometimes found in annexes of the CSR. GDPR and the storage and transfer of sensitive health information captured in IPD is a concern, but particularly if the patient population is small. Furthermore, GDPR regulation stipulates that the principle of the least amount required to be applied to the transfer and request of sensitive data, which includes health data.		The clause on raw data will be deleted.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We support the JCA to apply the same approach to trial data as it currently utilised by the EMA and most HTA agencies. EMA does not require IPD or raw data but instead the regulator review is based on the trial CSR, TLF (tables, lists and figures) and the SAP. The HTD processes and presents any additional analyses requested by EMA. Furthermore, EU HTA agencies such as IQWiG, INFARMED, NCPE, to name a few, do not currently request raw data. Takeda believes that the JCA should follow the same principles.		
Tanja Podkonjak, Takeda	7	193-195	Current text: However, in the JCA, each aspect (e.g., questionable applicability because of differences in patient population or control intervention) will only be commented on and briefly analysed, but without providing a conclusion on applicability. We question the current text and approach as it is not clear the role of the JCA assessors vs national HTA agencies. The statement above suggests that an analysis will be performed and comments made on the strengths and limitations of the data presented; however, comments on the analysis of the data are also envisioned as a part of a national HTA agencies assessment process. It is unclear what will be conducted at a EU vs local level.		In our view, the interface between EU-level and national assessment is described in sufficient detail.
EFPSI	7	181	Current wording: "Given that applicability is usually less relevant and more straightforward to assess compared with internal validity." In fact, transportability is often more difficult to assess than internal validity because a clear operationalization of the target population is lacking. As argued by e.g. doi:10.1093/aje/kwy228, considering external validity as secondary may be misleading.		The current guideline is about how to assess the internal validity of a clinical study, not about how to assess applicability.
EFPSI	7	191-194	Current wording: "This primarily includes any potential mismatch between the PICO of interest and the PICO examined in a clinical study. However, in		The current guideline is about how to assess the internal validity of a clinical

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the JCA, each aspect (e.g., questionable applicability because of differences in patient population or control intervention) will only be commented on and briefly analysed, but without providing a conclusion on applicability." This sentence suggests that mismatches between the PICO requested and the evidence provided can always be deferred to an applicability discussion at the member state level. Such a setup requires more involvement of the HTD in the scoping process to ensure alignment. Suppose that the HTD has an RCT versus one comparator in a class, but a member state wants data against another comparator in their PICO. Without more direct involvement of the HTD in scoping, the HTD may want to use the RCT in any case, in anticipation of an applicability discussion. If the member state does not agree with the HTD's arguments regarding applicability, the consequence may be that the JCA is not relevant for that member state, and that patient access is delayed. To avoid such situations, we recommend that the HTD be involved in the scoping process.		study, not about the scoping process.
EFPSI	7	206	Current wording: "[...] including raw data whenever useful." Is there a process in place to share individual patient data that respects privacy, confidentiality and integrity? Either this sentence should be removed or the process should be clarified.		The clause on raw data will be deleted
MTE	7	186	If there is an acknowledgement that PICO questions are different between EU member states, there needs to be reference back to the purpose of the JCA's overall. The issue of applicability of not only the PICO and associated evidence that is assessed but the final JCA evaluation needs to be referred to here if statements such as these		We see no need to explain why PICO questions can be different in the EU.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			are made.		
MTE	7	202-204	While reference to confidence intervals is appropriate, there are other methods to consider here (such as Bayesian methods). Consider changing this sentence to reflect that while 95% Cis are common, other approaches for consideration exist (include appropriate citations).		The statement that 95% confidence intervals are "preferably" reported includes the possibility of reporting other statistics. No change required.
François Houÿez (Eurordis)	7	204-06	*"To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data whenever useful."* For the use of raw data, do EUnetHTA21 /HTA bodies have the resources to re-analyse them? For large files, numerous data analysts and statisticians would be needed, and consideration for the time this would represent should be stated. The use of raw data should be piloted for feasibility / resources estimation.		The clause on raw data will be deleted.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ingo Hantke SKC Beratungs-gesellschaft mbH	7	185 - 195	Comment: The practical guideline states that the *applicability of a study can differ between member states* and therefore *the final judgment on the applicability of study results must be done for each member state*. The applicability of the study design is the key basis for the outcome of the HTA process, indicating that the JCA could lead in simple words to a success in one European country and simultaneously a complete failure in another European country. Conducting multiple trials each individually considering a different national context is among other things not in favor of a best possible and quick improvement of patient care. We recommend including that "applicability" is not necessarily a *black and white* question but rather a question on the "extent of applicability". In addition, it appears contradicting/not possible to "*describe and address*		In our view, the interface between EU-level and national assessment is described in sufficient detail. We agree that applicability is not a black or white question, but believe that this fact is self-explanatory. As stated elsewhere in the guideline, applicability "is usually less relevant" in comparison to internal validity. Therefore, it is quite unlikely that a new technology becomes "a success in one European country and simultaneously a complete failure in another" just because of applicability

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			specific issues in relation to the applicability in the JCA […] *without forestalling any national judgment*." In order to describe and address issues, the requirements and analysis parameters must be clear beforehand. Most of these parameters are affected by national differences. By analyzing the adequate handling of these parameters, a subsequent national judgement based on these factors is already strongly guided.		issues. No change in text required.
Ermisch – GKV-SV	7	166-167	In paragraph 1.1, the use of GRADE for HTA was discouraged. Thus, the reference to GRADE put here is unnecessary. Delete the clause „although the term „indirectness can be chosen in the context of GRADE methodology".		The use of GRADE was not discouraged. In fact, GRADE can "partially be applied". No change required.
Ermisch – GKV-SV	7	181	The statement „applicability is less relevant" lacks proper justification and should be deleted. It is highly important for follow-on decision making to know, whether the clinical studies and their results are applicable to the population of interest.		We agree that applicability is an important issue. In the authors' experience, however, issues of internal validity are by far more important than applicability issues.
Ermisch – GKV-SV	7	193-195	We agree that for applicability, the national context is very important. Given that one part of the JCA will fit PICOs of several member states in principle, details of the trial setting might be of different importance for national decision-making. Thus, a detailed description of the study setting might be more appropriate than just a comment and brief analysis.		The current guideline is about how to assess the internal validity of a clinical study, not about how to assess applicability or describe the available studies.
Bayer	7	164-198	The conclusion "Different aspects of applicability (primarily any PICO mismatch between assessment scope and clinical study) should be addressed in a JCA, but the final judgment on the applicability of study results must be left to the discretion of each member state" is not appropriate as it seems to be too generalized. It is specified for which reasons there could be national aspects that differ from the JCA. But it is one of the principles of JCA to consent on one or more PICOs and to allow additional assessment on member-state level only if a specific PICO is concerned that was not part of the JCA, for example. Therefore, a possible conclusion could be: "Different aspects of applicability should be		Applicability is largely determined by mismatches in the PICO questions, but other aspects of the healthcare setting may also be important. Therefore, it would not be appropriate to limit each member state's decision-making to only the PICO question. The current guideline in general has no right to define how member states should make use of the JCA report

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			addressed in a JCA, and the final judgment on the applicability of study results can be left only to the discretion of each member state if there is a mismatch (e.g. in PICOs) that had been addressed before" Furthermore, the following statement "Applicability is usually less relevant and more straightforward to assess compared with internal validity." should be removed, since external validity may be more difficult to assess than internal validity. In the case of RCTs: Firstly, RCTs balance both measured and unmeasured confounders on expectation. Secondly, there are well-established tools to evaluate internal validity such as the Cochrane "risk of bias" tools. Assessment of external validity is more complicated: Firstly, it depends on a strong assumption of all treatment effect modifiers being measured. Secondly, there are no well-established methods to assess or operationalize external validity with respect to a target population.		(including any national judgment on applicability).
Bayer	7	204-206	The statement "To increase the transparency and understandability of results, data submissions and analyses should contain counts and other types of descriptive statistics, including raw data whenever useful." is not sufficiently clear, with respect to cases in which raw data should be provided. The ambiguity of 'whenever useful' should be revised with examples provided, as cases where such data would be provided can be expected to be rare.		The clause on raw data will be deleted.
Karen Facey	7	Para starting l181	(It is welcomed that applicability will be judged qualitatively on a case by case basis, as checklists are never sufficient for all contexts.)		Thanks for this supportive comment.
Roche	7	206 / 2	Current wording: *"[...] including raw data whenever useful."* Suggested wording: *"[...]* ~~*including raw data whenever useful."*~~ Reason for change of wording:	X	The clause on raw data will be deleted.

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			1. **Proportionality** The request to submit patient data listings and / or raw data on a regular basis within the Joint Clinical Assessment ("JCA") submission dossier is disproportionate to the purpose of a clinical assessment based on the clinical relative effectiveness of the targeted (study) population(s). No justification or explanation is given why the use of aggregated data submitted by the HTD according to the PICOs defined by the member states is not sufficient to carry out the JCA appropriately and fully in line with the HTA Regulation. It is not apparent with which other data from these patients the "patient data listings" must be linked in order to properly conduct a JCA. The principle of proportionality also applies to the question of the right balance between public and private interests. The overall focus must be on the patient and on the access to innovative and affordable medicines (see the Pharmaceutical Strategy of the EU Commission, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0761). Furthermore it does not comply with the principle of data minimization according to article 5 of the European General Data Protection Regulation ("GDPR") which is also followed in patient consent forms for the participation in clinical trials to submit data not needed for the purpose of conducting a JCA.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			These principles have been followed by most HTA authorities to date. A respective decision of the German Ministry of Health from 2019 can be referenced here: The German Ministry of Health decided in 2019 that for reasons of data minimization according to Art. 5 I lit. c of the European General Data Protection Regulation and the principle of proportionality, "patient data listings" of CSRs must not be a general mandatory component of dossiers submitted by HTD. The Ministry concluded that pseudonymized individual patient data (in such listings) are generally not necessary for ensuring that the evaluation process is designed appropriately and functionally. At most, the admissibility of a case-by-case request (as anonymized data) to the HTD could be considered, should this be necessary for the health technology assessment in specific cases. (see GBA decision and communication: https://www.g-ba.de/downloads/40-268-5737/2018_03_16_2019-02-21_VerfO_Aenderung-Anlage-II_Kapitel-5_konsolidiert_BMG.pdf The decision of the Federal Ministry of Health is reflected in the following instructions for the german HTA dossier: Appendices containing individual patient information (patient data listings) or other individual personal information (e.g. information about investigators) need **not** be included and may be removed or made unrecognizable as appropriate . 2. **2. GDPR and Data Security** If the assessor and co-assessor have access to raw data including individual patient characteristics and individual patient results (which can be considered personal data) for the purpose of scientific evaluation, they are processing personal data and GDPR		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			is applicable. So far there is no adequate proof of concept presented to do this in a compliant way based on the following reasons. • There is no legal basis that justifies processing personal data for the European HTA assessment on EUnetHTA or Health Technology Assessment Regulation ("HTAR") level. • In that case the assessor and co-assessor are controllers of the data and thus obligated to inform the patients and responsible for possible data subject rights requests in accordance with GDPR. • Furthermore, technical and organizational measures must be taken to protect the data and audit rights to control compliance. If EUnetHTA or HTAR cannot provide an adequate proof of concept, case by case and only anonymous data shall be used for the purpose of scientific evaluation. **3. 3. Lack of Confidentiality framework** The additional data requirements laid out in the consultation document warrant an elaborated and with the HTD agreed upon confidentiality framework. As this is not yet developed for the European level and the specific purpose of the Joint Scientific Consultation ("JSC") and JCA procedures under the HTAR legal and compliance implications are currently difficult to oversee. In country settings in European member states the legal frameworks clearly regulate the appropriate handling of confidential data: which data will be published, which data will		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			not be published, which data can be handed over by the pharmaceutical company, which data can only be handed over redacted etc. This is not the case for the HTAR yet. In the absence of a clear confidentiality framework, higher data requirements in the EUnetHTA21 publication documents than in the past for JCA pilots create an unforeseeable risk for companies and patients. HTD should be involved in protecting their Commercial Confidential Information ("CCI") and in developing the confidentiality framework needed. 4. **4. Intellectual property of companies** Beneath the concerns around data protection rights the potential publication of confidential data at the given time point of the european HTA assessment risks to heavily affect the Intellectual Property (IP) of companies. As long as there is no agreed upon confidentiality framework including clarity which data will be published higher data requirements than in the past create an unforeseeable risk for companies and patients. There must be an appropriate procedure to protect the publication of information that could unreasonably prejudice the business interests and thus the protection of the property of individuals or companies. Otherwise, it will be impossible for HTD to bring innovative products to market in a sustainable manner.		
Roche	7	208	Current Wording: *"Statistical testing in a clinical study requires transparent and clear prespecification of hypotheses"*	X	The estimand concept will be shortly mentioned in the guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Proposal: Please add in a comment on the need for estimands to be specified within the hypotheses of the study, following the framework outlined in the ICH E9 estimands addendum.		
Silke Walleser Autiero Medtronic	7	186	If there is an acknowledgement that PICO questions are different between EU member states, there needs to be reference back to the purpose of the JCA's overall. The issue of applicability of not only the PICO and associated evidence that is assessed but the final JCA evaluation needs to be referred to here if statements such as these are made.		Duplicated comment
Silke Walleser Autiero Medtronic	7	202-204	While reference to confidence intervals is appropriate, there are other methods to consider here (such as Bayesian methods). Consider changing this sentence to reflect that while 95% CIs are common, other approaches for consideration exist (include appropriate citations).		Duplicated comment
Jasmine Toomey PHMR	7	Lines 187/188, section 2	It is not clear who will make the final judgment on applicability.		The guideline describes that "a final judgment on applicability can only be made at the national (or even regional) level by each member state itself". As it depends on national law, who will make the final judgment and decision, the guideline cannot contain such information on each member state.
François Houÿez (Eurordis)	7-8	207-216	One aspect is not discussed: when different interim analysis are performed throughout a study (it is discussed in D 4.5 though)		The guideline on Applicability of Evidence includes recommendations on how to deal with statistical multiplicity. No change required here.
Matias Olsen, EUCOPE	8	222-224	The proposed guideline suggests that there is no way to clearly evaluate the clinical relevance of the effect size and that this should be judged at the national level.		The current guideline does not prohibit that clinical relevance of effect sizes is

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			This is inappropriate as the clinical relevance is driven by some rules that are not country specific but based on widely accepted or established scientific criteria, such as MCID (minimum clinical important difference). For transparency and clarity purposes, the guideline should include references to the use of well-established MCID for a given outcome, or MCID, calculated according to approved methodologies, such as distribution or anchor point method, to inform the clinical relevance of effect size. The clinical relevance of the effect size should be part of the JCA and should not be done at MS level.		commented on and briefly analysed in a JCA (similar to surrogacy issues). The final decision, however, must be left to each member state, as some member states may want to apply stricter thresholds for defining clinical relevance, while other may accept all statistically significant effects regardless of their clinical relevance.
Matias Olsen, EUCOPE	8	230-233	Remove the sentence as this is arbitrary. The best available evidence should be used for evaluation. "~~For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p-value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30).~~"		We believe that highlighting very large effect sizes can support efficient decision-making at member state level. For the sake of transparency and reliability, this requires the definition of certain thresholds, even if they are arbitrary.
Mihai Rotaru - EFPIA	8	211-213/2 General Considerations	Current wording "Data-driven statistical tests provide results of low internal validity. Similarly, early stopping of clinical studies, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results" Suggested rewording "Data-driven statistical tests provide results of low internal validity. Similarly~~, early stopping of clinical studies, deliberate extension of recruitment, and~~ selective reporting of results ~~all~~ undermine the validity of study results."		We agree that early stopping of a clinical study is statistically appropriate, when specific methods were applied. We will insert the word "unplanned" before "early stopping" to make it clearer that this statement refers only to "data-driven" analyses, as described in the preceding sentence.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			[note: ~~strikethrough~~ denotes proposed deletion] Rationale There are statistically valid methods for sample size re-estimation and early stopping (for benefit or futility) that control overall type 1 error rates and thus do not undermine the validity of study results when implemented appropriately. For example: Early stopping: •DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. Stat Med. 1994 Jul 15-30;13(13-14):1341-52; discussion 1353-6. doi: 10.1002/sim.4780131308. PMID: 7973215. Sample size re-estimation: •Cui, L., Hung, H.M.J. and Wang, S.-J. (1999), Modification of Sample Size in Group Sequential Clinical Trials. Biometrics, 55: 853-857. https://doi.org/10.1111/j.0006-341X.1999.00853. Mehta, C.R. and Pocock, S.J. (2011), Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statist. Med., 30: 3267-3284. https://doi.org/10.1002/sim.4102		
Mihai Rotaru - EFPIA	8	214-216/2 General Considerations	Current wording "However, the rates of type I and type II errors in a clinical trial are not directly related to the validity of the observed treatment effects, because these errors are relevant only when interpreting the results of statistical tests" Suggested rewording ~~"However, the rates of type I and type II errors in a clinical trial are not~~		In our view, this sentence is useful to lead the reader back to the main topic of the guideline, which is the validity of a given study.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			~~directly related to the validity of the observed treatment effects, because these errors are relevant only when~~ ~~interpreting the results of statistical tests~~" [note: ~~strikethrough~~ denotes proposed deletion] <u>Rationale</u> This statement appears to be *a non sequitur*, as it does not seem to be related or connected with the rest of the content of the paragraph (except that type 1 and type 2 errors are concepts indirectly related to hypothesis testing and the paragraph is about hypothesis testing). EFPIA recommends that it is removed from the guideline to avoid any confusion in its interpretation. If included, it should be reworded to align with relevance to the paragraph.		
Mihai Rotaru - EFPIA	8	222-224	<u>Current wording</u> "Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context" <u>Suggested rewording</u> "Similarly, the ~~clinical~~ relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context." [note: strikethrough denotes proposed deletion]. <u>Rationale</u> There is no reason to assume that an effect measured by responder analysis, especially if the threshold for assignment as a responder or non-		In our view, stating in a JCA that an effect is (clinically) relevant already infringes each member state's national decision-making.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			responder, according to guidance 4.5, "corresponds to validated and consensus cut-off values[1]", should or could have different clinical relevance (or patient relevance) in different Member States. The validated/consented threshold already defines the clinical relevance of a response. Therefore, it doesn't follow that the *clinical* relevance is to be judged at the national context level. Rather, the relevance of the effect size of this (validated/consented) clinically relevant effect to the national care context evaluation is subject to national judgement. References: 1. EUnetHTA 21 – Individual Practical Guideline Document - D4.5 – Applicability of evidence – practical guideline on multiplicity, subgroup and post-hoc analyses, page 12, row 328.		
Mihai Rotaru - EFPIA	8	227-229/2 General Considerations	Current wording "Which effect sizes can be considered very large and which p values can be accepted as sufficiently low is an unresolved scientific question" Suggested rewording "~~Which effect sizes can be considered very large and which p values can be accepted as sufficiently low is an unresolved scientific question.~~ **There cannot be a consensus on which effect sizes can be considered very large and which p values can be accepted as sufficiently low as this is subjective and will depend upon the context, including endpoint, population and clinical relevance**." [note: **bold** and ~~strikethrough~~ denotes proposed inclusion and deletion, respectively] Rationale Effect sizes that are "large enough" and p-values that are "small enough" are not, per se, scientific questions. If they were, one could		We believe that highlighting very large effect sizes can support efficient decision-making at member state level. For the sake of transparency and reliability, this requires the definition of certain thresholds, even if they are arbitrary.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			hypothetically design an experiment and use the empirical evidence generated from said experiment to determine what effect size is large enough and what p-value is small enough. Rather, what can be considered a large effect size or small p-value is subjective and context dependent (i.e., different standards are used in different fields).		
Mihai Rotaru - EFPIA	8	226-234	Current wording "Nevertheless, in the context of a JCA, it might be helpful to highlight such situations, especially when no RCT evidence is available. For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a *p* value <0.01 (as an indicator of sufficient precision) was proposed as a "rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30). The JCA report will describe effect estimates, but without a conclusion on whether the certainty of results is increased, because this is best made at the national level." Suggested rewording ~~"Nevertheless, in the context of a JCA, it might be helpful to highlight such situations, especially when no RCT evidence is available. For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a *p* value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30).~~ The JCA report will describe effect estimates, but without a conclusion on whether the certainty of results is increased, because this is best made at the national level." [note: ~~strikethrough~~ denotes proposed deletion] Rationale EFPIA believes that describing an effect size that is large enough to safely exclude the possibility of no effect, when no RCT evidence is available, as stated in rows 226-234 of the guideline, corresponds to a value		We believe that highlighting very large effect sizes can support efficient decision-making at member state level. For the sake of transparency and reliability, this requires the definition of certain thresholds, even if they are arbitrary.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			judgement. The amount of uncertainty that is acceptable in such instances by individual Member States should be decided solely at national level. In addition, p-values are not indicators of precision, as they are influenced both by the absolute magnitude of the effect and the precision. In particular, one can obtain the same p-value from a small magnitude effect with high precision as a large magnitude effect with low precision[1]. Furthermore, state-of-the-art methodology, such as propensity scoring and matching-adjusted analyses can help address the challenges of non-randomised evidence and reduced potential bias. References 1. Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108		
Tanja Podkonjak, Takeda	8	229-232	Current text: Nevertheless, in the context of a JCA, it might be helpful to highlight such situations, especially when no RCT evidence is available. For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30). Proposed text: Nevertheless, in the context of a JCA, it might be helpful to highlight such situations, especially when no RCT evidence is available. ~~For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p value <0.01 (as an indicator of sufficient precision)~~		We believe that highlighting very large effect sizes can support efficient decision-making at member state level. For the sake of transparency and reliability, this requires the definition of certain thresholds, even if they are arbitrary. Highlighting very large effect sizes does not constitute an infringement of each member state's national decision process.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			~~was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30).~~ Rationale: The sentence above, suggesting a 'rule of thumb' or threshold for an acceptable size of relative risk is a value judgement – the size of treatment effect or relative risk reduction required for non-randomised evidence to be considered. According to the HTA Regulation, JCA should not contain any value judgment or the clinical added value of an assessed technology. Takeda would like to highlight the following articles from the HTAR [1]: Article 9 (14) "*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*" L 458/5 (28) "*The joint clinical assessment report should be factual and should not contain any value judgement, ranking of health outcomes, conclusions on the overall benefit or clinical added value of the assessed health technology*." Article 9 L 458/16 "*Joint clinical assessment shall result in a joint clinical assessment report that shall be accompanied by a summary report. Those reports shall not contain any value judgement or conclusions on the overall clinical added value of the assessed health technology and shall be limited to a description of the scientific analysis*" The size of the treatment effected or relative risk reduction or acceptable treatment effect for any technology, regardless of its supporting evidence base, should remain a decision at a national level based on the value judgement of that individual MS. We would like to highlight that we support the input of clinical experts to set indicators of sufficient precision;		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			however, this is best done with the clinical and patient experts for the given therapy area (vs. an arbitrary value as proposed) and as different societies hold diverse values in the EU, this is best done nationally with local experts; supported by the HTAR. [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).		
Hervé Tchala Vignon, Zomahoun/ INESSS	8	244-246/ 3.1	The text on the distinction between interventional and observational studies could be confusing. In fact, the routine clinical care could be considered as an intervention for a given patient group until a new intervention becomes available. Moreover, it could be difficult to identify a study as an interventional study without its study protocol considering the statement reported. However, one knows that the study protocols are not always available for the interventional studies. The intervention under assessment can also be assigned to participants as described or not in the study protocol for many reasons (e.g., inappropriate concealment). Therefore, here are two suggestions: 1. First, don't use the same term to name exposure in the interventional and observational studies. Exposure could be named: a. **Intervention** in the interventional studies b. **Exposure factor** or **risk factor** in the observational studies 2. Second, the main distinction between observational and interventional studies should be based on whether the investigators control or not the exposure at the beginning of study.		Exposure is now introduced in the new version of the guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			The present comment is also relevant for the sections **3.1.2 and 4.4** if considered. The following reference could be useful: Methods for the development of NICE public health guidance - Process and Methods. Sept. 2012 https://www.nice.org.uk/process/pmg4/resources/methods-for-the-development-of-nice-public-health-guidance-third-edition-pdf-2007967445701		
Hervé Tchala Vignon, Zomahoun/ INESSS	8	247/3.1.1	As presented for the observational studies, specific study designs should be described for the interventional studies in this section. The interventional studies are the most frequently used in the health technology assessment. So, a good clarification of different interventional studies could be a valuable contribution for the validity assessment of clinical studies on health technology: RCTs; cluster RCTs; controlled before-and-after trials; before-and-after trials; and interrupted time series. The following reference could be useful: Methods for the development of NICE public health guidance - Process and Methods. Sept. 2012 https://www.nice.org.uk/process/pmg4/resources/methods-for-the-development-of-nice-public-health-guidance-third-edition-pdf-2007967445701		We do not think specifying that randomization can be done at the subject or cluster level is very useful here as cluster RCTs are uncommon to be performed for the technologies addressed in the HTAR. Moreover, as comparative effectiveness assessment is the cornerstone of HTA, we do not think it adds a lot of info to mention controlled before-after trials.
Hervé Tchala Vignon, Zomahoun/ INESSS	8	248-249/ 3.1.	It would be helpful to report a clear definition of an interventional study.		It will be reworded for the next version of the draft.
EFPSI	8	212-213	Current wording: "early stopping of clinical studies [...] undermine the validity of study results." This sentence suggests that well-conducted group sequential trials have lower validity. Furthermore when trials are stopped early at about 2/3 of		This part of the sentence will be modified (see comments above).

Formatted: French (France)

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			information, the effect estimate is around the effect that was assumed for powering the trial. Also, the bias in effect estimates is typically negligible (https://journals.sagepub.com/doi/10.1177/1740774509102310 or https://pubmed.ncbi.nlm.nih.gov/27271682/). Thus we would appreciate to understand in what sense stopping a trial early might invalidate results or consider our proposal below. Proposed rewording: "early stopping of clinical studies, if not properly planned at the design stage [...] may undermine the validity of study results."		
EFPSI	8	227-229	Current wording: “Which effect sizes can be considered very large and which p values can be accepted as sufficiently low is an unresolved scientific question (29).” This is not a general scientific question, it depends entirely on the context. Importantly, in the context of the JCA, it is a value judgment. We suggest to remove this sentence.		We believe that highlighting very large effect sizes can support efficient decision-making at member state level. For the sake of transparency and reliability, this requires the definition of certain thresholds, even if they are arbitrary.
EFPSI	8	230-233	Current wording: “For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p value <0.01 (as an indicator of sufficient precision) was proposed as a ‘rule of thumb’ (i.e., an arbitrary rule based on expert opinion) (26,30).” Generally accepted thresholds for the magnitude of an effect size or the precision of a result do not exist. Therefore, the proposed thresholds of 5 or 0.2 for binary effect measures or a p value of less than 0.01 seem arbitrary and lack an underlying rationale. The decision on how the magnitude or the precision of an effect estimate impacts the certainty of the results is an individual, context specific question and should be addressed on a case-by-case basis.		See previous reply.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We suggest to remove this sentence.		
EFPSI	8	244-246	Current wording: "Distinction between interventional and observational studies depends on whether the intervention under assessment is assigned through the study protocol (interventional) or is given during routine clinical care (observational)." The protocol is not the only determining factor if a study is interventional or observational. It depends on how participants are assigned to the interventions. Proposed rewording: "Distinction between interventional and observational studies depends on how the intervention under assessment is assigned by the investigator or is given during routine clinical care."		It will be clarified for the next version of the draft.
EFPSI	8	Practical Guideline (requirement for JCA reporting)	"To describe statistical precision accurately, effect estimates should always be accompanied by the corresponding measures of variation, preferably CIs at a specified 1-α level of confidence, which is 0.95 (95%) in most cases." Please develop what is meant with "in most cases". Would there be exemptions based on study protocols? In addition, we should specify in what situations you can decide whether to use 1-sided or 2-sided.		Details on statistical analyses are out of the scope of the present guideline.
EFPSI	8	Sections 3.1. and 3.1.2	"Studies are classified into two categories: interventional studies and observational studies" Some more granularity in the definition may be needed, e.g., prospectively defined cohort study, where visits are mandated but not the treatment. This may require adjusting definition in Section 3.1.2 to "In observational studies, there is no forced change in routine care (except when the protocol requires visits at specific timepoints) and neither is the usual decision for intervention		It will added in the next version of the draft.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			affected by an observational study."		
MTE	8	230-233	This sentence is unclear. Using the term 'rule of thumb' does not appear to be a recommendation and therefore will be open to interpretation in any JCA.		The fact that a 'rule of thumb' has to be applied is uncritical, as the final decision has to be made at national level.
François Houÿez (Eurordis)	8	233-234	*"The JCA report will describe effect estimates, but without a conclusion on whether the certainty of results is increased, because this is best made at the national level."* One objective of assessing technologies jointly at the European level is precisely to develop efforts to align views from different HTA bodies. To refer the scientific assessment back to the national/regional level would represent some failure to work jointly. The explanation why EUnetHTA could not adopt a consensus on which effect size (e.g. relative risk superior to 5 (or inferior to 0.2)) and which precision (e.g. a p value <0.01) could be accepted is not convincing.		Developing a European consensus on this issue (very large effect size) appears to be a largely unnecessary and complex task.
Dr. Thomas Ecker, Ecker + Ecker GmbH	8	222-224	**Statement in guideline:** "Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context." **Comment:** Clinical relevance is commonly evaluated using minimal clinical important differences (MCID). MCIDs are an internationally accepted concept and are chosen on a scientific research basis and validated using established methods, such as anchor-based methods in a therapeutic area. The guideline should mention and encourage the use of established MCIDs. Furthermore, criteria should be given regarding when an MCID is		See reply to comment above.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			considered valid and will be used. The concept of MCID is independent from the national context. It should be mentioned in this guideline or at least in the EUnetHTA Practical Guideline Endpoints.		
Dr. Thomas Ecker, Ecker + Ecker GmbH	8	230-233	**Statement in guideline:** "For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a *p* value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30)." **Comment:** The guideline should not contain 'rules of thumb'.		See reply to comments above.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ingo Hantke SKC Beratungs-gesellschaft mbH	8	230 - 234	Comment: The description of a "Relative Risk superior to 5" is not sufficient to clearly state the required threshold since it does not include the respective confidence interval. A "p-value of <0.01" based on an "arbitrary rule of thumb" as part of an official document is scientifically highly misleading not in line with the general statistical requirements in the guidelines. We recommend excluding or at least changing these requirements to a consistent and vastly established threshold p value of $p < 0.05$.		See replies to comments above.
Sebastian Werner vfa	8	137 - 140	*"Nevertheless, there might be justification to not assess the evidence that ranges below a minimum level of internal validity, applicability, or statistical precision in detail, if the PICO question can be sufficiently answered on the basis of higher-certainty results."* Suggested rewording		Duplicated comment

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			"In a situation where higher and lower-level evidence is available to address the PICO questions, preference will be given to the higher-level evidence for a JCA given it has a higher level of internal validity, applicability, or statistical precision." All evidence should be considered, based on the context. Scientific and clinical rationale should drive the assessment and inclusion of available and generated evidence, taking into consideration acceptance by regulators, ethics committees and investigators in situations where RCTs are not suitable or feasible. Excluding relevant information from being considered in the assessment in the face of other higher validity evidence could be interpreted as a value judgement and may risk excluding Member States from giving due consideration to evidence that may be of relevance.		
Sebastian Werner vfa	8	211 - 213	*Data-driven statistical tests provide results of low internal validity. Similarly, early stopping of clinical studies, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results (23)."* Suggested rewording "Data-driven statistical tests provide results of low internal validity. Similarly~~, early stopping of clinical studies, deliberate extension of recruitment, and~~ selective reporting of results all undermine the validity of study results " Comment: First, there are statistically valid methods for sample size re-estimation and early termination (for utility or futility) that control the overall type 1 error rate and thus do not undermine the validity of study results when implemented appropriately. For example:		Duplicated comment

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Early stopping: DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. Stat Med. 1994 Jul 15-30;13(13-14):1341-52; discussion 1353-6. doi: 10.1002/sim.4780131308. PMID: 7973215. Sample size re-estimation: Cui, L., Hung, H.M.J. and Wang, S.-J. (1999), Modification of Sample Size in Group Sequential Clinical Trials. Biometrics, 55: 853-857. https://doi.org/10.1111/j.0006-341X.1999.00853. Mehta, C.R. and Pocock, S.J. (2011), Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statist. Med., 30: 3267-3284. https://doi.org/10.1002/sim.4102 Second, in the case of an extended recruitment period that results in more patients being enrolled than planned, an adjustment in the number of outcomes required and the associated delay in the final analysis may be quite reasonable. In contrast to the approval process, the HTA process is concerned with using continuously growing data from discontinued studies to have a better estimation for the assessment through the information gain. For example, the Cochrane Institute uses cumulative meta-analysis, in which meta-analyses are repeatedly updated to be able to base therapy decisions on the most up-to-date information possible. On this, Clarke et al. state "This large, unique collection of cumulative meta-analyses highlights how a review of the existing evidence might have helped researchers, practitioners, patients and funders make more informed decisions and choices about new trials over decades of research." Subsequent data cuts generally increase the database and thus the certainty of results. This should also be addressed in Guideline D4.6. Clarke M, Brice A, Chalmers I (2014) Accumulating Research: A Systematic Account of How Cumulative Meta-Analyses Would Have Provided Knowledge, Improved Health, Reduced Harm and Saved Resources. PLoS ONE 9(7): e102670. doi:10.1371/journal.pone.0102670		

EUnetHTA 21 Public Consultation Comments and Responses
***Of D4.6* – Validity of clinical studies**

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Sebastian Werner vfa	8	222 - 223	*"Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context."* Comment: The clinical relevance of an effect size should not depend on national considerations but on commonly aligned validated or established scientific criteria. Suggested change/addition: "Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), should be based on established or validated scientific criteria."		Duplicated comment
Sebastian Werner vfa	8 9	230 – 232 368 - 371	*"For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30)."* *"There are different methods that can be used to control for confounding (i.e., allowing if properly conducted, conditional exchangeability, e.g., design-based methods, such as stratification or matching, or modeling-based methods, such as adjustment or models of causal inference (e.g., propensity scores or g-computation)] within the trial."* Suggested rewording: Please remove. *"~~For effect sizes expressed as relative risks, the threshold of a relative risk superior to 5 (or inferior to 0.2) and a p value <0.01 (as an indicator of sufficient precision) was proposed as a 'rule of thumb' (i.e., an arbitrary rule based on expert opinion) (26,30).~~"*		Duplicated comment

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			In cases where an RCT cannot be conducted, the best available evidence should be used for evaluation. Describing an effect size large enough to safely exclude the possibility of no effect, when no RCT evidence is available, as stated in lines 226-234 of the guideline, corresponds to a value judgement. This arbitrary rule based on expert opinion should not be used here. Instead, to reduce the level of uncertainty, methods to control for confounding factors should be used as described in L 368 - 371 of this guideline.		
Ermisch – GKV-SV	8	217-225	The paragraph seems to open the possibility to prove added benefit using a single study that failed its primary endpoint. This can only be true in very selected cases when the primary endpoint is of no relevancy at all and the study's statistical power is sufficient for the endpoint of choice taking into regard multiplicity.		The paragraph does not contain any distinction between primary and secondary outcomes or analyses. Therefore, we do not share the commentator's concerns. Multiplicity issues will be addressed in another guideline.
Ermisch – GKV-SV	8	226-234	This paragraph seems to focus on endpoints and should be relocated to the appropriate guideline. The certainty "of a positive or negative effect will be higher if a very large effect size was found" might be misleading, as this will only be true if (all) other things are equal – the impression, that a stronger effect size might compensate in particular for systematic error/bias/lack of internal validity should be avoided.		We disagree with this comment. If a study shows a very large effect size, this may compensate for risk of bias, if it is unlikely that the amount of bias could have led to the observed effect.
Ermisch – GKV-SV	8	240-241	The use of the categories interventional studies and observational studies is problematic regarding single arm trials, as it indicates a difference to single cohorts that is irrelevant regarding their reliability. It indicates a difference to single cohorts that is irrelevant regarding their reliability (moreover, the theoretical distinction regarding the locus of the decision to intervene made in lines 383/384 might be difficult to assess in practice in uncontrolled studies). In many cases, studies with new active (medicinal) substances will not be observational simply because the new		As stated on lines 237-239, classification and definition can vary. We maintain our initial proposal.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			active substance was not part of routine clinical practice at the time of the study. This might also be true for trials with medical device. Thus, categorisation should be rethought.		
	8	222	"Similarly, the clinical relevance of an effect size, which can be assessed by comparing the effect size with a predefined threshold or by responder analyses (25), needs to be judged at the national context." One of the main goals of EU-HTA is harmonization. Therefore, the clincal relevance of an effect size must be defined and be valid for all Member States.		See replies to previous comments.
Roche	8	212 / 2	Current wording: *"Similarly, early stopping of clinical studies, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results."* This statement is not clear in its current form and might even be misleading. Typically, when trials are stopped early at about two-thirds of information the effect estimate is around the effect that was assumed for powering the trial. Also, the bias in effect estimates is typically negligible (https://journals.sagepub.com/doi/10.1177/1740774509102310 or https://pubmed.ncbi.nlm.nih.gov/27271682/). Suggested wording: *"Similarly, ~~early stopping of clinical studies~~, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results."*	X	This part of the sentence will be modified in order to clarify that this statement refers only to unplanned 'data-inspired' stopping of trials (see comments above and https://pubmed.ncbi.nlm.nih.gov/20332404/).
Roche	8	218 / 2	Please include more details around the definition and inherent design elements included within non-inferiority and equivalence studies. These parameters would be important to be included in the guideline as they are called out explicitly within the text.		As for superiority trial, we won't include more details for these designs.
Roche	8	244 / 3.1	Please consider acknowledging the grey areas in between interventional		Pragmatic trial are already discussed in

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			and observational, such as pragmatic or point of care trials.		section 5.
Roche	8	240-243 / 3.1	This needs more clarity on how this aligns with other EU regulations, such as CTR, where the definitions differ slightly.		As stated on lines 237-239, classification and definition can vary. We maintain our initial proposal.
Silke Walleser Autiero Medtronic	8	230-233	This sentence is unclear. Using the term 'rule of thumb' does not appear to be a recommendation and therefore will be open to interpretation in any JCA.		This is not a recommendation, but a part of introduction/context.
Jasmine Toomey PHMR	8	Line 215, section 2	Would be useful to define what type I and type II errors are in clinical trials.		They are already defined in other EUnetHTA 21 guidelines, as well as in the literature, and do not need to repeat here.
Jasmine Toomey PHMR	8	Line 257 section 3.1.1	Further information on blinding could be provided such as double, triple blinding etc.		We have deliberately chosen not to characterize blinding in this guideline.
Mihai Rotaru - EFPIA	8 and page 10	240-256 Figure page 10	Intervention and observational studies could be seen as ambiguous. In the scope of the consultation, the authors may want to consider terms including experimental, non-experimental, comparative, and non-comparative instead.		As stated in lines 237-243, we chose a definition, acknowledging others could exist.
Denis Lacombe EORTC	9	Table 3.1 General	The table is confusing as it seems to describe the characteristics of interventional studies but covers different types of studies i.e. by definition a single arm trial will not be randomized. It is suggested to refer to this table rather as definitions of terminologies used for interventional studies. The description of some of the forms and methods of data collection appear to be unprecise and should be aligned with the common understanding		As detailed in our guideline (figure 3.1), single arm trial is an interventional study.
Matias Olsen, EUCOPE	9	269-270	"Association does not imply causality". This is true for all studies including interventional studies. However, by design, interventional studies when randomized, blinded, and		We consider that the original sentence: 'It is generally important to remember that association does not necessarily 269 imply

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			when tight inclusion/ non-inclusion criteria have been applied, then the internal validity increases, and the association will imply causality. Increasing applicability will reduce the ability of the clinical study to imply causality. Achieving causality or applicability can't be achieved with the same study.		causality' is right, and we maintain it.
Mihai Rotaru - EFPIA	9	257, Table 3.1	For clarity, we recommend that it should be indicated that any one of these characteristics would be sufficient for the study to be classified as interventional. For example, a study can be interventional without it being controlled, or randomised, or blinded)		The figure 3.1 already has these elements.
Mihai Rotaru - EFPIA	9	264-265	Current wording "Observational studies can be either descriptive, that is, without a control group (case series and cross-sectional studies) or analytical (case-control and cohort studies) with a control group. Analytical studies provide a measure of the association between exposure (notably interventions) and outcome of interest. In a case-series, changes over time can be analysed (i.e., before and after the introduction of the treatment of interest); however, under usual circumstances, such before–after changes are unlikely to assess interventional effects." Suggested rewording "Observational studies can be either descriptive, ~~that is, without a control group (case series and cross-sectional studies) or analytical (case-control and cohort studies) with a control group. Analytical studies provide a measure of the association between exposure (notably interventions) and outcome of interest. In a case-series, changes over time can be analysed (i.e., before and after the introduction of the treatment of interest); however, under usual circumstances, such before–after changes are unlikely to assess interventional effects."~~ **that is, they neither aim to evaluate a causal relationship between a population characteristic and the occurrence parameter, nor to quantify the association**		We maintain our original proposal.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			between an exposure and an outcome), or analytical (when they aim to quantify the association between an exposure and an outcome)". Rationale The terminology "with a *control* group" could be confusing: for some, it refers to a control group in a case-control design; most often, for comparative cohort studies, the comparative cohort is not typically referred to as a control group. In addition, cohort studies are provided as an example of an analytic study, but they can also be a descriptive observational study. Furthermore, descriptive studies do not always exclude a control arm; one study can have two arms and still be descriptive. Whether a study is descriptive or not depends on the nature of its research question[1,2]. References 1. Doody O, et al., 2016. Setting a research question, aim and objective. Nurse Researcher. 23, 4, 19-23; 3. ENCePP Guide on Methodological Standards in Pharmacoepidemiology: https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml		
Mihai Rotaru - EFPIA	9	272-285	**Prospective and retrospective** The classification of studies as either prospective or retrospective is not straightforward, and these terms do not, *per se,* convey a clear message about the study. Early writers defined prospective and retrospective studies to denote cohort and case-control studies, respectively. The most important study feature that these terms highlight would be whether the disease could influence the exposure information in the study.		See previous comment.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			In particular, the study design can be highlighted by: a. The order in time of the recording of exposure information and the occurrence of disease. Assessing exposure by recall after disease has occurred is a feature of many case-control studies, which may explain why case-control studies are often labelled retrospective. However, not all case-control studies involve recall. For example, case-control studies that evaluate drug exposures have prospective measurement if the information on the exposures and other risk factors is taken from medical records that predate disease development. These case-control studies may be appropriately described as prospective, at least for exposure measurement. Also, some studies may combine prospective and retrospective measurement of variables. b. The timing of the accumulated person-time with respect to the study's conduct. Here, when the person-time accumulates before the study is conducted, it is said to be a retrospective study, even if the exposure status was recorded before the disease occurred. When the person-time accumulates after the study is conducted, it is said to be a prospective study; in this case exposure status is ordinarily recorded before disease occurrence, although there are exceptions. Cohort or case-control studies can ascertain events either prospectively or retrospectively from the point of view of the time that the study begins. According to this usage, prospective and retrospective describe the timing of events under study in relation to the time the study begins or ends; prospective refers to events concurrent with the study and retrospective refers to use of historical events. In summary, EFPIA notes that using the "prospective" and "retrospective" terms in the current version of the guideline conveys no additional		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			information and fails to highlight other important aspects of a study.		
Denis Lacombe EORTC	9	Table 3.1 General	The table is confusing as it seems to describe the characteristics of interventional studies but covers different types of studies i.e. by definition a single arm trial will not be randomized. It is suggested to refer to this table rather as definitions of terminologies used for interventional studies. The description of some of the forms and methods of data collection appear to be unprecise and should be aligned with the common understanding		Already answered.
Tanja Podkonjak, Takeda	9	264-265	Current text: Observational studies can be either descriptive, that is, without a control group (case-series and cross-sectional studies) or analytical (case-control and cohort studies) with a control group. Proposed text: Observational studies can be either descriptive for hypthesis generation, which is commonly used in case series and cross-sectional studies, or analytical for hypothesis testing with a control group which is generally used in case-control and cohort studies. Rationale: The main difference between descriptive studies vs analytical studies are whether it is for generating hypotheses or testing hypothese. It would be more clear to include this. [1] [1] https://www.cdc.gov/globalhealth/healthprotection/fetp/training_modules/19/desc-and-analytic-studies_ppt_final_09252013.pdf		We maintain the original proposal.
Tanja	9	272-275	It could be argued that both prospective and retrospective studies		We maintain the original proposal.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Podkonjak, Takeda			measure the exposure prior to the outcome. However, in prospective studies, the data on exposure is collected prior to the occurrence of the outcome, whereas in a retrospective study the data on both exposure and the outcome occurs after the occurrence of the outcome. In both types of studies, the association between the exposure and outcome is assessed.		
Tanja Podkonjak, Takeda	9	276-280	Current text: Retrospective data are usually collected from existing data sources. Thus, retrospective studies can be quicker to complete compared with prospective studies but are limited by the availability of the existing data." Proposed text: "Retrospective data are usually collected from existing data sources. Thus, retrospective studies can be quicker to complete compared with prospective studies but are limited by the availability of the existing data. For example. data elements needed to answer a research question may be missing and may lead to biases such as misclassification of outcome or exposure or inability to completely control for confounding." Rationale: The current wording seems vague. The additional sentence is added to provide more detail to the limitation of data availability in existing sources.		We do not think this example improve the quality of this section.
	9	257/Table 3.1	The success of a random allocation depends on both the randomisation method and the allocation sequence concealment, i.e., prevent participants and trial team members from knowing the forthcoming allocations until the participant assignment. Concealment here differs from blinding that consists of continue the masking of assigned intervention after randomisation. So, it would be interesting to add "Concealed" and its definition to Table 3.1 after information on "Randomised".		We agree on the high value of allocation concealment for correct randomization. Table 3.1, however, only shows the basic concepts. The RoB-2 tool is already recommended in the present guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Relevant reference: *Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng HY, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898. doi: 10.1136/bmj.l4898.*		
EFPSI	9	257	Definition of crossover studies "Comparison of two (or more) interventions in which patients are switched to the alternative treatment after a specified period (therefore, each patient receives each treatment) Patients get different interventions, not necessarily all treatments (example 2 out of 3 treatments) Suggest to remove "(therefore, each patient receives each treatment)" or change accordingly		It is correct that in the very special case of a three-group, two-phase cross-over trial study participants receive only 2 of 3 treatments. We will consider adding the words "in most cases" to the words between brackets.
EFPSI	9	257, Table 3.1	In Table 3.1, 4 clinical trial designs are defined. This may be understood that this guidance only considers those 4 types. Could the agency provide their views on other clinical trial designs or provide a reference where in the guidance documents they are discussed?		We do not understand what would be the other designs? In our view, other designs (e.g. controlled before-after designs) are not very relevant in the context of HTA.
MTE	9	260-262	The distinction made here is not always the case. New interventions are often subject to observational study designs (e.g., in the form of registries) as a requirement before their routine use. Please consider altering this sentence.		This is covered by 'this suggests'
Ermisch – GKV-SV	9	271	The statement on the ethics or feasibility of randomisation should be objectified be deleting the word "deemed", which leaves the question open, whose opinion would be decisive.		See general comment.
Karen Facey	9	Table 3.1	This table of study characteristics is rather old fashioned given the more	X	Timing of analysis is out of scope.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			complex designs we see coming through, particularly for oncology (e.g. adaptive designs, basket trials etc). These are described in section 5, but seem rather to be an add on, as opposed to the heart of the guideline. So in the table, the design row could be augmented and Section 5 on particularities could be brought up to appear before the current section 4. In terms of study characteristics, the implications of the timing of the analysis (and completeness of follow-up) are often a challenge for HTA of the technologies that will be first assessed in the HTAR. So there could be another row in the table that states something like: **Timing of Analysis:** **All planned patients: Data collection until a fixed number of patients have been followed up for a specified period, or experienced a certain event.** **Interim: Trial has been stopped at an interim analysis or a data cut-off applied with further data collection ongoing**.		
Karen Facey	9	Section 3.1.2	A lot of text is given to the description of observational studies. This could be shortened to a few sentences as it is a standard description and for most HTAs interventional studies will be the key source of relative effectiveness evidence.	X	We prefer to keep as it was proposed.
Roche	9 Figure 3.1	322 / 3.2	The figure is too simplistic. According to ICH E10 there are different types of control, external control being one of them, but this does not seem to be taken into account here. Proposal: The figure should be updated to reflect the ICH E10 addendum. This intends that the guidance should cover external controls and consider it within this figure as in ICH E10 it is clear that external data sets are a type of control arm.		External control is out of scope.
Silke Walleser Autiero	9	260-262	The distinction made here is not always the case. New interventions are often subject to observational study designs (e.g., in the form of		Duplicate

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Medtronic			registries) as a requirement before their routine use. Please consider altering this sentence.		
Matias Olsen, EUCOPE	9-10	276-281	Stating that the causality is unfeasible with retrospective data is excessive. For example, looking at the relationship between dose and adverse event may be very well documented in some specific databases where the analysis is retrospective. For example, excess mortality associated to the diet treatment from Mediator has been proven within retrospective study design using databases. Proving it in a randomized double-blind trial would lead to fatalities and a very long time to answer the research question. It is important to link the design to the research question and ensure the study design is optimal to answer the research question.		We did not state that 'causality is unfeasible with retrospective data' but 'In that case, the fundamental assumption that cause precedes effect can be violated, which implies that the study of causality between exposure and outcome of interest is unfeasible'.
Thomas Kanga-Tona, International Association of Mutual Benefit Societies - AIM	General		We welcome the practical guideline document's comprehensiveness. We also welcome that randomised clinical trials are recognised as the gold standard design for evaluating causal relationships between interventions and outcomes. We also welcome the fact that EUnetHTA explored a number of various study designs, particularities and discusses them. It is also very helpful that study designs that do not help generate appropriate data on treatments' therapeutic added value/relative effectiveness assessment (e.g. uncontrolled trials – 4.3) are addressed. It should be a requirement that health technology developers do not bring to HTA bodies evidence based on those clinical study designs that are irrelevant in the frame of REA.		Thank you. Details regarding the submission dossier by HTD is out of scope of this guideline and will be addressed in another guideline.
Daniel Widmer UEMO	general	3	Importance of the question of biases for internal validity and random errors for statistical precision. Those statistical technical questions belong to the general culture of GPs used in their activity of Continuing medical	x	We perfectly understand your point and the daily 'PICO mismatch'. However, this guideline is intended for assessor and co-

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			education (for example journal clubs for judging medical literature). The big topic for GPs is the applicability. We are living every day the "PICO mismatch" described in the paper, with our patient's population of aged multimorbid persons so different to the studies population We regret that in your chapter 3 (Clinical study design) nothing is said about qualitative research or mixed method. From our point of view these methods are essential for applicability. Ref. Pope C. Qualitative research in health care. BMJ Books 4th ed. Wiley-Blackwell 2020. And, Lebeau JP. Initiation à la recherche qualitative en santé. CNGE 2021 Excepting this comment the paper is very good.		assessor, and continuous medical education is a specific approach which is out of scope of this guideline. The same holds true for qualitative research.
Matias Olsen, EUCOPE	general		The assessment of the additional benefit of a drug has been segmented over a series of independent guidelines, which are interrelated. However, in this practical guideline there is a lack of harmonisation and consolidation that hinders the full appreciation of the relative effectiveness, which is a global assessment of one intervention over one or more reference intervention(s). In real life, there is a fine balance and trade-off between the internal validity and the applicability. Applicability should be restricted to external validity and potentially generalisability, which are overlapping concepts. The highest the internal validity, the lowest the external validity and the lowest the generalisability. Increasing the internal validity can reduce the external validity and vice versa. This is in part why all regulators aim to maximize the internal validity of clinical studies. The current guideline specifies that the medical context should not be considered when assessing the relative effectiveness (for example rarity or impossibility of blinding). This clearly implies that products based on single-arm studies will be concluded as not assessable, or assessments will conclude that there is no evidence of any relative effectiveness benefit over current available SoC. It is important to note that a vast majority of rare		See also main themes 1) and 2) at the beginning of the document. We believe that applicability could not be restricted to external validity and generalisability only. The JCA report should also discuss how the submitted evidence matches the PICO question. Again, we reaffirm that individual non-comparative studies (for example single arm study) do not allow relative effect assessment and are therefore of limited value. We also remind that evidence synthesis (such as indirect comparison using single arm trial data) is out of scope of this guideline, and not concerned by it. For applicability, it could vary between MS, due to different PICO question or

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			disease interventions and ATMPs are approved with single-arm trials and most of them recommended by HTA agencies. Under this new rule, although it may be inappropriate or not feasible to undertake a randomised control trial, single-arm trials would be systematically eliminated. This will lead to relevant evidence not being considered in the evaluation process and further delay patient access. Furthermore, the assessment of applicability and the clinical relevance of an effect size is left to be judged at the member state level without further methodological recommendations. The JCA needs to provide a final assessment for the agreed PICOs and discuss variability when applicable. This cannot be left open for the member states to decide. It has to be clearly discussed as part of this guideline within the JC. The EU HTA outcome has to be deliberative and transparent.		different healthcare setting. Therefore, a common judgment cannot be done at EU level, and should be let a national level.
Tanja Podkonjak, Takeda	General		A formatting error appears on the footnote throughout the document which states: Error! No text of specified style in document.	X	It will be corrected for the next version.
Tanja Podkonjak, Takeda	General		**Observational studies and RWE/RWD for causal inference, the distinguished role of observational IPD compared to RCTs** The guidance states the causality or causation several times in observational studies such as retrospective vs prospective, cohort studies, case-control, etc. It is known that observational studies based on real-world data may suffer from different bias, such as confounding, missing data, and misclassification [1]. Given the different nature of RCTs and observational studies, *a recommendation is needed for when and how observational studies could be used for causal inference*. The guidance mentions 'Given that observational studies are performed based on routine healthcare, this suggests that they allow the		This guideline is intended for assessor and co-assessor, when assessing the submitted data which answer to the PICO question. It provides recommendations for assessment, depending on study design. Define and list situations when RCT or observational studies are preferred/recommended is out of scope of this guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			assessment of relative effectiveness of only those interventions that are already used in medical practice, rather than of new ones.' Therefore, the role of observational studies is different from RCTs. However, the guidance does not discuss how observational IPD data could inform treatment effectiveness. We would like *to request a clear demonstration of the best applications of observational studies, potential biases, and recommended methodologies be added to the guidance.* [1]: Groenworld RH. Trial Emulation and Real-World Evidence. JAMA Network Open 2021; 4(3)		
Tanja Podkonjak, Takeda	General		**Single arm trials** The guidance states that "In the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness". Firstly, this statement is a value judgement which according to the HTA Regulation (HTAR), is left to the autonomy of each member state to decide. Secondly, uncontrolled clinical trials (i.e. 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. Therefore, single arm trials are useful in under-recognized patient populations, despite the absence of internal controls [1]. It is likely that some technologies in the first two phases of the HTAR implementation, oncology, ATMPs and rare diseases will only have single arm trial data available due to the constraints mentioned above. Therefore, it is imperative guidance be provided on an EU HTA level on how to analyse this data, including the RoB assessment tools, expected in line with the other study design methods described in this guidance. We agree,		We believe that stating that an uncontrolled clinical trial are of limited value to meet the comparative/relative effects assessment required from the HTAR (Article 2 and 9) is not a value judgment, but a description of the limitations of the available evidence, required by Article 9. The acceptance (or not) and the consequence of this limitation for national decision is left to the autonomy of each MS. Ethical issues or challenges are out of scope of this guideline, and are MS dependant. We remind that this guideline only concerns individual studies, and not

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			limitations of this data should be described in the submission dossier, however, without any guidance on how to present single arm data (as appears in the draft D4.6 guidance), HTD will present the evidence based on their own best judgement which will lead to inconsistencies. Therefore, *a recommendation of the role of single arm trials and how to analyse them in the context of an HTA is needed in this guidance.* [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		evidence synthesis, such as indirect comparisons, which are covered in two other guidelines.
EFPSI	General		This guidance, and the JCA process in general, asks for a lot of rigor in terms of the evidence that is needed in order to be successful, while at the same time being extremely open in terms of the number of questions/PICOs to be addressed. The guidance should acknowledge that drugs or devices that are beneficial to patients may fail to answer all possible questions because all PICOs will not be known to the HTD until after the trial has already concluded and read out. Thus we recommend that the HTD be involved and have the possibility to understand the questions earlier in the process in order to be able to provide the best possible evidence. In addition, JCA should be more open to innovative methods.		Discussion regarding the scoping process is out of scope and is addressed in the 4.2 'Scoping Process' guideline.
MTE	general	-	We welcome to obtain a comprehensive view on the validity of studies and clinical (including real world) data and look forward to a version without the current indicated restricted scope.		See other replies
MTE	general		**Role of non-randomised studies (NRS) and observational studies in JCA**		Duplicate from EFPIA comment

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			We believe that the guideline, in its current version, is a missed opportunity to recognise the critical role that NRS and observational studies play in addressing the challenges of evidence generation for some health technologies. In particular it will be important to account for : • MedTech, A classable double blinding may not be feasible, open label, pragmatic RCT might be a solutions We would like to emphasise that the methods applied need to be relevant for both the initial and any potential updates of JCAs (as per Article 14 of the EU Regulation 2021/2282). Particularly in such updates, NRS and observational studies might play an important role since it will reflect real world use (and the relatiev effectiveness of a given technology. The EU Regulation 2021/2282 acknowledges that for some new health technologies some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: "*Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. .....*".		
MTE	general	-	As per the previous guidance document, D4.6 reads less like a guideline and more like a glossary of terms and study taxonomy (although more work needs to be done here on classification of all study types described in this document). Consider also the addition of guidance on external validity and statistical precision beyond referring this step back to the responsibility of individual member states.		We consider the appraisal of the degree of statistical precision and some elements of applicability are MS dependant and therefore need to be left at the national level.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
MTE	general		**Value Judgements** It is important that the guideline does not inadvertently impose what could be a value judgement around validity considerations, including suggesting effect sizes, validity thresholds, and stating that some studies have limited value. Such judgements may mean that relevant evidence to some Member States is excluded from the JCA report. As stated in the 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.*'		Duplicate from EFPIA comment.
Marjorie Morrison, Lymphoma Coalition	General		**<u>Trial Design(s)</u>** While the guideline provides comprehensive, if not overly complex terminology and presentation for non-clinicians, descriptors of trial designs (clinical study designs, observational studies, and others), there are perceived gaps that may be of interest to note. For instance: when considering prospective and retrospective studies, it is feasible that variable factors such as **budget and demand on resources may be key considerations that present as barriers and/or challenges** despite clear advantages and/or disadvantages between methods. While this is also applicable to other study designs, economic considerations are of particular relevance in relation to these studies given that costs between the two will differ, with one being most costly than the other based on study techniques and other key factors. While it is important to address the differing trial design options, it is		Economic considerations and data collection process are not part of the HTA-R and are therefore out of scope of this guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			beneficial to concurrently appreciate the limitations of clinical designs in practice. Further, in relation to retrospective studies, the use of current pre-existing data are critical to determining conclusions regarding the outcome however, the **persistent lack of comprehensive data, data that is exclusionary of specific patient characteristics, or insufficient data collected or reported from registries or other classical data sources** are all of concern. Additionally, **data collection** – or rather, the processes for data analysis that differ between methods – is critical in determining methods, given that (for instance) prospective studies have a future-oriented focus while retrospective studies aim to expand on existing research, information or discovery. Thus, as the aforementioned is of relevance to the lymphoma community in relation to indolent given that therapy for incurable lymphomas are likely to be over time with the risk of adverse effects occuring that impact quality-of-life issues, these are key considerations in relation to trial design(s).		
EHA	General		We recommend earlier involvement of health care experts and patients regarding the individual appropriateness of study designs in context of very rare diseases. In line with recent health data discussions, including with EMA in the CMF, I development of novel/more appropriate external data comparators should be reviewed on a regular basis for their potential appropriateness		Involvement of clinical experts and patients are out of scope of this guideline and are addressed in a specific guideline. Choice for comparator(s) belongs to the MS only (see 4.2 'Scoping process' guideline) and are out of scope of this guideline.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			in the future. Especially EHA and other public efforts around accessible data repositories for well-annotated randomized controlled trial data could allow for comparator data with lower uncertainty than many current real-world evidence databases. There should be a cross-reference to sub-populations (mentioned in 4.5 for PICO) as potential rare or very rare entities. Some of these have very high unmet need (e.g. del(17p) AML) in context of an otherwise curable disease (large heterogeneity in disease outcomes) and could again be defined using consultation with healthcare professionals and patients. Again, a cross-reference to IVDR would be desirable, to ensure consistent identification of such high unmet need, very rare subgroups. Minor comments: On line 141, the sentence ' 117. Furthermore, the certainty of results is independent of the medical context of the PICO question.' does not seem entirely true. A study with few patients on a very rare disease will by definition have more uncertain results than a study on 40000 hypertensive patients. Same for a gene therapy tested on a very small population. Line 176: I think the sentence regarding external validity is wrong:		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			164. 'The key question is how well the evidence matches the elements of the PICO question' I suppose the authors mean that the study PICO matches the PICO relevant to stakeholders in a given country. Regarding the case controlled studies, the document also mentions (line 418) that they are not suited for more than 1 outcomes (eg survival I suppose) and this might also be problematic for rare hematological diseases.		
Dr. Thomas Ecker, Ecker + Ecker GmbH	general		Ecker + Ecker GmbH, a healthcare consultancy based in Germany with strong expertise in the early benefit assessment, welcomes the establishment of a European Health Technology Assessment (HTA) fostering closer cooperation between member states on health technology assessment by introducing a permanent framework for this joint work. The legal requirements for a European HTA have been determined as a legislative act by the end of 2021 with the EU regulation 2021/2282. From 2025, before placing innovative medicinal products on the market, oncology products and ATMP are subject to a European joint clinical assessment. In the next step, Orphan Medicinal Products (OMPs) will follow beginning in 2028 and from 2030, all medicinal products will have to go through the European assessment. While the regulation does not come into force until 2025, the process of implementation is already ongoing to ensure effective application from January 2025 onwards. At present, the development of a methodology for joint HTA work is facilitated by the European Network for Health		Ok.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Technology Assessment (EUnetHTA) 21 consortium.		
Dr. Thomas Ecker, Ecker + Ecker GmbH	general		Even if the scope of the guideline is the definition, classification and evaluation of certainty of study results, we would like to point out the necessity of clear guidance on the scope of evidence to be presented. Clear guidelines are required to ensure that all necessary data are presented.		Data presentation is out of scope of this guideline.
Prof. Matthias P. Schönermark, M.D., Ph.D. and Dr. Ingo Hantke SKC Beratungs-gesellschaft mbH	general		Comment: As expected, the sub-deliverable "D4.6 validity on clinical studies" does not offer any major surprises when providing insight into the clinical trial requirements. However, it is very unfortunate and not acceptable that the rare diseases and the respective (unmanageable) challenges in generating evidence are not considered in the entire document. In addition, potential ethical challenges in paediatric populations or life-threatening diseases are not considered. We recommend adding respective context into the relevant sections to consider the individual situation in the indication.		As previously answered, ethical consideration is MS dependant, and are out of scope of this guideline. We reaffirm (line 140-143) that certainty of result is independent of the medical context. Medical context could be used for national appraisal, but is out of scope of this guideline, and should be left at MS level.
Sebastian Werner vfa	General		This practical guideline is dedicated to the definition, classification, and certainty of results of studies leading to the statistical analysis of what is considered one data set. The Guideline provides a basic method frame for the assessment of validity and certainty in joint clinical assessments. For instance, the guidance recommends for internal validity distinguishing between different study designs. For the assessment of certainty, the guidance recommends standard study design-specific tools, i.e., the Cochrane Risk-Of-Bias-Tool for randomized controlled trials (ROB-2) and the Cochrane Risk-Of-Bias-Tool for non- randomized evidence (ROBINS-I). The provision of a basic method frame for validity and of certainty can promote harmonization of clinical assessments. However, many other suitable methodological aspects were excluded from incorporation into		See also main themes 1) and 2) at the beginning of the document. Single arm interventional studies with historical controls (indirect comparison) are out of scope of this guideline. They are covered in the guideline 4.3 'Direct and indirect comparison' guideline. Applicability (excluding mismatch between PICO and submitted evidence) is MS dependant and therefore need to be let to the national level. The same apply for appraising effect sizes.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			the frame. For instance, single-arm interventional studies using external ("historic") controls were not included. Further, the assessment of applicability (primarily any PICO mismatch between assessment scope and clinical study) is left to the member states, without recommending guiding principles. Similarly, the clinical relevance of an effect size, is left to be judged on the member state level without further methodological recommendations. The vfa recommends forming guidance on the assessment of one-arm interventional studies using external ("historic") controls, applicability, and clinical relevance. Single-arm interventional studies with historic controls are very valuable for new drug development on rare disease, highly targeted patient populations and in other special therapeutic situations with high medical need, when randomised controlled trails are not ethical or feasible. Methodological guidance for the assessment of the degree of uncertainty is needed. Guiding principles should be applied to the assessment of applicability. Applicability (transferability or external validity) should be generally assumed as long as no evidence-based medical reasons are known that argue against transferability. Guiding principles should be applied to the assessment the clinical relevance of effect size. The assessment of clinical relevance should consider valid and established thresholds or responds-levels. The vfa recommends establishing a harmonized European methodological framework for joint clinical assessment that provides a common methodological approach for data analysis and synthesis that member states accept and apply.		

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Sebastian Werner vfa	General		The guidance claims that in the context of HTA, uncontrolled clinical trials are of very limited value for estimating treatment effectiveness. Further, the guidance argues that it is deemed unnecessary to propose any formal rules for assessing Risk of Bias of single-arm trials, although some tools have been developed already in the past. The Guidance claims that the Risk of Bias would be very unlikely to be changed by formal assessment and thus this work would be dispensable. The vfa does not agree with these statements. Single arm studies are very valuable for new drug development on rare disease, highly targeted patient populations and in other special therapeutic situations with high medical need. In such situations randomised controlled trails might not be ethical or feasible. Single arm studies can be enhanced with an external control group (historic control). The external control groups might be comprised of historic interventional or observational studies, data from patient registries or other real world data sources. Single arm studies are often basis for regulatory approvals with valuable insights on effectiveness and safety. They ensure faster approval and faster patient access. Excluding single arm studies (with historic controls) as an "dispensable" part of the guidance foils regulatory approval and the objectives of the European HTA regulation to foster faster patient access. The claims set out in this draft guidance are stringent and unfeasible for many new technologies, particularly those included in the first waves of joint clinical assessment (Orphan Drugs, Oncology and ATMPs with high medical need). The vfa recommends forming elaborated guidance on the assessment of single-arm interventional studies using external ("historic") controls. The guidance should propose formal rules for assessing Risk of Bias of single-arm studies and of external controls, including factors that influence the degree of uncertainty. The guidance should consider that established		See previous answer.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			confounder adjustment methods such as Propensity Score Matching, or Matching Adjusted Indirect Comparisons can increase the certainty of results of the analyses. The joint clinical assessment shall be based on the best available evidence.		
Sebastian Werner vfa	General		The guidance claims that the certainty of results is independent of the medical context of the PICO question. The guideline argues that it would be methodologically inappropriate, for example, to take the rareness of a disease or the impossibility of blinding as an excuse to ignore or to euphemise the resulting uncertainties in the clinical evidence. The vfa does not agree with these claims. The vfa is convinced that the formal assessment of the degree of certainty should consider the medical context and must be adapted to consider the specificities of special therapeutic situations. The vfa also believes that these claims made in the guidance violate the intentions of the European HTA Regulation. The EU Regulation 2021/2282 acknowledges that for some new health technologies some data may not be available and new methods will be needed. The EU HTA Regulation explicitly states that: *"Methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to include specificities of new health technologies for which some data may not be readily available. This may be the case for, inter alia, orphan medicinal products, vaccines and advanced therapy medicinal products"*. That indicates that methodologies for performing joint clinical assessments and joint scientific consultations should be adapted to consider the medical context. In special therapeutic situations, such as rare diseases, high medical need areas, targeted patient groups, or paediatric patients, randomised controlled trails are often not ethical or feasible. Therefore, it is not appropriate to evaluate these situations according to the same standards applied in regular therapeutic situations in which randomized controlled		We reaffirm that certainty of results is independent of the medical context. However, we agree that medical context could be used for national appraisal, but is out of scope of this guideline, and should be left at MS level.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			trials are possible. The uncertainties of the clinical evidence in special situations should be treated more lenient than in regular situations where RCTs are possible. For that, the methods must be adapted to consider the specificities of the medical context, especially with respect to rare diseases, high medical need areas, targeted patient groups, or paediatric patients. The vfa recommends considering the medical context in the formal assessment of the degree of certainty. The methods must be adapted to consider the specificities of the therapeutic situations. Uncertainties of the clinical evidence in special situations should be treated more lenient than in regular situations. For instance, in the case of rare diseases, the probability of observed results should be meaningfully defined, considering the small number of patients by increasing the necessary p-value.		
Sebastian Werner vfa	General		It is important that the guideline does not inadvertently impose what could be a value judgement around validity considerations, including suggesting effect sizes, validity thresholds, and stating that some studies have limited value. Such judgements may mean that relevant evidence to some Member States is excluded from the joint clinical assessment report. As stated in the 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.'*		Already adressed issue.
Sebastian Werner vfa	General		In general, all sources of available evidence should be acknowledged. The corresponding evidence level, study type and RoB of the sources should be considered. Instead of complete rejection of existing evidence due to violation of		Already addressed issues.

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			strict criteria the general approach should be to use the best available evidence. Complete exclusion of existing evidence introduces another risk of selection bias and should therefore be avoided in the context of an evidence-based approach.		
Ermisch – GKV-SV	General	n/a	We support the general considerations regarding the different dimensions of the certainty of evidence and that favourable assessments in one dimension cannot substitute the lack of certainty in other dimensions. In particular, we support the notion that internal validity/relative freedom from risk of bias is the cornerstone and foundation of the certainty assessment, based on which applicability to the PICO question and safeguarding against random error are additionally required to come to a reliable conclusion regarding effectiveness. While the guideline gives valuable information and good advice at large, it partially loses its focus. One would expect to find advice on the validity of endpoints in the respective guideline, not within a guideline on validity of clinical studies as such.		Thank you for the comment. But, we have troubles to understand the last comment as it seems the guideline is focused on the validity of clinical studies and not on the validity of endpoints.
Bayer	general		The document excludes innovative fusion designs such as single arm studies/RCTs combined with external control arms, potentially sourced from real-world or historical data, or transportability analyses (transporting internally valid RCT inferences to a "real-world" population for external validity). The text suggests that these novel research designs would be classed as "evidence synthesis" because they borrow external data. We suggest reconsidering this position.		Studies that are single-arm trials coupled with a comparison using an external source of data are covered in the D4.3 guidelines on direct and indirect comparison. We understand this choice could be debatable, but we have decided they fall under the general case of evidence synthesis in the presence of disconnected networks.
Bayer	general		Generally, the document considers external validity to be less relevant than internal validity. This may be the case in the regulatory context, where relative efficacy is estimated in a controlled setting with a highly selected sample/ population. Conversely, the focus in HTA is on estimating "real-world" relative effectiveness. Therefore, external validity		What we meant by "gold standard" if the fact that RCTs are the simplest and most consensual design for allowing counterfactual reasoning. We will consider a better way of conveying this idea with

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			is equally as important as internal validity in the HTA context. External validity cannot be overlooked because it is instrumental to assess the relevance of results to the PICO research question. Assessment of "gold standards" in the evidence hierarchy is a function of both internal and external validity. Both dimensions should be included in discussions about "validity". A well-conducted RCT may have high internal validity but low external validity, if it is carried out in a sample that is not representative of the target population of policy interest. A pragmatic trial may have low internal validity, e.g., due to low adherence and high drop-out rates, but high external validity, e.g., by targeting a population that is representative of routine clinical practice through broad eligibility criteria.		less ambiguity in the next version of the draft. Regarding "external validity", as stated in the document, some elements are left at the discretion of the MS. Finally, we do not think the last example proposed here is sound. Indeed, if a study has low internal validity, it prevails on its external validity, as estimates are biased.
Karen Facey (Individual)	General	Overarching	This guideline explains the concepts related to critical assessment of clinical studies, but it feels as though it is something that was written 20 years ago when full HTAs were undertaken and clinical research only involved double-blind RCTs, with section 5 added to bring the guideline up to-date. However, in section 5, particularly in relation to RWE studies, the focus on evaluation of treatment effectiveness is lost and there is a restatement of an earlier recommendation that single arm studies are not relevant. So this guideline might be appropriate for health technologies that have a strong evidence base, such as medicines for larger populations. It will be harder to apply to high risk medical devices and the first tranche of medicines that will be assessed according to the HTAR (in oncology, ATMPs and then OMPs) as these products are being developed with more novel clinical trial designs, including single arm studies. . It would be helpful to test this guideline against a few recent examples of these kinds of medicinal products, to see how it manages the paucity of evidence that is typically the case for HTA given accelerated regulatory		According to clinical trials.gov, a lot of double-blind RCTs are still undertaken currently, so we do not think it's fair to imply these are designs of the past. Moreover, it's still a scientific consensus that principles of RCT designs are the simplest for allowing proper counterfactual reasoning, so it would be disturbing to not propose them as the cornerstone of comparative effectiveness assessment. This guideline is intended for assessors and co-assessors in helping them for reporting the necessary elements about the certainty of results. The approach of this guideline is to be based on scientific consensus and not on a best available

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			approvals.		evidence approach. We reaffirm single-arm studies are of limited value for relative effectiveness as they do not provide any form of comparison. This is factual and not a judgment. It does not mean single-arm trials will never have any weight into a JCA, but there is no need to describe their certainty of results using a standardized tool (even more as such highly consensual tool currently does not exist). In addition, we find it troubling the comment seems to consider single arm trials as "novel" designs.
Roche	General		Can the guidance include a table that lists all types of studies covered within the guidance and the associated RoB measure to be used?		As only two RoB instruments are mentioned (one for RCTs, one for NRS), a table appears unnecessary.
Roche	General		This document should provide additional guidance on the harmonization of evidence required at the pan European level and reflect a more centralized perspective. The current guideline provides many statements that reference the decentralized decision making at the member state level in regards to the acceptability of evidence. We suggest that the guidelines apply a broader, more centralized understanding of the evidence and flexibility on the methods should be included within the guideline.		As clearly stated in the HTAR, the JCA is based on the chosen parameters during the assessment scope (Article 9). As also clearly stated in the Article 8 (6): 'the assessment shall be inclusive and reflect Member States' needs in terms of parameters and of the information, data, analysis and other evidence to be submitted by the health technology developer'. Considering the above,

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
					assessment scope is based on MS needs, and different MS needs could exist. There is therefore no limitation or requirement for a harmonized methodology.
Roche	General		This guideline should include a reference to the ICH E9 (R1) Addendum on Estimands and Sensitivity Analyses. The Addendum allows for different treatment effects to be considered which may be more clinically meaningful and more relevant to patients and other stakeholders which is an important consideration for the evaluation of clinical validity. For the purpose of the JCA, there should be no prejudice in accepting and no downgrading of the evidence from RCTs that estimate treatment effects on the basis of strategies alternative to the ITT, if these strategies are suitable to address the research question of interest. The guidelines should acknowledge that there is a place where the highest methodological standards are applicable and there is a place where more methodological uncertainty is necessary (including possible statistical uncertainty).		We agree with this comment. But, sensitivity analyses are addressed in a specific guideline D4.5 in separate chapter. These chapters are written with the estimand framework in mind. Neither the D4.5 nor the current guideline implies an ITT strategy should always be the best or unique one to use for answering a PICO question.
Roche	General		The guideline should provide stronger language around the acceptance of studies (e.g. non-randomized studies (NRS), uncontrolled trials, observational studies) in small populations (e.g., rare diseases) for the case where RCTs would be unethical or not the most appropriate approach. For orphan drug development, clinical development is shaped by issues such as low disease prevalence, disease severity, small and heterogeneous patient populations, difficulties in patient recruitment, and limited knowledge of the natural history of disease, among others[1]. As a		The guideline is intended for assessors and co-assessors for helping them in reporting the necessary elements to assess regarding the certainty of results of the submitted evidence by HTDs. We reaffirm this assessment should be independent of the medical context. Nonetheless, the medical context could be

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			consequence, the design and analysis of clinical trials for these diseases becomes more difficult. Additionally, the EMA/CHMP guideline[2] suggests avoiding unnecessary clinical trials e.g., by extrapolation from a larger source population to a smaller target population when this is appropriate. The guideline should clarify the role of "underpowered" RCTs that can still generate valuable evidence, particularly when complemented with NRS and/or observational studies. 1. CHMP. Guideline on clinical trials in small populations. [Online] 2007. [Cited: February 1, 2013.] www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf. 2. Hee, S.W., Willis, A., Tudur Smith, C. et al. 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 12, 44 (2017). https://doi.org/10.1186/s13023-017-0597-1		considered by MS when they appraise the clinical added value of a health technology, including the difficulty (ethically or practically) in providing evidence such as RCTs. A list of specific situations where RCTs cannot be de facto conducted is out of the scope of this guideline. It will be appraised on a case-by-case basis at a national level. Moreover, there is no consensus in the scientific literature about such a list. For example, meta-epidemiological studies show RCTs are frequently conducted event in the context of rare diseases, even against placebo.
Roche	General		Joint scientific consultation (JSC) is a crucial component of the EU HTA framework and provides an excellent basis for the discussion on the validity of clinical evidence across the various stakeholders and decision makers (e.g. EMA, HTA, HTD, clinicians, patients). As such JSCs should be embedded into the guideline. The JSC should establish agreement between HTD, EMA, and HTA on the acceptability of the type of evidence and the totality of evidence required for a product/device.		We agree JSCs are important opportunities But are out of scope of this guideline (see dedicated EUnetHTA documents).
Silke Walleser Autiero Medtronic	general	-	As per D4.5, the titles and text in the boxes and how this information relates to the previous section is not clear in each instance. Consider clearer heading titles for each BOX that are more prescriptive and provide actual guidance on minimum evidence/data/analyses requirements for JCA authors.		We will consider if more clarity is needed regarding the box for the next version of the draft.
Silke Walleser	general	-	As per the previous guidance document, D4.6 reads less like a		Boxes are provided indicating what

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Autiero Medtronic			guideline and more like a glossary of terms and study taxonomy (although more work needs to be done here on classification of all study types described in this document). Consider also the addition of guidance on external validity and statistical precision beyond referring this step back to the responsibility of individual member states.		assessors and co-assessors need to report in the JCA. We reaffirm the appraisal of the degree of statistical precision and the appraisal of some elements of applicability is dependent of the national context and must be lest at the discretion of each MS.
Intuitive	General		Intuitive would like to raise the concern that the validity and applicability of clinical studies associated with medical devices may vary depending on the context of the health technology being reviewed. For medical devices, guidance should allow for flexibility in determining the type of evidence and studies that are used for decision making across different types of technologies. Generally speaking, medical devices face unique challenges associated with evidence generation when compared to pharmaceuticals. The delivery of care associated with complex devices and their corresponding outcomes are often impacted by differing levels of user and care team expertise and experience, the quality and availability of pre- and post treatment care, and the broader context of the institution or health system in which the technology is used, among other factors. As a result, we believe that the traditional evidence hierarchy should not be a barrier for determining the validity of clinical studies. In our past experience, RCTs are often prioritized for complex devices, even when biased or methodologically weak, simply because of their position within the evidence hierarchy. We believe this guidance should identify this challenge and explicitly allow for the consideration of other study designs or forms of evidence, which may have a more accurate representation of the technologies impact in real life clinical settings.		This guideline is intended for assessors and co-assessors in helping them for reporting the necessary elements about the certainty of results. The approach of this guideline is to be based on scientific consensus and not on a best available evidence approach. It does not imply the results of studies other than RCTs will never be considered. We reaffirm certainty of results can be factually assessed independent of the medical context. But, appraisal at the national level could use medical context into pondering the final decision in appraisal the clinical added value of a health technology. In addition, it's unfair to imply we consider the classical hierarchy of evidence as absolute. As stated in line 150-151

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
			Intuitive would like to reiterate that real-world studies are increasingly recognized as a valuable source of decision grade evidence provided the data is robustly analysed and elaborated. There has been a tremendous amount of work to develop recommendations and best practices for generating and reporting this type of data that mitigates many methodological concerns. HTA agencies such as NICE and HAS have developed recent RWE frameworks and published work related to demonstration projects and methods applicable to RWE studies. A holistic perspective should be used when evaluating the evidence base for a given technology, which includes both randomized and non-randomized evidence. In some scenarios, certain study designs may be more applicable than others in aiding a health technologies clinical evaluation. A "best evidence" approach should be applied to ensure that the best quality evidence is available for consideration, rather than potentially relying solely on poor quality RCTs simply based on adherence to the conventional evidence hierarchy. This is especially true when supplemental evidence and/or ongoing RWE can further aid decision making and ongoing evaluation.		"classification of study design alone is insufficient for a full assessment of internal validity".
Jasmine Toomey PHMR	NA	General	The guideline references many different documents which could be integrated to an appendix.		This will compromise the readability of the document.
Jasmine Toomey PHMR	NA	General	It needs to be clearer who performs the joint clinical assessments.		This is covered by other guidelines.
Thomas Kanga-Tona, International Association of	P 11	337	**Double-blind** randomised clinical trials: gold standard	x	We will consider the most appropriate title for the next version of the draft; the term "double blinding", however, was criticized for being ambiguous already 20 years ago

Comment from	Page number	Line/ section number	Comment and suggestion for rewording	Editorial comment?	HOG response
Mutual Benefit Societies - AIM					(https://pubmed.ncbi.nlm.nih.gov/11308438/). In our view, randomization is the key aspect of a trial.
Thomas Kanga-Tona, International Association of Mutual Benefit Societies - AIM	P 11	348 – 350	Nonetheless, depending on numerous factors, such as the quality of the design and conduct of the study, the certainty of results of a particular RCT can be questioned and biases can arise**, for instance in the case of single-arm randomised clinical trials – see below section 4.3** (39,40)	x	We do not think this is a sound proposition adding more clarity to the sentence, especially because single-arm trials cannot be randomized.
Thomas Kanga-Tona, International Association of Mutual Benefit Societies - AIM	P 8	212 - 213	Similarly, early stopping of clinical studies, deliberate extension of recruitment, and selective reporting of results all undermine the validity of study results (23) **and should be reported**.	x	The assessment of submitted evidence by HTDs is based on the assessment scope, not on study-level. Other guidelines such as the D4.5 guideline covers the issue of interim analyses and early stopping of a study.