

Research article

## Reporting of measures of accuracy in systematic reviews of diagnostic literature

Honest Honest\* and Khalid S Khan

Address: Academic Dept. of Obstetrics & Gynaecology, Birmingham Women's Hospital, Birmingham B15 2TG, United Kingdom

E-mail: Honest Honest\* - h.honest@bham.ac.uk; Khalid S Khan - k.s.khan@bham.ac.uk

\*Corresponding author

Published: 7 March 2002

Received: 5 September 2001

*BMC Health Services Research* 2002, **2**:4

Accepted: 7 March 2002

This article is available from: <http://www.biomedcentral.com/1472-6963/2/4>

© 2002 Honest and Khan; licensee BioMed Central Ltd. Verbatim copying and redistribution of this article are permitted in any medium for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** There are a variety of ways in which accuracy of clinical tests can be summarised in systematic reviews. Variation in reporting of summary measures has only been assessed in a small survey restricted to meta-analyses of screening studies found in a single database. Therefore, we performed this study to assess the measures of accuracy used for reporting results of primary studies as well as their meta-analysis in systematic reviews of test accuracy studies.

**Methods:** Relevant reviews on test accuracy were selected from the *Database of Abstracts of Reviews of Effectiveness* (1994–2000), which electronically searches seven bibliographic databases and manually searches key resources. The structured abstracts of these reviews were screened and information on accuracy measures was extracted from the full texts of 90 relevant reviews, 60 of which used meta-analysis.

**Results:** Sensitivity or specificity was used for reporting the results of primary studies in 65/90 (72%) reviews, predictive values in 26/90 (28%), and likelihood ratios in 20/90 (22%). For meta-analysis, pooled sensitivity or specificity was used in 35/60 (58%) reviews, pooled predictive values in 11/60 (18%), pooled likelihood ratios in 13/60 (22%), and pooled diagnostic odds ratio in 5/60 (8%). Summary ROC was used in 44/60 (73%) of the meta-analyses. There were no significant differences in measures of test accuracy among reviews published earlier (1994–97) and those published later (1998–2000).

**Conclusions:** There is considerable variation in ways of reporting and summarising results of test accuracy studies in systematic reviews. There is a need for consensus about the best ways of reporting results of test accuracy studies in reviews.

### Background

The manner in which accuracy of clinical tests is mathematically summarised in the biomedical literature has important implications for clinicians. Appropriate accuracy measures would be expected to sensibly convey the meaning of the study results with scientifically robust statistics without exaggerating or underestimating the clinical sig-

nificance of the findings. Lack of use of appropriate measures may lead authors of primary accuracy studies to draw biased conclusions.[1] In systematic reviews of test accuracy literature, there are many ways of synthesising results from several studies, not all of which are considered to be scientifically robust. For example, measures such as sensitivity and specificity commonly used in primary studies

are not considered suitable for pooling separately in meta-analysis.[2] Variations in reporting of summary accuracy and use of inappropriate summary statistics may increase the risk of misinterpretation of clinical value of tests.

A recent study evaluated a small sample of meta-analytical reviews of screening tests to demonstrate the variety of approaches used to quantitatively summarise accuracy results.[3] This study confined itself to a limited Medline search. It exclusively examined meta-analytical studies so reviews not using quantitative synthesis were excluded. It did not look at accuracy measures used to report results of primary studies separately from those used for meta-analyses. In order to address these issues, we undertook a comprehensive search to survey systematic reviews (with and without meta-analysis) of test accuracy literature to assess the measures used for reporting results of included primary studies as well as their quantitative synthesis.

**Methods**

We manually searched for relevant reviews in the *Database of Abstracts of Reviews of Effectiveness* (DARE).[4] In order to limit the impact of human error inherent in manual searching, we complemented it with electronic searching. DARE was searched electronically with word variants of relevant terms (diagnostic, screening, test, likelihood ratio, sensitivity, specificity, positive and negative predictive value) combined using OR. From 1994 to 2000 DARE[4] has identified 1897 reviews of different types by regular electronic searching of several bibliographic databases, hand searching of key major medical journals, and by scanning grey literature (search strategy and selection criteria can be found at [http://agatha.york.ac.uk/faq2.htm]). The structured abstracts of these reviews were screened independently by the authors to identify systematic reviews of test accuracy. The full texts were obtained of those abstracts judged to be potentially relevant. Reviews addressing test development and diagnostic effectiveness or cost effectiveness were excluded. Any disagreements about review selection were resolved by consensus.

Information from each of the selected reviews was extracted for the measures of test accuracy used to report the results of the primary studies included in the review. If a meta-analysis was conducted, information was also extracted for the summary accuracy measures. The various accuracy measures are shown in Table 1. We sought the following in the primary studies: sensitivity or specificity, predictive values, likelihood ratios and diagnostic odds ratio. For meta-analysis, we sought the summary measures pooling the above results and summary receiver operating characteristics (ROC) plot or values. All extracted data were double-checked. We divided the reviews into two groups arbitrarily according to time of publication; one group covering the period 1994–97 (50 reviews) and an-

**Table 1: Measures of accuracy of dichotomous test results**

---

<b>Measures for primary studies</b>	
<b>Sensitivity (true positive rate)</b>	The proportion of people with disease who are correctly identified as such.
<b>Specificity (true negative rate)</b>	The proportion of people with disease who are correctly identified as such.
<b>Positive predictive value</b>	The proportions of test positive people who truly have disease.
<b>Negative predictive value</b>	The proportions of test negative people who truly do not have disease.
<b>Likelihood ratios (LR)</b>	The ratio of the probability of a positive (or negative) test result in the patients with disease to the probability of the same test result in the patients without the disease.
<b>Diagnostic odds ratio</b>	The ratio of the odds of a positive test result in patients with disease compared to the odds of the same test result in patients without disease.
<b>Measures for meta-analysis</b>	
<b>Summary sensitivity, Specificity, predictive values, likelihood ratios, and diagnostic odds ratio</b>	Pooling of the above accuracy measures obtained from multiple primary studies (usually averaged and weighted according to size of individual studies).
<b>Summary receiver operating characteristics curve (ROC)</b>	A method of summarising the performance of a test as found in multiple primary studies, which takes into account the relationship between sensitivity and specificity.

---

other covering 1998–2000 (40 reviews). This allowed us to assess whether there were any significant differences in measures being used to report test accuracy results among reviews published earlier and those published later. As the approaches to summarising results are not mutually exclusive, we evaluated and reported the most commonly used measures and their most common combinations. We used chi-squared statistical test for comparison of differences between proportions.

**Results**

Of the abstracts available in DARE, 150 were considered to be potentially relevant. Excluding reviews that addressed test development and diagnostic effectiveness or cost, 90 reviews of test accuracy were left for inclusion in our survey. There were 45 reviews of dichotomous test results, 42 reviews of continuous results dichotomised by the original authors, and 3 reviews that contained both result types. Meta-analysis was used in 60/90 (67 %) reviews, 50 in 1994–97 and 40 in 1998–2000. (See Additional File: BMC\_IncludedRefList\_04032002 for a complete listing of the 90 reviews included in our study).

**Table 2: Measures of test accuracy reported in review of diagnostic literature (1994–2000)**

Measures of test accuracy	Time periods				
	1994–1997		1998–2000		p-value**
	n	% (95% confidence interval)+	n	% (95% confidence interval)+	
Included primary studies	50*		40*		
Sensitivity or specificity	35	70 (55–82)	30 <sup>a</sup>	75 (59–87)	0.77
Predictive values	13 <sup>b</sup>	26 (15–40)	13 <sup>c</sup>	33 (19–49)	0.66
Sensitivity, specificity and predictive values	12	24 (13–38)	12	30 (17–47)	0.69
Likelihood ratios	9	18(9–31)	11	28 (15–44)	0.41
Diagnostic odds ratios	0	0 (0–7)	0	(0–9)	-
<b>Meta-analysis</b>	<b>38*</b>	<b>76(62–87)</b>	<b>22*</b>	<b>55 (38–71)</b>	<b>0.16</b>
Independently pooled sensitivity or specificity	22	58 (41–74)	13 <sup>d</sup>	62 (36–79)	0.86
Pooled predictive values	9 <sup>e</sup>	24(11–40)	2 <sup>f</sup>	10(1–29)	0.29
Pooled likelihood ratios	5	13 (4–28)	8	38 (17–59)	0.08
Pooled diagnostic odds ratios	5	13 (4–28)	0	0(0–15)	0.20
Summary ROC plot or values	23	61 (43–76)	11	52 (28–72)	0.60

\* numbers do not add up to totals because some reviews used more than one measures of accuracy; \*\* chi sq. test with Yates' correction; + Exact (Clopper-Pearson) 95% confidence interval a includes study that only reported either sensitivity or specificity; b,c. includes study that only reported either positive or negative predictive value; d includes meta-analyses that only reported either pooled sensitivity or specificity; e,f. includes meta-analyses that only reported either pooled positive or negative predictive values

As shown in Table 2, sensitivity or specificity was used for reporting the results of primary studies in 65/90 (72%) reviews, predictive values in 26/90 (28%), and likelihood ratios in 20/90 (22%). For meta-analysis, independently pooled sensitivity or specificity was used in 35/60 (58%) reviews, pooled predictive values in 11/60 (18%), pooled likelihood ratios in 13/60 (22%), and pooled diagnostic odds ratio in 5/60 (8%). Summary ROC was used in 44/60 (73%) of the meta-analyses. There were no significant differences between reviews published earlier and those published later as shown in Table 2.

**Discussion**

Our study showed that sensitivity and specificity remain in frequent use, both for primary studies and for meta-analyses over the time period surveyed. Sensitivity and specificity are considered inappropriate for meta-analyses, as they do not behave independently when they are pooled from various primary studies to generate separate averages.[2] In our survey, separate pooling of sensitivities or specificity was used frequently in meta-analyses where summary ROC would have been more appropriate. [5–7].

Our findings about reporting of summary accuracy measures in meta-analyses are different to those reported previously.[3] We found a higher rate of use of summary ROC, though use of independent summaries of sensitivity, spe-

cificity and predictive values were similar. These differences may be due to differences in searching strategies (databases and time frames) and selection criteria. Our search was more recent and comprehensive, using DARE[4], which has covered seven different databases (Medline, CINAHL, BIOSIS, Allied and Alternative Medicine, ERIC, Current Contents clinical medicine and PsycLIT), and hand-searched 68 peer-reviewed journals and publications from 33 health technology assessment centres around the world since February 1994. Moreover, as we did not restrict our selection to meta-analytical reviews only, we were able to examine reviews summarising accuracy results of primary studies without quantitative synthesis, which constituted 33% (30/90) of our sample. Therefore, compared to the previous publication on this topic,[3] our survey provided a broader and more up-to-date overview of the state of reporting of accuracy measure in test accuracy reviews.

**Conclusions**

The use of inappropriate accuracy measures has the potential to bias judgement about the value of tests. Of the various approaches to reporting accuracy of dichotomous test results, likelihood ratios are considered to be more clinically powerful than sensitivities or specificities.[8] Crucially, it has been empirically shown that authors of primary studies may overstate the value of tests in the ab-

sence of likelihood ratios.[1] There is also evidence that readers themselves may misinterpret test accuracy measures following publication.[9] It is conceivable that the problem of inconsistent usage of test accuracy measures in published reviews, as found in our survey, may contribute to misinterpretation by clinical readership. The reason for variation in reported accuracy measures may, in part, be attributed to a lack of consensus regarding the best ways to summarise test results. It is worth noting that despite authoritative publications about appropriate summary accuracy measures in the past,[5,7,10] (we have only quoted a few references) inconsistent and inappropriate use of summary measures has remained prevalent in the period 1994–2000. Our paper highlights the need for consensus to support change in this field of research.

### Competing interest

None declared

### Additional material

#### Additional file

Reference list of the 90 reviews included in the survey

BMC\_IncludedRefList\_04032002

[<http://www.biomedcentral.com/content/supplementary/1472-6963-2-4-S1.pdf>]

### Acknowledgement

We wished to acknowledge J. Dinnes, J. Glanville and F. Song for their contribution in the searches and selection of the reviews for the survey.

### Funding

RCOG Wellbeing grant number K2/00

### References

1. Khan KS, Khan SF, Nwosu CR, Chien PFW: **Misleading authors' inferences in obstetric diagnostic test literature.** *Am J Obstet Gynecol* 1999, **181**:112-5
2. Shapiro DE: **Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test.** *Academic Radiology* 1995, **2**:S37-S47
3. Walter SD, Jadad AR: **Meta-analysis of screening data: A survey of the literature.** *Stat. Med* 1999, **18**:3409-24
4. NHS Centre for Reviews and Dissemination: **Database of Abstracts of Reviews of Effectiveness (DARE).** York: University of York, NHS Centre for Reviews and Dissemination 2001 [<http://ag-atha.york.ac.uk/darehp.htm>]
5. Irwig L, Macaskill P, Glasziou P, Fahey M: **Meta-analytic methods for diagnostic test accuracy.** *J Clin Epidemiol* 1995, **48**:119-30
6. Rutter CM, Gatsonis CA: **A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations.** *Stat. Med* 2001, **20**:2865-84
7. Moses LE, Shapiro D, Littenberg B: **Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations.** *Stat. Med* 1993, **12**:1293-316
8. Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group: **Users' guides to the medical literature, III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients?** *JAMA* 1994, **271**:703-7
9. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G: **Medicine Communicating statistical information.** *Science* 2000, **290**:2261-2
10. **Cochrane Methods Working Group on Systematic Reviews of Screening and Diagnostic Tests** *Screening and diagnostic tests: Recommended methods* 1996 [<http://www.cochrane.org/cochrane/sadtdoc1.htm>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMedCentral will be the most significant development for disseminating the results of biomedical research in our lifetime."

Paul Nurse, Director-General, Imperial Cancer Research Fund

Publish with **BMC** and your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours - you keep the copyright



BioMedcentral.com

Submit your manuscript here:

<http://www.biomedcentral.com/manuscript/>

[editorial@biomedcentral.com](mailto:editorial@biomedcentral.com)