Funding: PS is supported by a contract from the Agency for Healthcare Research and Quality to the Southern California Evidence-based Practice Center. The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Executive Department of Health. The Health Services Research Unit and Centre for Health Services Research are members of the MRC Health Services Research Collaboration.

Competing interests: JMG is a member of the Guidelines Advisory Committee for the National Institute for Clinical Excellence and the methodological adviser to the Scottish Intercollegiate Guidelines Network. MPE is chairman of the Guidelines Advisory Committee for the National Institute for Clinical Excellence. SHW is a member of the US Preventive Services Task Force and other practice guideline panels involved in updating.

1 Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: Potential benefits, limitations, and harms of clinical guidelines. *BMJ* 1999;318:527-30.

2 Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines. Developing guidelines. *BMJ* 1999;318:593-6.

3 Fields WS, Maslenikov V, Meyer JS, Hass WK, Remington RD, Macdonald M. Joint study of extracranial arterial occlusion. V: Progress report of prognosis following surgery or nonsurgical treatment for transient cerebral ischemic attacks and cervical carotid artery lesions. *JAMA* 1970;211:1993-2003.

4 North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *N Engl J Med* 1991;325:445-53.

5 European Carotid Surgery Trialists' Collaborative Group. MRC European Carotid Surgery Trial. Interim results for symptomatic patients with severe (70-99%) or with mild (0-29%) carotid stenosis. *Lancet* 1991;337:1235-43.

6 Platelet Receptor Inhibition in Ischemic Syndrome Management in Patients Limited by Unstable Signs and Symptoms (PRISM-PLUS) Study Investigators. Inhibition of the platelet glycoprotein IIb/IIIa receptor with tirofiban in unstable angina and non-Q-wave myocardial infarction. *N Engl J Med* 1998;338:1488-97.

7 Lincoff AM, Califf RM, Moliterno DJ, Ellis SG, Ducas J, Kramer JH, et al. Complementary clinical benefits of coronary-artery stenting and blockade of platelet glycoprotein IIb/IIIa receptors. Evaluation of Platelet IIb/IIIa Inhibition in Stenting Investigators. *N Engl J Med* 1999;341:3109-27.

8 Campbell SE, Walker AE, Grimshaw JM, Campbell MK, Lowe GDO, the TEMPEST Group, et al. The prevalence of prophylaxis for venous thromboembolism in acute hospital trusts [abstract]. *J Epidemiol Community Health* 1999;53:669.

9 Eccles M, Freemantle N, Mason J. North of England evidence-based guideline development project: summary version of guidelines for the choice of antidepressants for depression in primary care. North of England Anti-depressant Guideline Development Group. *Fam Pract* 1999;16:103-11.

10 Konstam MA, Dracup K, Baker DW, Bottorff MB, Brock NH, Dacey RA, et al. *Heart failure: evaluation and care of patients with left-ventricular systolic dysfunction. Clinical practice guideline No 11.* Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, US Department of Health and Human Services, 1994. (AHCPR publication No 94-0612.)

(Accepted 12 March 2001)

Systematic reviews in health care

Systematic reviews of evaluations of diagnostic and screening tests

Jonathan J Deeks

This is the third in a series of four articles

Imperial Cancer Research Fund/NHS Centre for Statistics in Medicine, Institute of Health Sciences, Oxford OX3 7LF

Jonathan J Deeks *senior medical statistician*

Correspondence to: J J Deeks J.Deeks@icrf.icnet.uk

Series editor: Matthias Egger

BMJ 2001;323:157–62

Tests are routinely used in medicine to screen for, diagnose, grade, and monitor the progression of disease. Diagnostic information is obtained from a multitude of sources, including imaging and biochemical technologies, pathological and psychological investigations, and signs and symptoms elicited during history taking and clinical examinations.[1] Each of these items of information can be regarded as a result of a separate diagnostic or screening "test." Systematic reviews of evaluations of tests are undertaken for the same reasons as systematic reviews of treatment interventions: to produce estimates of test performance and impact based on all available evidence, to evaluate the quality of published studies, and to account for variation in findings between studies.[2–5] Reviews of studies of diagnostic accuracy involve the same key stages of defining questions, searching the literature, evaluating studies for eligibility and quality, and extracting and synthesising data. However, studies that evaluate the accuracy of tests have a unique design requiring different criteria to appropriately assess the quality of studies and the potential for bias. Additionally, each study reports a pair of related summary statistics (for example, sensitivity and specificity) rather than a single statistic (such as a risk ratio) and hence requires different statistical methods to pool the results of the studies. This article concentrates on the dimensions of study quality and the advantages and disadvantages of different summary statistics for combining studies in meta-analysis. Other aspects, including searching the literature and further technical details, are discussed elsewhere.[6]

Summary points

Systematic reviews of studies of diagnostic accuracy differ from other systematic reviews in the assessment of study quality and the statistical methods used to combine results

Important aspects of study quality include the selection of a clinically relevant cohort, the consistent use of a single good reference standard, and the blinding of results of experimental and reference tests

The choice of statistical method for pooling results depends on the summary statistic and sources of heterogeneity, notably variation in diagnostic thresholds

Sensitivities, specificities, and likelihood ratios may be combined directly if study results are reasonably homogeneous

When a threshold effect exists, study results may be best summarised as a summary receiver operating characteristic curve, which is difficult to interpret and apply to practice

Summary statistics of diagnostic accuracy

Sensitivities and specificities
The rates of correct identification of patients with and without the disease are known as test sensitivity and test specificity, respectively.[7] For a test to be useful at ruling out a disease it must have high sensitivity, and for it to be useful at confirming a disease it must have high specificity

Likelihood ratios
Positive and negative likelihood ratios describe the discriminatory properties of positive and negative test results, respectively.[8] Likelihood ratios state how many times more likely particular test results are in patients with disease than in those without disease. Positive likelihood ratios above 10 and negative likelihood ratios below 0.1 have been noted as providing convincing diagnostic evidence, whereas those above 5 and below 0.2 give strong diagnostic evidence.[9] Likelihood ratios can be directly applied to give probabilistic statements concerning the likelihood of disease in an individual (box)

Diagnostic odds ratios
The diagnostic odds ratio is a convenient measure when combining studies in a systematic review (it is often reasonably constant regardless of the diagnostic threshold) but is difficult to apply directly to clinical practice. The diagnostic odds ratio describes the odds of positive test results in participants with disease compared with the odds of positive test results in those without disease. A single diagnostic odds ratio corresponds to a set of sensitivities and specificities depicted by a receiver operating characteristic curve

Studies of diagnostic accuracy

Studies of test performance (or accuracy) compare test results between groups of patients with and without the target disease, each of whom undergoes the experimental test as well as a "gold standard" diagnostic investigation to ascertain disease status. The relation between the test results and disease status is described using probabilistic measures, such as sensitivity, specificity, likelihood ratios, diagnostic odds ratios (box), and receiver operating characteristic curves (box).

Receiver operating characteristic curves

Receiver operating characteristic curves are used in studies of diagnostic accuracy to depict the pattern of sensitivities and specificities observed when the performance of the test is evaluated at several different diagnostic thresholds. Figure 1 is a receiver operating characteristic curve from a study of the detection of endometrial cancer by endovaginal ultrasonography.[8] Women with endometrial cancer are likely to have increased endometrial thicknesses: few women who do not have cancer will have thicknesses above a high threshold whereas few women with endometrial cancer will have thicknesses below a low threshold. This pattern of results is seen in figure 1 , with the 5 mm threshold showing high sensitivity (0.98) but poor specificity (0.59) and the 25 mm threshold showing poor sensitivity (0.24) but high specificity (0.98).
The overall diagnostic performance of a test can be judged by the position of the receiver operating characteristic line. Poor tests have lines close to the rising diagonal, whereas the lines for perfect tests would rise steeply and pass close to the top left hand corner, where both the sensitivity and specificity are 1.
Receiver operating characteristic plots are used in systematic reviews to display the results of a set of studies, the sensitivity and specificity from each study being plotted as a separate point in the receiver operating characteristic space

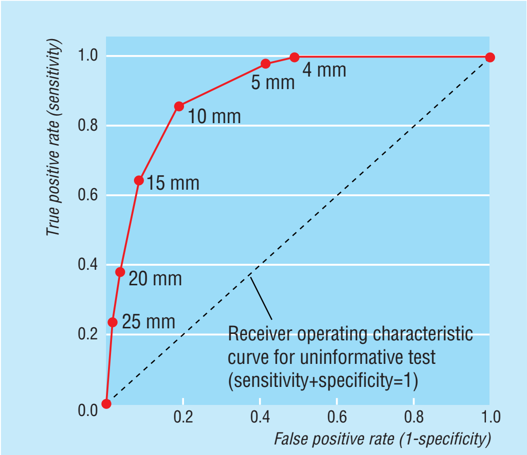


Fig 1 Receiver operating characteristic plot of endovaginal ultrasonography for detecting endometrial cancer

Dimensions of study quality

The quality of a study relates to aspects of the study's design, methods of sample recruitment, the execution of the tests, and the completeness of the study report, as summarised in table 1.[4–6 10–12]

To be reliable a systematic review should aim to include only studies of the highest quality. Systematic reviews may either exclude studies that do not meet these criteria and are susceptible to bias or include studies with a mixture of quality characteristics and explore the differences.[3 5] Whichever approach is adopted, it is essential that the quality of the studies included in the review is assessed and reported, so that appropriately cautious inferences can be drawn.

Empirical evidence

A recent empirical study evaluated which aspects of design and execution listed in table 1 are of most importance.[13] The most notable finding related to the design of the study. Studies that recruited participants with disease separately from those without disease (for example, by comparing a group known to have the disease with a group of healthy controls) overestimated diagnostic accuracy when compared with studies that recruited a cohort of patients unselected by disease status and representative of the clinical population in which the test was used. Studies that used different reference tests according to the results of the experimental test also overestimated diagnostic performance, as did unblinded studies. Omission of specific details from the report of the study was also associated with systematic differences in results.

Meta-analysis of studies of diagnostic accuracy

Meta-analysis is a two stage process involving derivation of summary statistics for each study and

Table 1 Framework for considering study quality and likelihood of bias

Study feature	Qualities sought
Sample of patients	Consecutive or randomly selected sample, recruited as single cohort unclassified by disease state, recruited from clinical setting and point in referral process where test would be used, selection and referral processes fully described, clinical and demographic characteristics fully described, complete
Reference diagnosis	Method and tests described in detail, positive and negative diagnoses clearly described, diagnosis likely to be close to truth, available for all patients, based on same tests and information in all patients, blinding procedures used to prevent knowledge of result of experimental test influencing the reference diagnosis, made before treatment commenced
Experimental test	Application of test described in detail, positive and negative test results clearly described, blinding procedures used to ensure that test is undertaken without knowledge of reference diagnosis, test undertaken before treatment commenced, results reported for all patients, including those with "grey zone" results

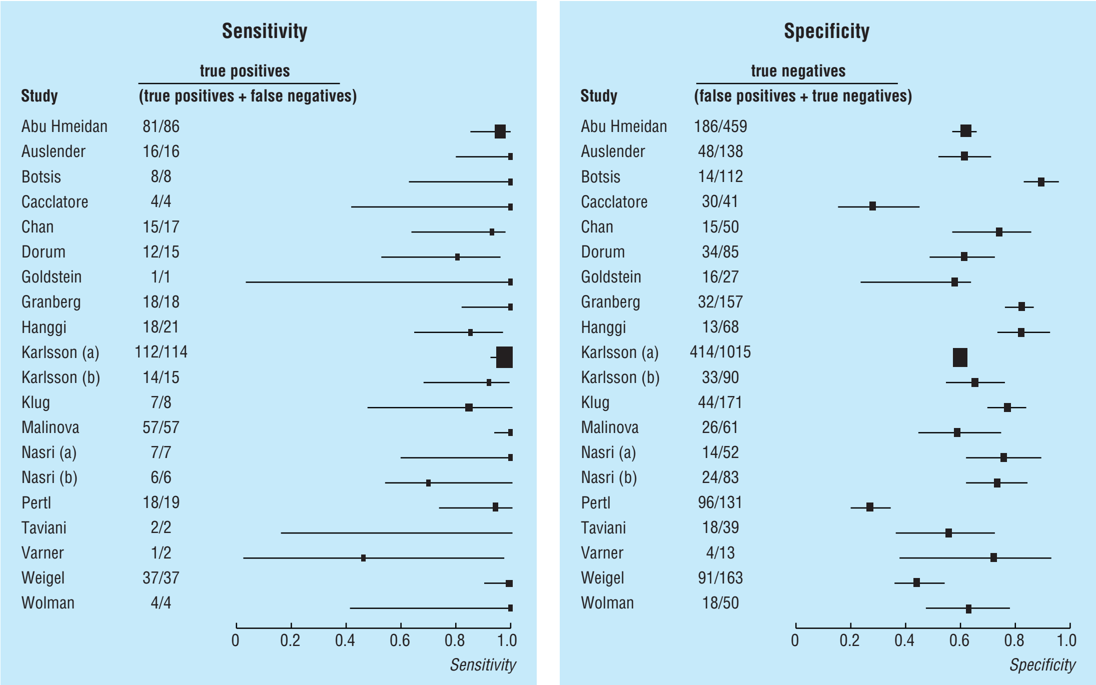


Fig 2 Estimates from 20 studies of sensitivity and specificity of measurement of endometrial thicknesses of more than 5 mm using endovaginal ultrasonography for detecting endometrial cancer.[15] Points indicate estimates of sensitivity and specificity. Horizontal lines are 95% confidence intervals for estimates. Size of points reflects total sample size

computation of a weighted average of the summary statistics across the studies.[14] I illustrate the application of three commonly used methods for pooling different summaries of diagnostic accuracy with a case study.

As with systematic reviews of randomised controlled trials, meta-analysis should be considered only when the studies have recruited from similar patient populations (it is problematic to combine studies from general practice with studies from tertiary care), have used comparable experimental and reference tests, and are unlikely to be biased. Even when these criteria are met there may still be such gross heterogeneity between the results of the studies that it is inappropriate to summarise the performance of a test as a single number.

Case study

Detection of endometrial cancer with endovaginal ultrasonography

Smith-Bindman et al published a systematic review of 35 studies evaluating the diagnostic accuracy of endovaginal ultrasonography for detecting endometrial cancer and other endometrial disorders.[15] All studies included in the review were of prospective cohort designs and used the results of endometrial biopsy, dilation and curettage, or hysterectomy as a reference standard. Most of the studies presented sensitivities and specificities at several endometrial thicknesses detected by endovaginal ultrasonography (the receiver operating characteristic curve in figure 1 is from one of these studies). The case study is based on the subset of 20 studies from this review that considered the diagnostic accuracy of endovaginal ultrasonography in ruling out endometrial cancer with endometrial thicknesses of 5 mm or less. Figure 2 shows the sensitivities and specificities for the 20 studies.

Sources of heterogeneity

The choice of meta-analytical method depends in part on the pattern of variability (heterogeneity) observed in the results. Heterogeneity can be considered graphically by plotting sensitivities and specificities from the studies as points on a receiver operating characteristic plot (fig 3). Some divergence of the results around a central point is to be expected by chance, but variation in other factors, such as patient selection and features of the study's design, may increase the observed variability.[16]

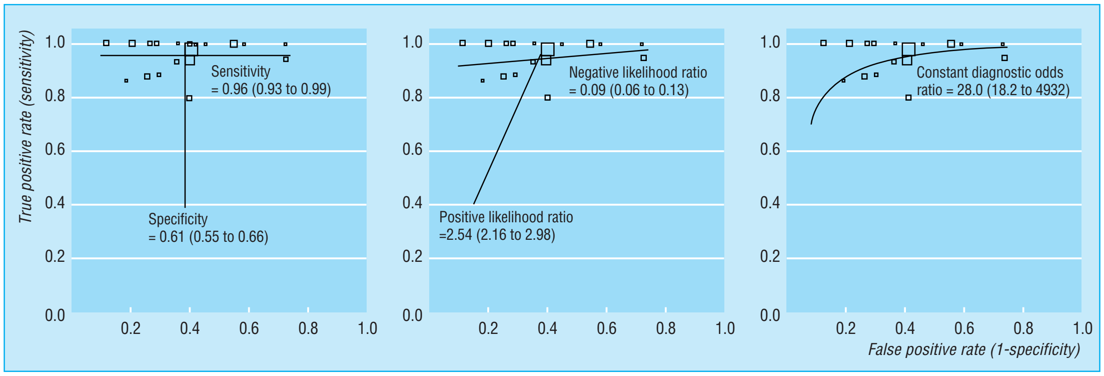


Fig 3 Receiver operating characteristic plots showing three approaches to meta-analysis of 20 studies of diagnostic accuracy of endovaginal ultrasonography for detecting endometrial cancer. Results of studies are indicated by squares. Area of squares is proportional to study sample size. Fitted lines indicate (left) average sensitivity and specificity, (centre) average positive and negative likelihood ratios, and (right) average diagnostic odds ratios. Figures in brackets are 95% confidence intervals for summary estimates

One important extra source of heterogeneity is variation introduced by changes in diagnostic threshold. Studies may use different thresholds to define positive and negative test results. Some may have done this explicitly—for example, by varying numerical cut-off points used to classify a biochemical measurement as positive or negative, whereas for others there may be naturally occurring variations in diagnostic thresholds between observers, laboratories, or machines. The choice of a threshold may also vary according to the prevalence of the disease—when the disease is rare a more extreme threshold may have been used to avoid large numbers of false positive diagnoses. Unlike other sources of variability, variation of the diagnostic threshold introduces a particular pattern into the receiver operating characteristic plot of study results, such that the points show curvature (fig 1).

If there is no heterogeneity between the studies, the best summary estimate of test performance should be a single point on the receiver operating characteristic graph. The first two methods estimate such a summary, first by pooling sensitivities and specificities then by pooling positive and negative likelihood ratios. The third method is more complex and pools diagnostic odds ratios to take account of possible heterogeneity in diagnostic threshold.

Application of a likelihood ratio

The probability of endometrial cancer in a woman with an endometrial thickness of 5 mm or less measured by endovaginal ultrasonography can be computed with Bayes' theorem[1 17]:
Post-test odds = pretest odds×likelihood ratio
Assuming that the study samples are representative, an estimate of the pretest odds can be calculated from the prevalence of endometrial cancer across the studies (13%)

$$\text{Pretest odds} = \frac{\text{prevalence}}{1\text{–prevalence}} = \frac{0.13}{0.87} = 0.15$$

Applying Bayes' theorem to the summary negative likelihood ratio:
Post-test odds = pretest odds×negative likelihood ratio = 0.15×0.09 = 0.014
and converting the post-test odds to a probability:

$$\text{Post-test probability} = \frac{\text{post-test odds}}{1 + \text{post test results}} = \frac{0.014}{1 + 0.014} = 0.013$$

we estimate that only 1.3% of women with an endometrial thickness of 5 mm or less measured by endovaginal ultrasonography will have endometrial cancer. Knowledge of other characteristics of a particular patient that either increase or decrease their prior probability of endometrial cancer can be incorporated into the calculation by adjusting the pretest probability accordingly[1]

Pooling sensitivities and specificities

The pooled estimate of sensitivity is 0.96 (95% confidence interval 0.93 to 0.99) and is depicted by the horizontal line on the receiver operating characteristic plot in figure 3 (left). The overall estimate of mean specificity is lower: 0.61 (0.55 to 0.66).

Heterogeneity is, however, clearly evident in figure 3 (left): although the study points lie reasonably close to the summary sensitivity (test for heterogeneity, P = 0.04), the results of many studies lie some distance from the summary specificity (test for heterogeneity, P < 0.001).

Regardless of the causes of the heterogeneity, the overall high estimate and relative consistency of the sensitivity results does suggest that a negative test result could be of potential clinical use in ruling out endometrial cancer. As there is heterogeneity between specificities, however, it is more appropriate to note the range of specificities (0.27 to 0.88) rather than to quote the average value of 0.61. It is difficult to draw a conclusion about test specificity: the observed values vary considerably and there is no understanding from this analysis as to the reasons for the variation.

Pooling likelihood ratios

For the case study the pooled estimate of the positive likelihood ratio was not particularly high (2.54, 2.16 to 2.98), and the values varied significantly between the studies (test for heterogeneity, P < 0.001). In figure 3 (centre) it is clear that the summary positive likelihood ratio lies some distance from many of the values. Again it is debatable whether reporting the average value of such heterogeneous results is sensible, but it is unlikely that a positive test result could provide convincing evidence of the presence of endometrial cancer as the positive likelihood ratios are all below 10 (data not shown).

The negative likelihood ratios show no evidence of significant heterogeneity (test for heterogeneity, P = 0.09), the pooled estimate being 0.09 (0.06 to 0.13), with the summary line on the receiver operating characteristic plot in figure 3 (centre) lying close to the results of most of the studies. This finding again shows that a measurement of an endometrial thickness of 5 mm or less made by endovaginal ultrasonography can provide reasonably convincing evidence to rule out endometrial cancer.

Although these conclusions concerning potential diagnostic use are similar to those obtained by pooling sensitivities and specificities, the summaries obtained by pooling likelihood ratios can be more easily interpreted and applied to clinical practice. The box describes how the summary negative likelihood ratio can be applied to estimate the probability of endometrial cancer in a woman with a negative test result.

Diagnostic odds ratios and summary receiver operating characteristic curves

If the observed heterogeneity between the studies arises due to variation in the diagnostic threshold, estimates of summary sensitivity and specificity or summary positive and negative likelihood ratios will underestimate diagnostic performance.[18] In this situation the appropriate meta-analytical summary is not a single point in the receiver operating characteristic space but the receiver operating characteristic curve itself. Methods of deriving the best fitting summary receiver operating characteristic curve are necessarily more complex.[2–5 18–21]

How is a summary receiver operating characteristic curve estimated? The simplest approach involves calculating a single summary statistic for each study—the diagnostic odds ratio (box). Each diagnostic odds ratio corresponds to a particular receiver operating characteristic curve. If the studies in a review all relate to the same curve they may have consistent diagnostic odds ratios even if they have variable sensitivities and specificities. Table 2 gives examples of diagnostic odds ratios corresponding to particular sensitivities, specificities, and positive and negative likelihood ratios.

In the case study it is possible that some of the observed heterogeneity could be explained by a threshold effect, perhaps due to differences in calibration of the ultrasound machines. The estimate of the summary diagnostic odds ratio is 28.0 (18.2 to 43.2) and is reasonably consistent across the studies (test for heterogeneity, P = 0.3), suggesting that the points indeed could have originated from the same receiver operating characteristic curve. The summary diagnostic odds ratio can be interpreted in terms of sensitivities and specificities by consulting table 2 (for example, a diagnostic odds ratio of 29 corresponds to a sensitivity of 0.95 and a specificity of 0.60 and to a sensitivity of 0.60 and specificity of 0.95) or by plotting the corresponding summary receiver operating characteristic curve (fig 3 (right)). This method does not yield a unique joint summary estimate of sensitivity and specificity: it is only possible to obtain a summary estimate of one value by specifying the value of the other. This greatly limits its clinical application.

Discussion

Systematic reviews of diagnostic accuracy have not, as yet, made the same impression on the practice of evidence based health care as have systematic reviews of randomised controlled trials. Reasons relate to reliability, heterogeneity, and clinical relevance.

Table 2 Examples of diagnostic odds ratios corresponding to particular pairings of sensitivity and specificity and positive and negative likelihood ratios

	Sensitivity						
Specificity	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**0.95**	**0.99**
0.5	1	2	2	4	9	19	99
0.6	2	2	4	6	14	29	149
0.7	2	4	5	9	21	44	231
0.8	4	6	9	16	36	76	396
0.9	9	14	21	36	81	171	891
0.95	19	29	44	76	171	361	1881
0.99	99	149	231	396	891	1881	9801

Negative	Positive likelihood ratio						
likelihood ratio	**1**	**2**	**5**	**10**	**20**	**50**	**100**
1	1	2	5	10	20	50	100
0.5	2	4	10	20	40	100	200
0.2	5	10	25	50	100	250	500
0.1	10	20	50	100	200	500	1000
0.05	20	40	100	200	400	1000	2000
0.02	50	100	250	500	1000	2500	5000
0.01	100	200	500	1000	2000	5000	10 000

Are systematic reviews of diagnostic studies reliable?

Many meta-analyses of the accuracy of diagnostic tests are hindered by the poor quality of the primary studies: most published evaluations of the accuracy of diagnostic tests having at least one flaw.[12] Headway has been made in understanding the importance of particular features of a study's design and in improving quality, but for many diagnostic tests few high quality studies have been undertaken and published.[13]

The reliability of a review also depends crucially on whether the included studies are an unbiased selection. As with all reviews, systematic reviews of diagnostic tests are susceptible to publication bias, and this may be a greater problem than for randomised controlled trials.[2 3] No investigations, however, have been conducted to estimate rates of publication bias for studies of diagnostic accuracy.

MARK OLDROYD

How useful are systematic reviews to a practising clinician?

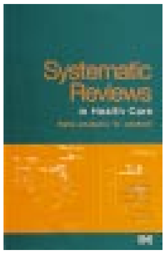
Heterogeneity of the results of studies of diagnostic accuracy is common but in itself does not prevent conclusions of clinical value from being drawn.[22] Despite heterogeneity being observed in the case study, it was still possible to draw a conclusion of clinical value—that an endometrial thickness of 5 mm or less can rule out endometrial cancer.

Diagnostic odds ratios and summary receiver operating characteristic curves are, however, often promoted as the most statistically valid method for combining test results when there is heterogeneity between studies, and they are commonly used in systematic reviews of diagnostic accuracy.[2–4] Unfortunately summary curves are of little use to practising healthcare professionals: they can identify whether a test has potential clinical value, but they cannot be used to compute the probability of disease associated with specific test outcomes. Their use is also based on a potentially inappropriate and untested assumption that observed heterogeneity has arisen through variation in diagnostic threshold. In the case study, whereas the diagnostic odds ratio was a reasonably consistent summary statistic across the studies, there was no evidence to suggest that the observed heterogeneity arose through variations in diagnostic threshold (all included studies had a 5 mm threshold for endometrial thickness). Variation in referral patterns, sample selection, and study methods may be more likely explanations for the heterogeneity. There is no clear statistical advantage in using a summary receiver operating characteristic approach to synthesise the results over pooling sensitivity and specificity or likelihood ratios unless there is a threshold effect. Empirical research is urgently required to find out whether the simpler methods for pooling sensitivities, specificities, and likelihood ratios are likely to be seriously misleading in practice and whether apparent threshold effects are really due to variations in diagnostic threshold rather than alternative sources of heterogeneity.

Are studies of diagnostic accuracy clinically relevant?

Systematic reviews of the accuracy of tests do not always answer the most clinically relevant question. New tests are often evaluated for their ability to replace or be used alongside existing tests. The important issues are comparisons of tests or comparisons of testing algorithms: these would be best addressed in properly designed comparative studies, rather than by synthesising studies of diagnostic accuracy separately for each test.

The evaluation of the diagnostic accuracy of a test is also only one component of assessing whether it is of clinical value.[23 24] Treatment interventions are recommended for use in health care only if they are shown on average to be of benefit to patients: the same criterion should also be applied for the use of a diagnostic test, and even the most accurate of tests can be clinically useless or do more harm than good. It should always be considered whether undertaking a systematic review of studies of diagnostic accuracy is likely to provide the most useful evidence of the value of a diagnostic intervention.



Systematic Reviews in Health Care: Meta-analysis in Context can be purchased through the BMJ Bookshop (www.bmjbookshop.com); further information and updates for the book are available on www.systematicreviews.com.

I thank Rebecca Smith-Bindman for providing the data for the case study.

Competing interests: None declared.

1 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*, 2nd ed. Boston: Little, Brown, 1991.
2 Irwig L, Tosteson AN, Gatsonis CA, Lau J, Colditz G, Chalmers TC, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667-76.
3 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytical methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-30.
4 Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests. *Recommended methods* [updated 6 Jun 1996]. www.cochrane.org/cochrane/sadtdoc1.htm (accessed 27 March 2001).
5 Vamvakas EC. Meta-analyses of studies of diagnostic accuracy of laboratory tests: a review of concepts and methods. *Arch Pathol Lab Med* 1998;122:675-86.
6 Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context*, 2nd ed. London: BMJ Books, 2001.
7 Bland JM, Altman DG. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 1994;308:1499.
8 Deeks JJ, Morris JM. Evaluating diagnostic tests. *Baillière's Clinical Obstetrics and Gynaecology* 1996;10:613-30.
9 Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. 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.
10 Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. How to use an article about a diagnostic test. A: Are the results of the study valid? *JAMA* 1994;271:289-91.
11 Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing the quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;4:288-95.
12 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;274:645-51.
13 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
14 Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context*, 2nd ed. London: BMJ Books, 2001.
15 Smith-Bindman R, Kerlikowske K, Feldstein VA, Subak L, Scheidler J, Segal M, et al. Endovaginal ultrasound to exclude endometrial cancer and other endometrial abnormalities. *JAMA* 1998;280:1510-7.
16 Devillé W, Yzermans N, Bouter LM, Bezemer PD, van der Windt DAWM. Heterogeneity in systematic reviews of diagnostic studies. *Proceedings of the 2nd symposium on systematic reviews: beyond the basics*. Oxford, 1999. Abstract available at on www.ihs.ox.ac.uk/csm/talks.html#p21 (accessed 27 March 2001).
17 Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in clinical medicine*, 3rd ed. New York: McGraw-Hill, 1994:26-50.
18 Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;2:37-47S.
19 Moses LE, Littenberg B, Shapiro D. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytical approaches and some additional considerations. *Stat Med* 1993;12:1293-316.
20 Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Meth Info Med* 1990;29:12-22.
21 Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytical method. *Med Decis Making* 1993;13:313-21.
22 Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematically reviewing studies of diagnostic tests. *Clin Chem Lab Med* 2000;38:577-88.
23 Deeks JJ. Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence. *Ann Oncol* 1999;10:761-8.
24 Guyatt GH, Tugwell P, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *Can Med Assoc J* 1986;134:587-94.

Endpiece

Nothing like experience

We live in an age of mass loquacity. We are all writing it or at any rate talking it: the memoir, the apologia, the cv, the cri de coeur. Nothing, for now, can compete with experience—so unanswerably authentic, and so liberally and democratically dispensed. Experience is the only thing we share equally, and everyone senses this.

Martin Amis, *All from experience*,
London: Jonathan Cape, 2000