Reviewer's report

Title: Didactic Guidelines for Conducting Systematic Reviews of Studies Evaluating the Accuracy of Diagnostic Tests

Authors:

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

Version: 3 Date: 5 Mar 2002

Reviewer: Dr Carolyn Rutter

Level of interest: not specified

Advice on publication: Unable to decide on acceptance or rejection until the authors have responded to the compulsory revisions

This article provides a 'how-to guide' for systematic review of diagnostic testing. The general idea is good and the format is reasonable. The guidelines presented cover most the issues that need to be addressed. However, there are problems with this paper, described below, that need resolution before publication.

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

Given changes in technology, how relevant are older publication? 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.

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.
Page 6, para 2: Characteristics other than sample size might be more strongly associated with spectrum bias, for example, use of a convenience sample.

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.
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.

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

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 models 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.

Section on Statistical Pooling
Separate pooling of sensitivity and specificity 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 models".

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 / description are also needed for the random effects models based on ROC parameter estimates.

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.

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

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".

Competing interests:
None declared.