JA2- WP7- SG 3 – **SAG and Public** consultation on the methodological guideline "Internal Validity non-randomised studies (NRS) on interventions"
2nd version of guideline

Number of comment	Page	Line	Comment	Character of comment "major"[1] "minor"[2] "linguistic"[3]	Author/Draft Group reply
1	All		It is clear that this guideline is only looking at internal validity of NRS (and not the assessment of external validity). It seems that the focus of the tool for internal validity only is to assess whether the NRS can be considered equivalent to a randomised trial (e.g. the low risk of bias defined in ACROBAT-NRSI is defined as "the study is comparable to a well-performed randomised trial") and not consider NRS in supporting randomised trials in a HTA report. However, the NRS is often undertaken to be able to provide evidence on the generalisability (assessed by the external validity) of the intervention of interest in HTA reports. Are these studies included in the scope of this guideline?	☒ major ☐ minor ☐ linguistic	It is correct that NRS can be useful to examine generalisability. Studies with high external validity (but poor internal validity) can provide information on how interventions are being used or on who tends to receive which intervention in practice, yet good internal validity is a prerequisite for a reliable estimate of the expected effects in clinical practice (effectiveness). As the guideline does not address external validity (or generalisability), this also means that instruments for assessing risk of bias (RoB) were not included if they focused on external validity.
2	All		The ACROBAT-NRSI tool is the recommended tool for the RoB assessment of NRS (4th recommendation). It was developed by the Cochrane Colloquium and their purpose is to assess the risk the bias of NRS for inclusion in their systematic reviews of interventions. It is not clear if the tool is appropriate for use to provide an assessment of a NRS for purposes of HTA decision making. The context of the guideline is to assess if the NRS can be a replacement for a randomised trial, which may not be the reason for including NRS as part of an HTA submission.	☒ major ☐ minor ☐ linguistic	The inclusion of NRS covered by this guideline is restricted to questions on inclusion of NRS to answer questions on relative efficacy or effectiveness of medical interventions. Internal validity is an overarching quality aspect of a clinical study. Therefore, the assessment of internal validity of studies included in a systematic review is independent of the user context, may it be an HTA, Cochrane or other type of systematic review.

[1] "major" indicates that a comment points to a highly relevant aspect and that the author / the draft group is expected to give a thorough answer
[2] "minor" means that a given comment does not necessarily have to be answered in a detailed manner
[3] "linguistic" labels problems with grammar, wording or comprehensibility

					The reasons for or against NRS inclusion to answer questions on efficacy or relative effectiveness will differ according to context, but including NRS and appraising RoB are two different steps. This is already stated several times in the guideline, but a sentence has been added to the summary to already clarify at this point any possible misunderstandings regarding the purpose of this guideline (please see comment on page 5)
3	All		It is unclear where pragmatic studies fit. These studies are randomized so should not be part of "NRS". However, NRS seem to be RCT only. Clarification is needed.	☒ major ☐ minor ☐ linguistic	A pragmatic study design should not be considered a design of its own. We agree with the statement that pragmatic studies should be randomized (see: Raymond et al.; J Clin Epidemiol 2014; 67: 1150-6). As such, the internal validity of pragmatic trials should be assessed using the same methods as applied for any other RCT. Thus, in our opinion no change in the current guideline is required.
4	General		The purpose of the SAG is to provide constructive input into the EUnetHTA methodological guidelines to help ensure their quality and credibility. Below are general and specific comments on the draft guideline for '*Internal validity on non-randomised studies (NRS) on interventions*'. As an overall summary of this guideline we are concerned that support of the specific tool (ACROBAT-NRSI) seems arbitrary and promotional, especially in light of one author of this guideline also being associated with this specific tool. The lack of justification for exclusion of some of the other tools and the lack of validation and testing of the proposed tool is additionally problematic. To maintain the credibility of the EUnetHTA methodological guidelines process, we therefore suggest that the current report should be subject to a rigorous reconsideration of (i) its scope,	☒ major ☐ minor ☐ linguistic	This comment raises general criticism but fails to identify any specific examples of RoB tools that were inappropriately described or evaluated in the EUnetHTA guideline. The current version of the guideline contains exact reasons justifying the exclusion of each of the other RoB tools. The comment also speculates that scientific conflicts of interests influenced the selection of the Cochrane tool. The guideline authors do not consider themselves as having a conflict of interest, because they did not participate in the development of the Cochrane tool. Only one of the guideline authors took part in piloting the tool – thereby testing the ease of use of this new tool. Nevertheless, the lack of data on reliability and validity is admittedly a problem of the Cochrane tool. Thus, a re-evaluation and reconsideration of the current

			(ii) the literature review and (iii) the quality of the process that led to the current recommendations as in its current form prior to release for public consultation.		guideline should take place within the next few years.
5	General		It is made clear that this guideline is only looking at internal validity of NRS and that assessment of external validity is out of scope. However, to assess the quality of NRS, the external validity is equally important as the NRS is often undertaken to be able to provide evidence on the generalisability of the intervention of interest. The focus of the tool for internal validity only is to assess if the NRS can be considered equivalent to a randomised trial. This is a very specific and, uncommon, use of NRS and if this is indeed the intent of this guideline, then this should be made more explicit. Additionally, from ACROBAT-NRSI, the low risk of bias is defined as "the study is comparable to a well-performed randomised trial", however, where the purpose of the study is to address generalisability then NRS would be expected to outperform studies conducted in a controlled clinical environment.	☒ major ☐ minor ☐ linguistic	We agree that the evaluation and discussion of external validity and the relative importance of internal and external validity is outside the scope of this guideline. We disagree with the argument that "external validity is equally important" as internal validity (Windeler, Z Evid Fortbild Qual Gesundhwes 2008; 102: 253-9). To cite Dekkers et al. (Int J Epidemiol 2010; 39: 89-94), "internal validity is a prerequisite for the external validity. Study results that deviate from the true effect due to systematic error lack the basis for generaliza-bility." In a similar vein, it is very doubtful whether NRS evidence is to be "expected to outperform" RCT evidence, when looking at both internal and external validity. As RCTs are accepted as the 'gold standard' in thera-peutic research, it is not a farfetched idea to examine how well an NRS emulates the 'ideal' RCT, even if the RCT may be impossible to perform in certain cases. Thus, these arguments do not lead to changes in the guideline.
6	6	16-18	The ACROBAT-NRSI tool was developed by the Cochrane Colloquium for the purpose of assessing the risk the bias of NRS for inclusion in their systematic reviews of interventions. It is not clear if the tool is appropriate for use to provide an assessment of a single NRS for purposes of HTA decision making. The context of the guideline is to assess if the NRS can be a replacement for a randomised trial, which may not be the	☒ major ☐ minor ☐ linguistic	According to this comment, a drawback of ACROBAT-NRSI is that it assumes "the evaluation of NRS to be performed subjectively as opposed to explicitly and quantitatively assessing the sources of bias in the actual study". To the best of the guideline authors' knowledge, there are currently no methods available which allow a quantitative and objective assessment of bias. If it were possible to quantitatively assess bias, we would be able

			reason for including NRS as part of an HTA submission. ACROBAT is based on the premise of maximizing internal validity and identifies potential biases and threat to it. It does so at the expense of addressing issues of generalizability (applicability) to populations excluded from the study. However, observational studies are often undertaken specifically to address this issue. As a result, they introduce sources of heterogeneity (i.e., bias) that are important to understand and consider in making optimal decisions on a new technology. In addition, ACROBAT proposes the use of a "generic target Randomized RCT". This is an interesting concept but it assumes the evaluation of NRS to be performed subjectively as opposed to explicitly and quantitatively assessing the sources of bias in the actual study. This is a potential weakness of the entire proposal.		to adjust the effect estimate accordingly and calculate a bias-free effect. Unfortunately, no reliable, objective method of bias correction exists. Assessment of RoB inevitably covers several domains of bias and involves some degree of subjective judgment. The other arguments presented here are already addressed in the two preceding replies. In our opinion no change in the guideline text is required.
7	6	19 on (Recommendation # 1)	**EDMA does not agree with the concepts on recommendation # 1.** <u>The efforts required to include NRS in HTA are of equal magnitude as the ones required to includ RCT</u>, the difference would lay in the instruments required to measure internal validity, and it is possible that knowledge building will be required inside HTA agencies as part of their internal quality framework, but this should be encouraged as part of the continuouly evolving HTA paradigm. Both types of evidence (RCT and NRS) will require adequate and specific knowledge (or knowledge building) on how to assess its internal validity and other charachteristics of key importance,	☒ major ☐ minor ☐ linguistic	Admittedly, the respective workload associated with the inclusion of NRS and RCTs in systematic reviews has not yet been compared. However, it is unlikely that identifying and assessing 10 NRS can be done in the same time as required for 10 RCTs. According to the vast practical experiences within the Cochrane Collaboration (Higgins et al., Res Syn Meth 2013; 4: 12-25), "there are currently no sensitive search filters for identifying studies on the basis of design labels or design features" (except for RCTs). If cohort and case-control designs cannot be reliably identified in bibliographic searches, much more effort is required for study selection. More effort is also necessary in the assessment of RoB. According to Higgins et al.,

			such as what kind of evidence is pragmatically possible to be achieved according to the specificity of the technologies and the time-point of assessment in the life cycle of the product. This is of key importance for in vitro diagnostics, as the consequences of test aplication on patient outcomes (impact/outcomes) are indirect through the influence of the information provided by the test on changes in the healthcare pathways and should not be regarded as dependant only on the IVD device but also interdependant on the context the device is applied to. This complexity is increased with the fact that in real life there are multiple decision points impacting those healthcare pathways through the interaction of more than one test, those pathways being highly dependant on the context were the patients are taken care of, converting the assessment of IVDs into a complex intervention scheme not suitable in most instances to be assessed through a RCT, as in this highly controlled environment it would be extremely difficult to recreate these complexities highly dependant on the context and the performance in real life conditions. On top of that the quantity of patients required for obtaining sufficient statistical power to obtain a given significant difference in outcomes assessed related to diagnostic technologies would in many instances prevent the undertaking of such studies, and even pose ethical issues related to the magnitude of patients required to be included. The concept stated in the recommendation that NRS often fail to increase the validity of the report's conclusion is not supported by EDMA or by the last review of the Cochrane		"assessment of risk of bias in NRS is more complicated than in randomized trials", because considerable epidemiological skills and content expertise is required. Very similarly, AHRQ authors noted that "the inclusion of data from observational studies increases the time and resources required to complete a comparative effectiveness review" (Norris et al., J Clin Epidemiol 2011; 64: 1178-1186). Thus, common experience from Cochrane, AHRQ and EUnetHTA researchers indicates that identification and assessment of evidence is more time-consuming for NRS than for RCTs. It is true that meta-epidemiological studies found, on average, no or only small differences between the results of NRS and RCTs. However, this averaging over many different research questions blurs the individual discrepancies noted in the primary studies, as differences cancel each other out. Several years ago (BMJ 1998; 317: 1185-90), Kunz and Oxman coined the term 'unpredictability paradox' to describe the problem that NRS often but not always give the same treatment effect as RCTs. This paradigm still holds true, as Dahabreh and Kent explain (JAMA 2014; 312: 129-130): "It is the unpredictability of the disagreement [between NRS and RCTs] that most undermines the credibility and limits the application of observational results." Thus, it may be true that NRS have been evolving over time towards higher quality standards, but they still do not reach the same internal validity as RCTs. The complexity or difficulty of designing and undertaking high-quality RCTs in some areas such as the example of in-vitro diagnostics mentioned by the stake-holders does not preclude that also for the evaluation of diagnostic tests – as mentioned before - the RCT is

			collaboration, where it concluded that results across all reviews (pooled ROR 1.08) are very similar to results reported by similarly conducted reviews. As such, they have reached similar conclusions; on average**, there is little evidence for significant effect estimate differences between observational studies and RCTs**, regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmaco-logical interventions (Anglemyer A. et al. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Editorial Group: Cochrane Methodology Review Group. Published Online: 29 APR 2014. Assessed as up-to-date: 11 JAN 2014-DOI: 10.1002/14651858.MR000034.pub2). Correlations of treat-ment effects may depend on the quality of studies chosen for comparison. Comparing RCTs with high-quality observational studies was more likely to yield similar results than comparisons of studies of mixed quality. For many techno-logies (particularly diagnostic devices), data from NRS of good quality is of key importance in order to be able to assess completely and accurately its effects and should definitely be sought for. Uniformed person strategy for treatment allocation, may overcome some of the issues associated with 'confounding by indication'. Restrictive cohort designs harness some features of RCTs, such as baseline assessment of prognostic risk factors, well defined eligibility criteria and intention to treat strategies. Subsequent application of appropriate statistical techniques such as multivariate analyses and propensity score methods		considered to be the gold-standard for the evaluation of patient-relevant outcomes (Lijmer & Bossuyt, J Clin Epidemiol 2009; 62: 364-73. Ferrante di Ruffano et al., BMJ 2012; 344: e686).

			may account for more than one influential variable, and provide enough statistical control to draw observational studies nearer to experimental validity. Literature review (please see biblio below) supports also the notion that NRS studies have been evolving over time towards higher quality standards and statistical methods for reducing risk of bias (such as propensity score, etc) have also evolved and are increasingly being applied in NRS production and assessment, diminishing or closing the gap between effect estimates derived from RCT and NR. In fact half of RCTs conducted achieve their recruitment targets and even less are completed on time. 60% of interventional trials registered at the ClinicalTrials.gov website (from 2000 through 2010) had anticipated enrollment of less than 100 participants, opening important question regarding the validity of results derived from such undertakings, even when they are grouped under metanalysis form at a later stage. Furthermore out of 137 000 trials recorded in the ClinicalTrial.gov database (accessed december 2012) less than 10% had published results. • Anglemyer A. et al. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Editorial Group: Cochrane Methodology Review Group. Published Online: 29 APR 2014. Assessed as up-to-date: 11 JAN 2014-DOI: 10.1002/14651858.MR000034.pub2). • Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000; 342:1878–1886. • Britton A, McKee M, Black N, et al. Choosing between randomised and nonrandomised studies: a systematic review. Health Technol Assess Winch Engl 1998; 2:1–6 • Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–1892.		

			• Ligthelm RJ, Borzı` V, Gumprecht J, et al. Importance of observational studies in clinical practice. Clin Ther 2007; 29:1284–1292.		
8		1-39	The previous statement base the fact that the arguments under - Possible reasons against inclusion of non-randomised studies (NRS) e.g The inclusion of NRS will nearly always increase RoB, – **are not valid to diagnostic devices** (or any medical technology in fact) and thus EDMA respectfully asks they are removed from the text. In fact the statement -The inclusion of NRS evidence might mislead researchers into the false belief that RCTs are not worthwhile to perform. Thus, HTA might act as a barrier in finding out the 'true' effect of an intervention-is challenged by the latest Cochrane report mentioned earlier (Anglemyer) as the true effect of an intervention could be derived both from RCT or NRS given that both study types are of adequate quality. Furthermore, interpretation of 'true associations' obtained from RCTs when small datasets are used resulting in inappropriately selected significance thresholds, suboptimal power, early termination strategies and flexible analyses. The effect sizes are likely to be inflated and subsequent reports biased. In the case of diagnostics, there will be many times where in fact the most advantageous and efficient design will be observational, with the IVD test working in real life conditions, drawing data from real patients (e.g registry type data), taking advantage of the newest and powerful sources of information and data analytic tools such as cloud computing and big data analytics based on electronic health records or other sources. Observational designs are a more affordable and efficient form of investigation in a clinical setting when diagnostic devices are concerned, and comprehensive administrative databases are useful tools to access large	☒ major ☐ minor ☐ linguistic	As stated in section 1.3, it was not the aim of the guideline to elaborate on the evaluation of diagnostic interventions. It is nevertheless true that observational studies, such as test accuracy studies, may be very valuable for the evaluation of diagnostic interventions, but assessment of their validity requires specific RoB tools (e.g. QUADAS-2; Ann Intern Med. 2011; 155: 529-36). The guideline would apply only to comparative studies of diagnostic interventions, where some patients undergo diagnostic testing while others do not. As stated by EDMA, it may well be that NRS are "more affordable" than RCTs. However, affordability was not considered relevant for the guideline. We agree that the following statement was misleading: "The inclusion of NRS will nearly always increase RoB". The statement lacks a clear comparison against which to gauge the additional value of NRS. In those cases where RCT data are lacking, inclusion of NRS does not increase RoB, because the alternative to NRS inclusion would be not to answer the HTA question. The sentence was changed as follows: "The inclusion of NRS as the sole information source will nearly always ~~increase RoB, thus~~ prevent~~ing~~ the results from being 'definitive'."

			representative populations as well as small subpopulations, enable long-term follow-up of patients, while maintaining ethical guidelines.		
9	7	Recommendation # 4	Regarding : At present, ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) should be used for the RoB assessment of NRS. EDMA asks that EUnetHTA provides evidence on validation of this instrument in the European setting applied to observational data in general and evidence from registries in particular, before recommending it as the prime instrument for assessing ROB in NRS.	☒ major ☐ minor ☐ linguistic	In the current version of the guideline, it is stated that ACROBAT-NRSI is a new instrument. As experience with ACROBAT-NRSI is therefore very limited, the current guideline should be reassessed in a few years' time in order to include more information on this and other RoB tools. The developers of ACROBAT-NRSI are currently starting a larger-scale evaluation of ACROBAT-NRSI (Higgins J. Personal communication. June, 2015). Currently, only face validity or criterion validity of ACROBAT-NRSI is evident, as this tool addresses all bias domains. However, as no other instrument showed better validity (and reliability), it appears appropriate to recommend this as the best available RoB tool.
10	8	2	There needs to be a definition of external validity as this concept is discussed in the document as is often a key reason why NRS are used given that RCTs often have poor external validity.		A definition of applicability (i.e. external validity) has now been added.
11	8	17-18	Editorial comments are not necessary in a document that is providing recommendations.	☐ major ☒ minor ☐ linguistic	As the first sentence only defines the term "confounders", the second sentence is necessary to explain "confounding".
12	8	27	Given the heterogeneity of backgrounds and skills of HTAs , wouldn't be helpful to state that this studies are usually called observational? Or refer to other existing Glossaries. Also mentioning the growing use of the term Real World Evidence (RWE) as NRS would be also clarifiying for guideline's users	☐ major ☐ minor ☒ linguistic	As suggested, the term "observational" was added to the definition. The term "real-world evidence" was not added, as this term implies high applicability and thus is not neutral.

13	9	3	While randomised controlled trials (RCTs) provide the most robust evidence regarding the question of an interventions efficacy and safety, they do this by maximising internal validity over external validity. This is built into the design of the RCT in order to meet the specific questions of drug regulatory authorities. However, the payer is usually also interested in understanding external validity and therefore how the intervention is going to perform in the real world where conditions are not controlled. This is a weakness of RCTs and is where NRS can provide complementary evidence given that they are often designed to maximised external validity at the expense of internal validity. See for example Rawlins M (2008) De testimonio: on the evidence for decisions about the use of therapeutic interventions. Lancet 372: 2152-2161.	☐ major ☒ minor ☐ linguistic	As already stated in the reply to comment No. 5, applicability is outside the scope of the present guideline. Due to their lower internal validity, NRS are only seldom able to provide useful complementary evidence on effectiveness. The same, albeit to a somewhat lesser extent, is also true for safety. Let us assume, for example, that RCTs from highly experienced centres show that a new interventional procedure lowers mortality in well-selected patients. In this situation, it is not very helpful to know from NRS that mortality is higher in less qualified centres and unselected patients. The treatment effect itself and possible effect-modifying variables are primarily relevant for decision-makers. If NRS show high mortality rates, neither the true reasons for mortality nor the true treatment effect in this setting can be delineated. The assessment of the intervention should thus only be affected if event rates or treatment effects in NRS greatly differ from those observed in RCTs.
14	9	11	There is no mention of the real world environment within which most NRS studies are conducted, offering insight to how products perform in patients that were not eligible for RCTs.	☒ major ☐ minor ☐ linguistic	NRS can be used to find out which patients might be eligible to receive a new intervention. If some patient groups (e.g. pregnant women, patients mentally incapable to provide informed consent) are not eligible for RCTs, NRS certainly offer insight into likely treatment effects. Still, RCT evidence is to be preferred, unless such evidence is unobtainable (c.f. section 1.2 first bullet point).
15	9	11 and 21	Not clear „inclusion" to what? Should state this is about using RCT and NRS for HTA at the beginning of the 1st para, e.g, at "… relative advantages of **using** randomised … **in HTA** has …"	☐ major ☐ minor ☒ linguistic	The inclusion of NRS covered by this guideline is restricted to questions on inclusion of NRS to answer questions on relative efficacy or effectiveness of medical interventions. This is stated in the scope and summary of the guideline.

16	9	12	The parenthetical comments highlight the bias of the authors of this document.	☐ major ☒ minor ☐ linguistic	Awareness of the strengths and limitations of different study designs based on the current knowledge and experience in evidence-based medicine is a pre-requisite for being able to judge risk of bias.
17	9	15-16	Quite the contrary, due to the large sample sizes often available in NRS, small effects can be detected with sufficient power, that would have gone undetected in small RCTs; despite the RoB	☒ major ☐ minor ☐ linguistic	Large sample sizes may detect statistically significant effects, but the credibility of these effects may be diminished due to RoB. This applies to any study design, but is a more inherent problem of NRS compared to RCTs
18	9	17	Rather than, ‚best guess', perhaps, ‚best available of evidence'	☐ major ☐ minor ☒ linguistic	As the best-available evidence approach is a more structured stepwise approach in evidence-based medicine, we feel that 'best guess' better explains the issue.
19	9	22-23	, thus preventing the results from being 'definitive'' is not necessary, as RCTs are not 'definitive' either.	☐ major ☐ minor ☒ linguistic	The word 'definitive' was already placed in quotation marks to indicate that no absolute truth can be drawn from RCTs.
20	9	26-27	'Increasing workload' should not be a reason to not evaluate all available evidence.	☐ major ☒ minor ☐ linguistic	Efficiency is a necessity in HTA, as human and financial resources are limited.
21	9	28-30	Not sure if this should be a reason. The researcher should have the responsibility of not being misled.	☐ major ☒ minor ☐ linguistic	Although researchers are also responsible for performing clinical trials on adequately selected topics, in practice RCTs are much more difficult to perform once HTA agencies and reimbursement bodies have accepted a new intervention (Hamilton et al., J Clin Oncol 2010; 28: 5067-73; Walker et al., Value Health 2012; 15: 570-579).
22	9	31-32	Worth noting that the RoB for NRS will largely depend on whether intended or unintended effects are being assessed.	☒ major ☐ minor ☐ linguistic	The guideline does not state that RoB of NRS depends on whether intended or unintended effects are being assessed.
23	9	31-	It seems the document (here and page 19) suggests not to use RCT and NRS together in one HTA. I feel as a document for RoB	☐ major ☒ minor	The guideline authors were asked to add some ideas on the crucial decision when to include NRS in HTA. There-

		34	assessment for NRS, there is no need to include such a recommendation.	☐ linguistic	fore, page 9 contains several reasons for and against NRS inclusion. Thus, it is also worthwhile to comment on the role of NRS as additional evidence to RCTs. All arguments on page 9, however, address only internal validity.
24	9	32-34	HTA's want to consider the generalizability of interventions and want more information on how interventions will perform in the "Real world" setting, outside of clinical trials, therefore it is not appropriate to assume that NRS will not be seen as additional source of information. This shows a bias from the authors of this guideline where they only focus on the hierarchy of evidence, and not consider NRS in supporting randomised trials	☒ major ☐ minor ☐ linguistic	The comment assumes that the "real world" exists only "outside of clinical trials". Many clinical trials, however, mirror the real world very well, and NRS do not neces-sarily provide better applicability than RCTs. As explained in Point 5 of this document, internal validity is a prerequisite for the external validity. We have added the words "on effectiveness and safety" after "information", in order to highlight the guideline aim and to open the door for using NRS evidence specifically for an assessment of applicability.
25	10	4	In my opinion one of the two main questions, i.e. "How to classify NRS evidence according to study design?" is not really addressed in this draft guideline. It would be useful to add more info on this topic or change guideline scope.	☒ major ☐ minor ☐ linguistic	In sections 2.5 and 2.7, it is explained that "at least a basic classification of study design (e.g. cohort study, case-control study)" is required for ACROBAT-NRSI (and also several other tools). Thus, the second question of the guideline is implicitly answered by recommending this RoB tool. In addition, the guideline stresses the importance of correctly identifying the design of a given study (page 16, line 34).
26	10	4	Is „evidence" needed? Should we add "conduct and reporting"?	☐ major ☐ minor ☒ linguistic	As NRS provide evidence, they should be named so ("NRS evidence"). Conduct and reporting of NRS cannot be classified. Furthermore, reporting of NRS is only indirectly linked to internal validity.
27	**10**	17-21	The choice of tool and the importance of diferent biases can only be assessed in the context of how the NRS is to be used. This guideline appears to just consider the NRS in the absence of randomised trials. If that is the scope then the criteria for the tool is acceptable, but if the tool is to be used to assess all NRS,		Internal validity is a construct that is independent of the purposes for which the evidence is used. It is not logical that the existence of RCTs might affect the validity of NRS. Accordingly, the present guideline selected a tool for assessing the internal validity of NRS (excluding case series), independent of whether RCTs are available or

			then the criteria is not appropriate.		not.
28	**10**	23-25	External validity should be considered along side Internal validity, so the two guidelines should be combined, and preferably use just one tool to assess the quality of NRS.		As most researchers recommend separate assessments of internal validity and applicability, the guideline already contains the statement that "it is important not to mix up internal and external validity".
29	10	37-42	Not sure why case series studies are considered different from case-control studies, all estimate subject specific treatment effects.	☐ major ☒ minor ☐ linguistic	Case series lack a comparison group. Therefore, a treatment "effect" can only be estimated by looking at pre-post-changes or comparing the results with external (e.g. literature) controls. Case-control studies compare patients with an outcome (i.e. cases) and those without the outcome (i.e. controls). They examine whether cases are more likely than controls to have received an intervention (or exposure).
30	11	5	Why „RCT"? Are we talking about NRS?	☐ major ☐ minor ☒ linguistic	Section 1.4 provides information on related documents. It is important that ACROBAT-NRSI is not used to assess the validity of RCTs, as better RoB tools are available for this purpose.
31	12	3	Key criteria that high-quality RoB tool should fulfil, are defined . However ,there is no quotation on who defined the criteria neither based on what elements, principles or dimensions are described.	☒ major ☐ minor ☐ linguistic	Four criteria (criteria a-d) were selected by the guideline authors. The criteria aim to ensure that the tools are fit for purpose by assessing their applicability (a and b), their distinction between quality of reporting and quality of research as well as internal and external validity (c) and their comprehensiveness regarding the most important types of bias (d). The principles/dimensions for defining five important types of bias are explained in more detail in the guideline for RoB in RCTs and by a review, both cited in the text. In our opinion no change in the text is required.
32	12	5	Should these be two types of designs, including variations of both case-control and cohort? Where cross-section studies go,	☐ major ☒ minor ☐ linguistic	The role of cross-sectional studies is in fact not yet fully defined, but cross-sectional studies are *per se* less able to analyse a cause-effect-relationship than other NRS.

			as a special case of cohort?		Cross-sectional studies present an additional domain of bias, which relates to the sequence of cause and effect. While cause clearly precedes effect in cohort and case-control designs, the parallel ascertainment of cause and effect in a cross-sectional study makes it more difficult to rely on this basic tenet of clinical research.
33	12	8-9	By excluding 'external validity' the authors excluding the primary benefit of including NRS. Further, a tool should not be excluded if it includes both Internal and External validity, even if it is agreed that External validity is out of scope (i.e. The internal validity section may still be superior to other tools)	☒ major ☐ minor ☐ linguistic	Using only the internal validity section of an existing RoB tool means to create a new tool, which would require extensive evaluation and piloting. Therefore, it appears better to recommend the best currently available tool, which is ACROBAT-NRSI. The problems associated with modifications of existing RoB tools are now explained in section 2.1, where a sentence was added.
34	12	8-9	Why should a tool assessing both Internal and External validity be excluded? The internal validity section of a broader tool may be better to tools assessing only the internal validity.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
35	12	10	Please be consistent across the guideline document as sometimes it is reported there are 5 different bias domains, sometimes 6 bias domains (cf. 6 domains mentioned on: p6 third recommendation; Table 1 p15; p19 line 7).	☐ major ☒ minor ☐ linguistic	The difference between 5 and 6 bias domains arises from the fact that selection bias and bias due to confounding are often seen as just one bias domain, because in many observational studies inclusion of patients in the study and allocation to an intervention are not clearly separated from each other. In the text, we now explain this issue in more detail and consistently refer to 5 bias domains. The recommendation has been changed accordingly.
36	6/12	10	Define „domain" or give refs	☐ major ☐ minor ☒ linguistic	In psychometry, quality-of-life research, and medical statistics, a domain is a set of individual variables that are closely related to other, because all variables essentially measure the same underlying construct (Scientific Advisory Committee of the Medical

					Outcomes Trust, Qual Life Res 2002; 11: 193-205). Health-related quality of life, for example, includes the domains of physical, mental, and social health.
37	12	11	It is stated that the four criteria mentioned above are considered mandatory. However it is not quoted by whom? Based on what (there is no quotation)	☒ major ☐ minor ☐ linguistic	- please see reply to comment No.31 -
38	**12**	21-27	The point made regarding quality of reporting and quality of research is true, but a poorly reported study may make it difficult to assess the quality of research. If the assessment of NRS internal bias is being made from a publication then the quality of reporting may influence that assessment and so it may be important to understand if it is due to poor reporting that the study is rated of high risk of bias, rather than the actual different types of internal bias.	☐ major ☒ minor ☐ linguistic	Poor reporting indirectly affects the assessment of RoB. High-quality RoB tools make it transparent whether a study was affected by an overt methodological short-coming or inadequately reported methodological details. In ACROBAT-NRSI, the response options to all items include a "no information" category.
39	12	21-27	The point regarding quality of reporting and quality of research is true, but a poorly reported study may make it dificult to assess the quality of research. If the assessment of NRS internal bias is being made from a publication, then the quality of reporting may influence that assessment, and so it may be important to understand if it is due to poor reporting (and many time, lack of reporting) that the study is rated of high risk of bias, rather than the actual diferent types of internal bias.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
40	12	24	"study" instead of "trial"	☐ major ☐ minor ☒ linguistic	Changed as proposed.
41	**12**	31-	An assumption is made that "Because external validity is dependent on the clinical setting, it is in the eye of the beholder	☒ major ☐ minor	The assessment of applicability (or external validity) was outside the scope of the present guideline. In

		34	and can never be proven on a ubiquitous level. Therefore, the assessment of RoB does not include external validity, and any RoB tool containing items on external validity was deemed less suitable for recommendation". This seems to be a flawed assumption. We propose that external validity needs to be considered. It does not need to be solely a subjective assessment (i.e., "in the eyes of the beholder") given that a definitive (external) population for the questions and technology being addressed does exist. We would propose that the target population forms a comparison by which the external validity of any study can be assessed. Furthermore, once such a target population is defined, the unique characteristics of the study population can be used to look for sources of heterogeneity or biases that matter in the specific decision and target population.	☐ linguistic	addition, the comment lacks criteria how to define the target population of a study. HTA agencies in one European country may want to address the needs of one target population, while another target population is considered important in another country. Thus, external validity is in the eye of the beholder.
42	12	37-38	Prior section notes 'Bias due to confounding predominates in observational studies'	☐ major ☒ minor ☐ linguistic	As the term 'selection bias' in a broader sense also included bias due to confounding, the two sentences are not contradictory.
43	12	37-40	Not clear what "latter" and "former" refer to, biases due to sampling participants and assignment of treatments?	☐ major ☐ minor ☒ linguistic	Yes, "latter" refers to assignment of treatments.
44	13	1	Please be consistent across the guideline document as sometimes it is reported there are 5 different bias domains, sometimes 6 bias domains (cf. 6 domains mentioned on: p6 third recommendation; Table 1 p15; p19 line 7).	☐ major ☒ minor ☐ linguistic	- please see reply to comment No.35 -
45	13	3 and	The non mandatory criteria, are clearly subjective. Should be mentioned . So the assessment results shoud be interpreted	☒ major ☐ minor ☐ linguistic	Ease-of-use could be measured objectively by recording the time required to complete one NRS assessment. Acceptance among systematic reviewers could be

		4	with caution and susceptible of appeals by sponsors.		quantified by analysing HTA reports or systematic reviews.
46	13	6	The proposed list of RoB tools is impressive. However, they all rely on subjective assessments of process. Attention should be given to assessing the outcomes (e.g., was bias a real issue and was it adequately addressed by the process). This can, for select issues, be done objectively via quantitative methods. Such a more objective and definitive approach should be considered and use when feasible. More specifically, the diagnostics of the actual methods use can be useful in determining if they succeeded in addressing the bias or other concern for which they were employed.	☒ major ☐ minor ☐ linguistic	The fact that all RoB tools rely on subjective judgment shows that RoB assessment is an inevitably subjective process. Already in 2003 it was stated that "study quality is a rather subjective concept, open to different interpretations depending on the reader" (Deeks et al., Health Technol Assess 2003; 7: iii-x, 1-173). It remains unclear what the comment refers to when describing objective and quantitative methods. Methods for bias-adjusted meta-analysis have been proposed (Turner et al., J R Stat Soc Ser A Stat Soc 2009; 172: 21-47), but their validity is uncertain and their use is very limited.
47	13	11	Please specify end of research period (I believe it is November 2013).	☐ major ☒ minor ☐ linguistic	The search period ended in 2013, as the literature search was performed in early 2014 (as described in Annexe 2). The sentence was not changed, as the connection to the previous sentence would no longer be logical.
48	13	15-19	It is questionable if the assessment of bias tools is not biased in the way it has been conducted. The ACROBAT-NSRI tool was not identified within the Literature search, and this raises the question as to whether the developers of the guideline were influenced by being asked to pilot a new tool, and not able to assess the other tools without prejudice.	☒ major ☐ minor ☐ linguistic	- please see reply to comment No.4 -
49	13	15-19	The way ACROBAT-NSRI tool was chosen may contain some bias. The ACROBAT-NSRI tool was not identified within the literature search and the developers of this guideline, by being involved in	☒ major ☐ minor ☐ linguistic	- please see reply to comment No.4 -

			the pilot of this tool, may have been biased towards this tool.		
50	14	5-9	The ISPOR questionnaire meets all of the key bias domains (see Table 1), but has been excluded because it also includes a section on quality of reporting. However the way the ISPOR questionnaire is designed, these questions can be omitted if it is not felt relevant for HTA purposes. So the rational for excluding this tool is weak.	☒ major ☐ minor ☐ linguistic	The article by Berger et al. (Value Health 2014; 17: 143-156) does not mention the option of omitting irrelevant questions. The authors clearly state that they developed "a single questionnaire consisting of 33 items". As this instrument also includes questions on applicability and on statistical precision of results, the rationale for excluding this tool is strong.
51	14	5-9	The ISPOR questionnaire meets all of the key bias domains (see Table 1), but has been excluded because it also inclusdes a section on quality of reporting. However the way the ISPOR questionnaire is designed, these questions can be omited if it is not felt relevent for HTA purposes. So the rational for excluding these tool is weak.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
52	14	7-8	Having more than what is needed does not seem like a rational rationale for excluding from evaluation. Easy adaptations could be made to address the concern.	☐ major ☒ minor ☐ linguistic	- please see reply to comment No.50 and 33 -
53	14	15-21	The Rationale for excluding the Downs-Black checklist appears weak, given the openning statement "has received widespread international recognition", especially in comparison with ACROBAT which has not had much widespread use, and as only just finished pilot testing.	☒ major ☐ minor ☐ linguistic	Acceptance among systematic reviewers was only an additional criterion for selecting the best RoB tool.
54	14	15-21	The Rationale for excluding the Downs-Black checklist appears weak, given the openning statement "has received widespread international recognition", especially in comparison with	☒ major ☐ minor ☐ linguistic	- please see previous reply -

			ACROBAT which has not been used outside the pilot testing.		
55	16	3	Four tools should have been considered for reliability and ease of use, (ACROBAT, ISPOR, Downs-Black and RoBANS), this would provide a clearer rationale for the final choice of tool given ISPOR and Downs-Black are more established tools. It would allow the assessment of reliability data in these two more established tools.	☒ major ☐ minor ☐ linguistic	As set out in the previous comments, the Berger/ISPOR and the Downs-Black instrument both do not fulfil the essential requirements to be recommended.
56	16	3	More tools should have been considered for reliability and ease of use. Given ISPOR and Downs-Black are established tools and that they are not included in the list to choose from, the rational of the choice for ACROBAT sounds weak.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
57	16	4	Due to novelty of RoB tools, reliability data are sparse (RoBANS) or absent (ACROBAT-NRS). Based on what are these instruments recommeded in page 7 ?? Please mention it	☒ major ☐ minor ☐ linguistic	As reliability was not among the criteria for selecting RoB tools, ACROBAT-NRSI and RoBANS were both considered suitable for use. The validity of a RoB tool is clearly more important than its reliability.
58	16	4-5	The lack of validity and reliability testing of the proposed instruments is concerning. Only face and content validity has been assessed. This is necessary but not sufficient. An assessment should be performed of these critical properties for all identified instruments. We believe that some may fare better than the proposed instruments IF others have better reliability and validity then they may be more preferable to use.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
59	16	16	The authors indicate that both ACROBAT-NRSI and RoBANS are new and no literature on ease of use. Does the novelty of both RoB tools compromise their reliability? Have the tools been tested enough to be trustworthy? Please comment in the text	☒ major ☐ minor ☐ linguistic	In section 2.4, the guideline clearly describes that data on reliability and ease-of-use are essentially lacking for ACROBAT-NRSI and RoBANS. Therefore, in our opinion no change is necessary.

			on their limitations in assessing risk of bias in non-randomized studies		
60	16	16-24	Any reason why only ACROBAT-NRSI was piloted regarding this aspect (and not both?) The assessment of ease-of-use was biased in this guideline, as the team developing the guidelines undertook the pilot testing of ACROBAT and then also made the assessment of ease of use. All four potential tools should have been assessed by an independent group of researchers, who did not have the advantage of participating in the pilot testing.	☒ major ☐ minor ☐ linguistic	The Berger/ISPOR and the Downs-Black instrument both did not fulfil essential requirements, primarily due to insufficient content validity. Ease-of-use had no influence on the decision to exclude these two instruments and to include the two other instruments. Therefore, the selection of RoB instruments was not biased.
61	16	16-24	The assessment of ease-of-use is biased in this guideline, as the team developing the guidelines, undertook the pilot testing of ACROBAT and made the assessment of ease of Use. All tools that fully covered the key bias domains should have been assessed by an indpendent group of researchers, independent of the pilot testing.	☒ major ☐ minor ☐ linguistic	- please see previous reply -
62	16	23	It is mentioned that a key advantage of ACROBAT-NRS is the availability of methodological guidance. Should be explicetely mentioned *„evidence based* methodological guidance"? I presume is reliable , but should it be explictely mentioned or there are known limitations? If yes mention it explicitely	☒ major ☐ minor ☐ linguistic	The guidance document on ACROBAT-NRSI does not claim to be evidence-based. Therefore, the text in the present guideline is appropriate.
63	16	30	'is'	☐ major ☐ minor ☒ linguistic	Thanks for noting this omission, which is now corrected.
64	16	32-35	Based on this, there will be discongruence between what is reported in the literature as the study design and what is being evaluated? This will add another level of subjectivity and bias to	☐ major ☒ minor ☐ linguistic	Because so many clinical studies are wrongly labelled with regard to study design, checking the study design is necessary in order to reduce errors. Of course, the

			the assessment tool.		assessor of a study also errs occasionally, but this does not produce bias, because the assessor should be free from conflict of interests.
65	16	36-38	The sentence "This represents…" is not clear to me.	☐ major ☐ minor ☒ linguistic	We have added a sentence which uses the examples of cohort vs. case-control studies and prospective vs. retrospective designs in order to explain more clearly what is meant here.
66	16	39-45	In general NRS are not intended to be a replacement for RCT, they are usuallt designed to complement the results from RCT's. There is a risk that the ACROBAT tool will misclassify	☒ major ☐ minor ☐ linguistic	When RCT and NRS are used for different assessment aims (such as effectiveness and applicability), the two types of evidence complement each other. As the present guideline is restricted to internal validity (which is essential when assessing effectiveness and safety), it addresses the question whether and how both types of evidence can be used for the same aim.
67	16	39-45	The intention of the para. Is not clear. The 1st sentence says "researchers should clearly define whether RoB of NRS can reach the RCT level…" but the rest says they shouldn't. Also should define be "determine" (minor)?	☐ major ☒ minor ☐ linguistic	Thanks for noting this lack of clarity. The question "whether RoB of NRS can reach the RCT level" should indeed not be decided for each single HTA report. The authors of ACROBAT-NRSI warn against putting NRS on the RCT level and state that "most NRS will be judged as at least at moderate overall risk of bias".
68	16	43-45	Editorial comments are not necessary in a document that is providing recommendations. There are plenty of references that can present a different opinion than 'Barton'.	☐ major ☐ minor ☒ linguistic	The comment was replaced by an instruction taken from the ACROBAT-NRSI manual.
69	17	1	It is not clear to me how to reach an overall assessment of evidence using the 7 domains. What if for instance the studies had large magnitude effect and yet suffer from publication bias?	☐ major ☒ minor ☐ linguistic	According to GRADE, all the criteria for down-grading the quality of the evidence should be assessed prior to considering an up-grade of the quality of the evidence. If any serious limitations have been identified for the former criteria, the quality of the evidence should only rarely be upgraded (Guyatt et al., J Clin Epidemiol 2011; 64: 1311-6).

70	17	4	I guess "certainty of results" here means the certainty of have any treatment effects rather that of quantitative results.	☐ major ☐ minor ☒ linguistic	Yes, certainty refers to the presence of a positive or negative effect - regardless of quantitative effect size.
71	17	19-30	This section discusses the importance of assessing other domains for HTA decision making that relate to external validity, and reference seperate EUnetHTA guidelines. A tool which provides the option to assess both internal and external validity would be prefereable than having to use too seperate tools. So the rational for excluding tools because they include domains on external validity is not consistent with this section of the guidelines.	☒ major ☐ minor ☐ linguistic	Most researchers recommend separate assessments of internal validity and applicability (Higgins et al., BMJ 2011; 343: d5928). Thus, the guideline already contains the statement that "it is important not to mix up internal and external validity". A very similar statement was already made in the previous EUnetHTA guideline on RCTs. In our opinion no changes are required.
72	17	25	EDMA respectfully asks EUnetHTA to provide evidence which sustains the following argument: Registry analyses come with the promise of minimal selection bias, minimal attrition bias, and a good ability to control for confounding, but their true internal validity may well be lower than that of conventional cohort studies- Given previous presented arguments, and many examples even in Europe (please see bible bellow) of the high quality and reliability of registries data when drawing information for assessing medical technology, EDMA respectfully asks this paragraph to be revised. -Petersson et al. Internal validity of the Swedish Maternal Health Care Register. BMC Health Services Research. http://www.biomedcentral.com/1472-6963/14/364.	☒ major ☐ minor ☐ linguistic	Registry data are often collected without specific research questions in mind or for administrative purposes. Given their different purpose, it may well be that losses to follow-up, co-interventions, unmeasured confounders, selective reporting or other sources of bias are more prevalent in the registry analysis. Therefore, it appears fully tenable to state that registry analyses may be less valid as compared to other observational studies. It is noteworthy that the guideline does not state whether registry analyses in general provide higher or lower internal validity when compared to other observational data.
73	17	28	'Registry studies' should be defined – according to the definitions section on this document, registries would meet the definition of a 'cohort study'	☐ major ☒ minor ☐ linguistic	For good reason, the vague term 'registry studies' was avoided in the document. In recommendation No.6, the word 'registry studies' was now replaced by 'registry analyses'.

74	17	19-30	This section discusses the importance of assessing other domains for HTA decision making that relate to external validity, and reference seperate EUnetHTA guidelines. A tool which provides the option to assess both internal and external validity would be prefereable than having to use too seperate tools. So the rational for excluding tools because they include domians on external validity are not consistent with this section of the guidelines.	☒ major ☐ minor ☐ linguistic	- please see reply to comment No.4 -
75	18	12	It is stated that RoB assessment is an inevitable subjective process.... I suggest the following rewording: Given the lack of evidence on reliability of the tools for RoB assessment and the inevitability of subjective value juddgemt in the assessment . Detailed reporting of the assessment results ,transparent description of assessors profiles/credentials and right of appeal from sponsors are recomneded. This statement shoul also be included in pag 7 of summary od recomendations	☒ major ☐ minor ☐ linguistic	The guideline does not address the question of how HTA agencies and external stakeholders should interact. Nevertheless, most HTA agencies routinely contact trial sponsors or authors in order to obtain additional information on study conduct and results. Many agencies also offer the opportunity that external stakeholders can comment on a version of the HTA report. This essentially represents a "right of appeal", as suggested in the comment.
76	19	13-16	What if some of the evidence comes from cross-sectional studies? If I understood it correctly, RoBANS provide assessment of cross-sectional study designs while ACROBAT-NSRI does not. Are we then supposed to use ACROBAT-NSRI for case-control and cohort studies and RoBANS for cross-sectional ones in one project?	☐ major ☒ minor ☐ linguistic	- please see reply to comment No.32 -
77	19	17	There is a mention that ACROBAT-NSRI is the best tool for assessment....of interventions. Suggest the use of „non-randomised observational studies" instead of inteventions.	☐ major ☒ minor ☐ linguistic	The difference between "non-randomized studies on interventions" and all non-randomized (or observatio-nal) studies is important here, as the guideline does not cover any observational studies on test accuracy, prognosis, prevalence or incidence.

78	29	3	Why are the publications for the tools not identified in the literature search, i.e. ISPOR, EPHPP, GRACE and NOS. All were published after 2005. Their absence suggests that other tools could also have been missed.	☒ major ☐ minor ☐ linguistic	The articles on the ISPOR and the GRACE checklist were both published in March 2014, which was after the bibliographic searches for the present guideline were performed. Both the NOS and the EPHPP tool were never published in a journal article. Thus, literature searches appear to be sufficiently sensitive. In this context, it is worth mentioning that in May 2015 authors working for NICE proposed a new RoB assessment tool called QuEENS (Quality of Effectiveness Estimates from Non-randomised Studies; http://www.nicedsu.org.uk/Observational-data-TSD%282973296%29.htm). This checklist contains items on reporting quality ("Have the results of the study been compared to others in the literature?") and statistical precision ("Is the sample size relatively large?"). Therefore, the QuEENS checklist does not meet the essential requirements to be used for RoB assessment.

EFPIA - European Federation of Pharmaceutical Industries and Associations

EDMA – European Diagnostic Manufacturers Association

EFSPI HTA SIG - European Federation of Statisticians in the Pharmaceutical Industry (EFPSI) Health Technology Assessment (HTA) Special Interest Group (SIG)