Author's response to reviews

Title: Conducting Systematic Reviews of Diagnostic Studies: Didactic Guidelines

Authors:

Walter L Deville (w.deville@nivel.nl)
Frank Buntinx (frank.buntinx@med.kuleuven.ac.be)
Lex M Bouter (lm.bouter.emgo@med.vu.nl)
Victor M Montori (montori.victor@mayo.edu)
Henrica CW de Vet (hcw.devet.emgo@med.vu.nl)
Danielle DAWM van der Windt (dawm.van_der_windt.emgo@med.vu.nl)
P D Bezemer (pd.bezemer.biostat@med.vu.nl)

Version: 6 Date: 22 May 2002

PDF covering letter
Amsterdam, May 21 2002

Dear Editor,

We are pleased to resubmit our revised manuscript together with our comments to both referees who reviewed our manuscript.

We adapted the manuscript also according the 30 point checklist provided by your office.

We are hopeful that the referees will agree with our answers to their comments.

Sincerely,

For all authors,
Walter Devillé
MD, PhD
Answer to comments of reviewer Dr J Mant

We thank Dr. Jonathan Mant for his constructive review of our guidelines. We were happy to read that he had no major problems and found the review didactical.

We have the following answers to his comments:

1. P5: Inclusion criteria: an introductory sentence here would help prepare the reader -it is not clear that the heading refers to how to decide what papers to include.

   Answer: We changed the subheading as suggested: ‘Criteria for inclusion of studies’ (page 6, line 3).

2. P6: methodological quality: It would be useful to discuss the problems with case control studies here and perhaps refer to incorporation bias (i.e. is result of diagnostic test incorporated into gold standard test?)

   Answer: The problem of case-control studies is referred to on page 7 where we mention the relative overestimation of diagnostic accuracy in case-control studies and refer to the paper of J. Lijmer for more details. In addition, in response to the comments of the reviewer, we have briefly explained the main sources of potential bias in case control studies with a reference. (page 9, lines 15-16)

   As to incorporation bias we added the following: “In some conditions results of the index test may be incorporated into the diagnostic criteria, what may lead to incorporation bias and overestimation of accuracy.” (page 9, lines 6-8)

3. P11 : Do the authors have any advice (and why) as to which of the different measures of test accuracy should be reported? The DOR needs more explanation than simply its formula, as many (myself included!) will not be familiar with it. What happens to the DOR if either sensitivity or specificity are 100%?!

   Answer: We would recommend to present at least sensitivity and specificity of the original papers in a systematic review, and have added this advice to our guidelines (page 12, lines 15-16). Furthermore, we have added the following to explain the meaning of a diagnostic odds ratio: “The DOR is a measure for the discriminative power of a diagnostic test: the ratio of the odds of a positive test result among the diseased to the odds of a positive test result among the non-diseased. For the calculation of the DOR based on the study-specific sensitivity and specificity we refer to Littenberg [15], Midgette [16] or Moses [17] The potential problems associated with sensitivities and specificities of 100% are solved by adding 0,5 to all cells of the diagnostic 2x2 table.” (Page 12, lines 19-21; page 13, lines 1-4)

4. P12: Give a worked example of how to test for the presence of a cut off point effect.

   Answer: We have added an example on how to test for the presence of a cut off point effect. (page 14, lines 10-23)

5. P13: Odd example to use looking at positive predictive value to test for heterogeneity, since it is known that predictive value is a consequence of prevalence. Better to use an example where there is heterogeneity in sensitivity or specificity.
Answer: We agree with the reviewer on this point. In the revised manuscript we refer to the example given in Figure 1, that demonstrates heterogeneity in sensitivity and specificity of test results (page 15, lines 13-15).

6. P17: Need to expand on meta-regression, what it is and what it is for. The last sentence of this paragraph (A problem that is encountered in diagnostic research) is not clear, and could be re-written.

Answer: We added a sentence to clarify the term meta-regression: “This type of analysis is usually referred to meta-regression, and refers to (multivariate) regression of the summary estimate DOR of primary research as the dependent variable, and characteristics of the included studies as independent variables.” (page 19, lines 12-15). We rewrote the last sentence into: ‘This may be a problem when comparing the pooled accuracy of different tests, and has not yet been solved’ (page 19, lines 23-24).

7. P19: Data presentation: what about confidence intervals? (referred to in the example, but not stated as important 'up-front'.

Answer: We have stressed the importance of presenting confidence intervals on page 13, line 5 and line 9.
**Answer to comments of reviewer Dr Carolyn Rutter**

We would like to thank dr. Carolyn Rutter for her critical and detailed comments on our review, and are happy that she approves the general idea and finds the format reasonable. Below we discuss her comments in detail.

1. **Section on Search Criteria:**
   Last sentence of first paragraph: what other resources would be used to extend the search?

   **Answer:** With these resources we referred to reference checking and correspondence with experts, which is explained in subsequent paragraphs. We have clarified this in the revised manuscript (page 4, line 11).

2. Given changes in technology, how relevant are older publications? Definition of a reasonable timeframe is important, because test characteristics are likely to change as the test is developed then used at academic institutions and finally incorporated into general clinical practice.

   **Answer:** We are very well aware of all changes that can occur within a certain timeframe. That is why we advocate to take note of all information describing the variation in patient population, clinical setting and level of care. Indeed, also the timeframe has to be taken into account because of technical improvements, changing performance and changing indications for the test. It should be explored whether time of data collection can explain differences in accuracy of the diagnostic test across studies. If test characteristics have changed over the years as a result of changing methods or technological evolution, one has to consider the inclusion of only studies that used the new version of the testing. We have mentioned the importance of time of data collection in the revised manuscript and added a new reference 33, (page 6, lines 20-22; page 7, line 1). Nevertheless, as in other domains of clinical research, and as is advocated in the Cochrane Collaboration, we would like to recommend including all research published in a systematic review of diagnostic research. In the analysis of the review the association with time and other study characteristics can be studied and explained if present.(page 6, lines 20-22; page 7, lines 1-3)

3. **Section on Inclusion Criteria:**
   Do the authors mean 'spectrum bias' when they refer to 'selection bias'? or do they mean 'publication bias'? Either way, a one sentence definition of selection bias would be helpful.

   **Answer:** Selection bias occurs as a result of a specific way of sampling in stead of a random sample. Thus by selection bias we refer to the sampling procedure, which is often insufficiently described in primary research. Spectrum bias may be, indeed, a part of selection bias. Spectrum bias may, for example, occur as a consequence of the referral filter used in a specific system of health care. We changed the text into: “Although large samples are no guarantee against selective patient sampling, small samples seldom result from a consecutive series of patients or a random sample, but often constitute a convenience sample. Small samples are, therefore, very vulnerable to selection bias” (page 7, lines 9-12).

4. **Section on Heterogeneity**
   The first paragraph is confusing. The first sentence defines heterogeneity in terms of study characteristics. Later in the paragraph, the authors discuss tests for heterogeneity in sensitivity and specificity. The authors need to better delineate the ideas of heterogeneity in study characteristics and heterogeneity in study outcomes.
Answer: There is indeed a difference between ‘clinical heterogeneity’, ‘methodological heterogeneity’ and ‘statistical heterogeneity’ (Thompson, Lancet 1991 and BMJ 1994). Clinical and methodological heterogeneity may result in statistical heterogeneity. Existing statistical heterogeneity may be explained by looking at the factors explaining clinical heterogeneity. We have modified the text (page 13, lines 7-11).

5. The discussion of heterogeneity in outcomes could be improved by adding a brief description of ROC analysis, that is, how sensitivity and specificity vary together when tests are continuous or ordinal.

Answer: We have briefly mentioned the ROC curve on page 14, line 2-4).

6. I am not convinced that the proposed method for testing for a cut-off effect works. This seems to be a way of screening for difference that arise solely due to differences in cut-points across studies. The authors should explain that this is an ad-hoc approach or provide a reference if it exists.

Answer: We have added a reference to the revised manuscript as suggested by the reviewer (page 14, line 7). We repeated this approach in several meta-analyses, and it seemed to work at weaker correlations between sensitivity and specificity (see example added, page 14, lines 10-23). We agree with dr. Rutter when she likes to see more evidence for this approach, and have added a remark on the limited evidence for this approach to the text of the manuscript (page 15, lines 1-2).

7. Section on Dealing with Heterogeneity
   It would be useful to describe heuristically how the random effect model works, and to compare the fixed and random effect models. There continues to be confusion about heterogeneity. For example, clarify the meaning behind the last paragraph on page 15 (i.e., what do "parameters" and "their results" refer to?) It is not clear why pooling is a last resort, given the work that has been done on meta-analytic mode's for diagnostic tests. In this case, heterogeneity in study characteristics can be seen as a benefit, since it allows one to examine the effect of these characteristics on estimated test performance.

Answer: We would like to plea for practical guidelines, that can be read and used by clinical researchers. The problems in developing systematic reviews are not avoided, but we try to avoid going into details in all statistical issues. Clinical readers are advised to look for the assistance of a biostatistician. (sentence added page 3, lines 11-12). Therefore, we feel that a heuristically description of random and fixed effect models is outside the scope of this paper. We have added references for interested readers (page 17 line15, reference no 20 and 42).

8. Section on Statistical Pooling
   Separate pooling of sensitivity and specificity is generally a bad idea and should not be recommended. "Meta-regression" is not a commonly used phrase, and should probably be replaced with some other phrase, such as "meta-analysis" or "meta-analytic regression mode's".

Answer: Although it may rarely occur, we think that sensitivity and specificity can be pooled in case of homogeneity (perhaps eventually in subgroups). This is not trivial as the meaning and interpretation of a diagnostic odds ratio remains difficult for clinical practice. We added a sentence to clarify the term meta-regression: “This type of analysis is usually referred to meta-regression, and refers to (multivariate) regression of the summary estimate DOR of primary research as the dependent variable, and characteristics of the included studies as independent variables” (page 19, lines 12-15).
9. Re: pooling of ROC curves via weighted linear regression: please provide additional information or references describing weighted regression models for continuous test outcomes and ordinal regression models for ordinal tests. Additional references I description are also needed for the random effects models based on ROC parameter estimates.

Answer: We have added references 19 and 44 in a new sentence, in which weighted regression models for continuous or ordinal test results are explained (page 20, lines 8-10):

“If primary studies present ordinal test results and they use the same number of categories, ROC curves can be constructed for each study and pooled as below or by using ordinal regression methods [19, 44].” Furthermore, reference no. 45 mentioned on page 21 refers to the use of random effects models based on ROC parameter estimates.

10. Discussion Section
What is the statement "the studies are often poorly reported..." based on and is this related to the issues surrounding methodological quality. Please make clear what is based on observations/ anecdote and previously reported work.

Answer: With this statement we meant to explain that primary diagnostic research often lacks an adequate description of research methods, test characteristics and study population. We have clarified this statement and added references (page 23, lines 10 and 15)

11. What does this sentence mean? "The reader should remember that evidence about the influence of validity of studies on diagnostic tests is still limited.” The supporting references seem to be related to the quality of individual studies.

Answer: With validity we refer to aspects of internal and external validity of the primary research that may be associated with differences in results of diagnostic accuracy across studies. We adapted the sentence as follows: “.. that evidence about the effect of different aspects of internal or external validity on the results of diagnostic accuracy is still limited” (page 24, lines 7-8).

12. The statement that "any minimum set of methodologic criteria is largely arbitrary" seems extreme. However, these standards may vary with the tested condition. For example, existence of reference standard is a minimum standard, though the criteria for a good reference standard will vary by disease. While standards vary, I do not believe they are "arbitrary".

Answer: As we mentioned earlier, we feel that we have arguments supporting this statement, as the evidence of specific validity criteria on the accuracy of diagnostic tests is still limited. We have clarified this sentence: “Consequently, it is difficult to recommend a strict set of methodological criteria at this moment, recognizing that there is, as yet, insufficient evidence to support the use of any minimum set of criteria” (page 24, lines 8-10).