JA2- WP7- SG 3 – **SAG and External** consultation on the methodological guideline "Meta-analysis diagnostic test accuracy studies"
1st version of draft guideline

Number of comment	Page	Line	Comment	Character of comment "major"[1] "minor"[2] "linguistic"[3]	Author/Draft group reply
1	13-14	Whole section (1.1.5)	ROC curves may not be available for some areas as threshold values may not exist when they are either too high or too low. It would be useful to note this.	☐ major ☒ minor ☐ linguistic	Text has been added to the paragraph following Figure 3. It should be noted that this issue was already addressed in the later text in section 2.2.4 in the context of SROC curves.
2	14	307	The line of symmetry and Q estimate could be added to the plot of ROC Curve for clarification purposes. CRD guidance for undertaking reviews in health care. Jan 2009. Chapter 2: Clinical Tests //www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/2_2_DIAGNOSTIC_TESTS.htm	☐major ☐minor ☐**Xlinguistic**	The ROC curve in Figure 3 has been amended to include the line of symmetry and Q*. The subsequent text also explains the relevance of Q*.
3	15	342	Where it says "The predictive values are referred to as specific **un**conditional measures of test accuracy" it should say "The predictive values are referred to as specific **conditional** measures of test accuracy" as they depend on disease prevalence.	x☐**major** ☐minor ☐linguistic	Text has been corrected.

[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

4	15	367-368	Due to economic and practical reasons RCTs cannot be conducted for all tests and for all their intended uses as • They need a lareger number of patients than treatment trials • They can be more time consuming • They can be more expensive • They are often problematic to design and undertake • They are potentially wasteful of resource and uninformative (Ouwendijk R et al. Imaging peripheral arterial disease: a randomized controlled trial comparing contrast-enhanced MR angiography and multi-detector row CT angiography, Radiology. 2005 Sep;236(3):1094-103) The opinion of decision makers and payers should be taken into account because they are important stakeholder with regards to what extent they need and value data from potential outcome studies. RCT cost would need to be reflected in test cost and thus impact decision makers' budgets. For the sake of efficiency of health care systems more consensus is needed on in which cases an RCT is required and when a linked evidence approach is sufficient: Merlin at. al point out that "… namely, that LEA[4] may be inadequate to act as a proxy for direct evidence in instances where there are spectrum of disease	☐major ☐minor ☐**xlinguistic**	The text referred to here deals with the difficulties in associating diagnostic test performance with longer term outcomes. There are other options, such as the linked evidence approach. However, the guideline is restricted to meta-analysis of diagnostic test accuracy, and longer term outcomes are beyond the scope of the document.

[4] linked evidence approach

			differences between the tested population and the treated population (i.e., the test identifies new cases that cannot be identified with existing tests); and where there is an imperfect reference standard against which to determine test accuracy. … No problems were identified using LEA for "rule in" tests." Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care. 2013 Jul;29(3):343-50. Modern methods like directed-acyclic graphs can be used to assess potential confounding (Emily Cox et al. Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II, VALUE IN HEALTH, Volume 12, Number 8, 2009). They can enhance the contribution of non-RCT types of evidence to the decision before the necessity of RCTs is concluded.		
5	15	367-368	Due to economic reasons RCTs cannot be conducted for all tests and for all their intended uses. The opinion of decision makers and payers should be taken into account because they are important stakeholder with regards to what extent they need and value data from potential outcome studies. RCT cost would need to be reflected in		See response to comment number 4.

			test cost and thus impact decision makers' budgets. For the sake of efficiency of health care systems more clarity is needed on in which cases an RCT is required and when a linked evidence approach is sufficient: Merlin at. al point out that "… namely, that LEA[5] may be inadequate to act as a proxy for direct evidence in instances where there are spectrum of disease differences between the tested population and the treated population (i.e., the test identifies new cases that cannot be identified with existing tests); and where there is an imperfect reference standard against which to determine test accuracy. … No problems were identified using LEA for "rule in" tests." Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care. 2013 Jul;29(3):343-50. Modern methods like directed-acyclic graphs can be used to assess potential confounding (Emily Cox et al. Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II, VALUE IN HEALTH,		

[5] linked evidence approach

			Volume 12, Number 8, 2009). They can enhance the contribution of non-RCT types of evidence to the decision before the necessity of RCTs is concluded.		
6	16	373	More detailed information on the **linked evidence approach** and inputs when evidence is scarce would improve guideline applicability, e.g. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care. 2013 Jul;29(3):343-50.	x☐**major** ☐minor ☐linguistic	See response to comment number 4. The suggested reference is already cited in this section.
7	16	374	An **analytic framework to test evaluation like "Figure 2-2: Example of an analytical framework within an overarching conceptual framework in the evaluation of breast biopsy techniques"** could be incorporated to the guideline: "organizational frameworks are useful in categorizing key questions and which types of studies would be most useful for specific questions in the review. They may be useful in clustering studies to be reviewed together, and this may improve the readability of a review document. No specific framework is recommended, and indeed the categories of most organizational frameworks at least approximately line up with the analytic framework and the PICO(TS) elements…" Agency for Healthcare Research and Quality. Methods Guide for Medical Test Reviews [posted November 2010].	☐**Xmajor** ☐minor ☐linguistic	See response to comment number 4. We appreciate that different study types may be appropriate for answering questions regarding intermediate and longer term outcomes, but these issues are outside the scope of this guideline.

			http://www.effectivehealthcare.ahrq.gov/tasks/sites/ehc/assets/File/methods_guide_for_medical_tests.pdf		
8	15	360	Intermediate outcomes, mainly the **impact of test results on patient management** could also be included as a strategy for linking evidence when clinical outcomes are scarce. It will be good to add some **guideline on adequatelly reporting** these outcomes, its rationale, assumptions, and quality of input studies (Staub LP. .Int J Technol Assess Health Care. 2012 Jan;28(1):52-8).	☐**Xmajor** ☐minor ☐linguistic	See response to comment number 4. Such guidance may have to form part of a more general guideline on diagnostic tests. This guideline is restricted to the meta-analysis of diagnostic test accuracy studies.
9	16	395	Expanding explanation and practical tools for implementing **Tiered Model of diagnostic efficacy** approach would be useful.	☐**Xmajor** ☐minor ☐linguistic	As the guideline is restricted to the meta-analysis of diagnostic test accuracy, most of the components of the tiered model do not apply to the guideline. The purpose of its inclusion was merely to show where diagnostic test accuracy sits in the context of overall diagnostic efficacy.
10	17	416	"Scope: The guideline will not address issues relating to systematic reviews and meta-analysis that are not restricted or unique to diagnostic test accuracy studies"**.** As results of meta-analyses not only depend on sources of biases but also upon bibliographic search particularities, it would increase applicability and comprehensibility if this could be included in the guideline for example similar like in CRD guidance for undertaking reviews in health care. Jan 2009: 2.2.2.2 Database searching https://www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/	x☐**major** ☐minor ☐linguistic	As stated in the guideline, issues relating to systematic reviews and meta-analysis that are not restricted or unique to diagnostic test accuracy studies will not be addressed in the guideline. These issues include: bibliographic searching and study types. So, a discussion of the bibliographic search is beyond the scope of this guideline. However, the reference to the CRD guidance has been added alongside the reference to Cochrane.

			2_2_DIAGNOSTIC_TESTS.htm		
11	17	430	"It should be noted that the EUnetHTA guidelines were developed for the relative effectiveness assessment of pharmaceuticals. However, the principles contained in the guidelines are also relevant to the meta-analysis of diagnostic test accuracy studies". There are many cases in the diagnostics field specially IVD (like biomarkers essays) for which the previous statement is not accurate, and a **fit for purpose approach for diagnostic testing** is required. At least, it should be stated exactly which principles of the REA guidelines do apply to diagnostics: This statement is by far too broad. Lee J. Method validation and application of protein biomarkers: basic similarities and differences from biotherapeutics. Byoanalisis 2009. http://www.future-science-group.com/_img/pics/Method_validation_and_application_of_protein....pdf	☐**Xmajor** ☐minor ☐linguistic	We have clarified that the "extent to which the principles contained in the guidelines are also relevant to the meta-analysis of diagnostic test accuracy studies will depend on the nature of the diagnostic test being evaluated. In general, the guideline on the applicability of evidence is most likely to be of use when pooling data from diagnostic test accuracy studies."
12	19	454	"Separate random-effects meta-analyses of sensitivity and specificity: This approach may be appropriate when there is evidence of no correlation between sensitivity and specificity across studies". But this is opposed by Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy , Chapter 10 Analysing and Presenting Results , 2010, where it says	☐**Xmajor** ☐minor ☐linguistic	We do state that separate pooling "may be appropriate when there is evidence of no correlation", rather than advocating this approach. We have reworded the text to state: "This approach has been suggested for situations when there is evidence of no correlation between sensitivity and specificity across studies. However, situations where plausibly there is no correlation may be highly unlikely in practice." The text on pooling likelihood ratios already notes the

			"Methods that are not routinely included in Cochrane re views are commonly encountered in the literature for diagnostic meta-analysis**. Separate pooling of sensitivity and specificity** estimates fails to account for the trade-off between sensitivity and specificity, which m ay lead to underestimates of test accuracy (Deeks 2001). Similarly **separate pooling of likelihood ratios** ignores correlations between positive and negative likelihood ratios, and theoretically c an produce estimates which are impossible (Zwinderman 2008)."		difficulties and indicates that, should pooled likelihood ratios be required, they can be computed from the other methods which are advocated.
13	20	514	It is noted that "the following parameter definitions include a common covariate Z which affects both parameters". However, it is not clear how those covariates can be included in the model". On the contrary, for Bivariate model, the document explicitly explains in line 533-535, how covariates can be included in the model. Also, on page 29, line 794 onwards, it mentions both of these two models can incorporate study level covariates. So it might be good to mention in line 514 how to incorporate those covariates for HSROC model.	☐ major ☒ minor ☐ linguistic	Clearly the guideline is not intended to be a comprehensive how-to document for practitioners. Hence we have tried to keep the detail to a level that gives some indication of how the methods work without overwhelming the reader. How covariates are incorporated into the model may also differ by software, and the user is advised to look at the software documentation in the first instance.
14	29	754	"Simulation studies have, however, indicated that the effect of publication bias on meta-analytic estimates of the Diagnostic Odds Ratio (DOR) is not likely to be large.[40] It has been demonstrated that the unique features of the test accuracy study make the application of the Begg,	☐xmajor ☐minor ☐linguistic	We have added the suggested text regarding interpretation of funnel plot tests for publication bias. The text now reads: "It is possible that publication bias may be more prevalent in studies of test accuracy than in studies of clinical effectiveness.38 There are a number of approaches available for estimating funnel

			Egger, and Macaskill tests of funnel plot asymmetry potentially misleading.[40] An alternative approach uses funnel plots of (natural logarithm (ln) DOR) vs. (1/Veffective sample size) and tests for asymmetry using related regression or rank correlation tests.[40] It should be noted that the power of all statistical tests for funnel plot asymmetry decreases with increasing heterogeneity of DOR. It should also be noted that factors other than publication bias, for example aspects of study quality and population characteristics, may be associated with sample size. Given the limitations of current knowledge, to ignore the possibility of publication bias would seem unwise, however, its assessment in reviews of test accuracy is complex." CRD guidance for undertaking reviews in health care. Jan 2009. Chapter 2: Clinical Tests https://www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/2_2_DIAGNOSTIC_TESTS.htm "There are a number of approaches available for estimating funnel plot asymmetry, each of which may give different results in a given context.": It would improve applicability if they were mentioned and briefly described, with main advantages & disadvantages.		plot asymmetry, each of which may give different results in a given context. The unique features of the test accuracy study make the application of the Begg, Egger, and Macaskill tests of funnel plot asymmetry potentially misleading for typical DOR values.39 Alternative funnel plots using the natural logarithm DOR and functions of the effective sample size may be useful for evaluating publication bias. 39 The power of any of these statistical tests for funnel plot asymmetry decreases with increasing heterogeneity of DOR. Other factors may also be associated with sample size and hence may impact on the results of publication bias tests. Furthermore, where the number of included studies is small, the statistical methods available may be underpowered to detect asymmetry. As such, funnel plot asymmetry should be used but interpreted with caution.37"
15	31	877	Applicability would improve with a deeper explanation of individual patient data analysis techniques.	☐major ☐**Xminor** ☐linguistic	It is not clear that IPD can be included in a random effects meta-analysis or in the alternative approaches. For discrete factors, subgroup analyses may be possible, but simply adding summary statistics (e.g. by using the covariate Z) into the meta-analysis might be hampered by ecologic bias due to ignoring correlation. Until such time as there is clear evidence of the appropriateness and applicability of IPD analysis in diagnostic test accuracy meta-analysis, we will not expand the

					corresponding section in the guideline.
16	32	898-899	The cited publication by Altman does not mention a "strong risk of publication bias", but rather "Negative (non-significant) results may not be reported (publication bias)" (page 226). Is recommended to cite more closely according to the original source.	☐**xmajor** ☐minor ☐linguistic	Your quote comes from the short BMJ paper by Altman, rather than book chapter quoted in the guideline. The chapter states (page 230): "It is probable, and is supported by anecdote, that there is considerable publication bias, such that studies showing a strong (often statistically significant) prognostic ability are more likely to be published." However, your point is well made and we have edited the test to read: "More so than diagnostic test accuracy studies, there is a relatively high risk of publication bias."
17	32	915	The HuGE Review Handbook could be mentioned as an example of possible standarization of reporting for Meta-analyses of gene-disease association studies (www.med.uottawa.ca/.../HuGE_Review_Handbook_V1_0.pdf).	☐major ☐**xminor** ☐linguistic	It is not clear where this would most appropriately sit in the document, because HuGE refers to meta-analysis of genetic association studies and not to diagnostic studies.
18	33-34	Whole section (2.7)	HSROC model can have convergence problems when the sample sizes are small or too much heterogeneity exists. It is important to examine the validity of the model.	☐ major ☒ minor ☐ linguistic	Text has been added to the Software description section: Application of the HSROC model can be associated with convergence problems when the sample sizes are small or there is too much heterogeneity. It is important to examine the validity of the model and the software used should support such investigation.
19	34	994	It would be best to add links for software download.	☐major ☐**xminor** ☐linguistic	We have a preference not to given that URLs for free software often change. In any event, the listed packages are easy to locate using a standard internet search engine.
20	35	1000	"Diagnostic test accuracy is not a measure of clinical	☐**xmajor**	Such a recommendation could only follow from inclusion and discussion in the main text. The use of

			effectiveness and improved accuracy does not necessarily imply improved patient outcomes" A similiar sentence as follows could be incorporated: "Measurements of intermediate outcomes such as the **impact of test results on patient management** could be included in the analysis".	☐xminor ☐linguistic	intermediate outcomes was outside the scope of the guideline, and hence they are not formally discussed in the document. Suggesting their inclusion in analysis would have to be supported by guidance on what sorts of outcomes could be included, and how they might be incorporated.

Karsten Berndt, Chair of EDMA HTA Task Force, EUnetHTA Stakeholder Advisor in cooperation with SAG counsel: Torsten Strunz-McKendry

Chrissie Fletcher MSc. BSc. CStat CSci, Executive Director & Regional Head, Global Biostatistical Science (GBS); Amgen Ltd