Interpreting diagnostic accuracy studies for patient care

Susan Mallett,[1] Steve Halligan,[2] Matthew Thompson,[1] Gary S Collins,[3] Douglas G Altman[3]

[1]University of Oxford, Department of Primary Care Health Sciences, Oxford OX2 6GG, UK
[2]University College London, Centre for Medical Imaging, London, NW1 2BU, UK.
[3]University of Oxford, Centre for Statistics in Medicine, Oxford, OX2 6UD
Correspondence to: S Mallett
susan.mallett@phc.ox.ac.uk
Accepted: 17 May 2012
Cite this as: *BMJ* 2012;344:e3999
doi: 10.1136/bmj.e3999

bmj.com
Previous articles in this series
- Comparative effectiveness research in cancer screening programmes (*BMJ* 2012;344:e2864)
- Assessing the value of diagnostic tests: a framework for designing and evaluating trials (*BMJ* 2012;344:e686)
- Clinical prediction rules (*BMJ* 2012;344:d8312)

A diagnostic test accuracy study provides evidence on how well a test correctly identifies or rules out disease and informs subsequent decisions about treatment for clinicians, their patients, and health-care providers. The authors highlight several different ways in which data from diagnostic test accuracy studies can be presented and interpreted, and discuss their advantages and disadvantages.

Studies of tests that aim to diagnose clinical conditions that are directly applicable to daily practice should present test results that are directly interpretable in terms of individual patients— for example, the number of true positive and false positive diagnoses. We do not examine measures used for early experimental (exploratory) studies, in which diagnostic thresholds have not been established.

Results obtained from a diagnostic test accuracy study are expressed by comparison with a reference standard of the "true" disease status for each patient. Thus, once a clinically relevant diagnostic threshold has been established, patients' results can be categorised by the test as true positive (TP), false positive (FP), true negative (TN), and false negative (FN) (fig 1).

Diagnostic accuracy can be presented at a specific threshold by using paired results such as sensitivity and specificity, or alternatively positive predictive value (PPV) and negative predictive value (NPV) (see fig 1). Other methods summarise accuracy over a range of different test thresholds—for example, the area under the receiver operator curve (ROC AUC, see fig 1).

Despite the simplicity of the 2×2 structure, the presentation and interpretation of tests and comparisons between them are not straightforward. Graphical presentation can be highly informative, in particular an ROC plot, which is a plot of sensitivity against 1–specificity (or false positive rate). Figure 2 shows an ROC plot of test accuracy of a single test at different thresholds. ROC plots are also used within studies to compare different tests, to compare different groups of patients, and to investigate variability between different test observers (readers). ROC plots are useful in systematic reviews to present results from multiple studies.

Several concepts need to be considered carefully in the interpretation of data from a diagnostic accuracy study:

- How does accuracy change with different diagnostic thresholds?
- If paired outcomes (such as sensitivity and specificity) are compared for different scenarios, they often change in opposite directions. For example, sensitivity is often higher in one test and specificity higher in the other. Which is more important?
- What are the clinical consequences of a missed (false negative) diagnosis or a false positive diagnosis? Can these risks be presented together—for example, as a relative benefit?
- What is the best way to include disease prevalence in the summary of clinical benefit?
- Are results presented in terms of what happens to individual patients, which are often the easiest for clinicians (and their patients) to understand?[1]

SUMMARY POINTS

Diagnostic test accuracy studies should present data in a way that is comprehensible and relevant to clinicians, their patients, and healthcare providers when making clinical management decisions

The most relevant and applicable presentation of diagnostic test results allows inclusion of four key components: interpretation in terms of patients; clinically relevant values for test threshold(s); realistic disease prevalence; and clinically relevant relative gains and losses in terms of patients (that is, true positive and false positive diagnoses)

Presenting diagnostic accuracy as paired measures, such as sensitivity and specificity, or as net benefit summaries with component paired measures, allows inclusion of these four components whereas using the area under the ROC curve as a diagnostic performance measure does not

Reporting test accuracy at different thresholds

Presenting results at a single threshold

When a test has only a single threshold or cutpoint value (for instance, positive or negative for disease, such as a biopsy), results are naturally presented in pairs, usually sensitivity and specificity, or PPV and NPV (see fig 1). Although PPV (and NPV equivalently) allow easy comprehension of the probability that a patient with a positive test result has the disease, when tests are compared in the same patients it is not straightforward to use these measures because the calculation of confidence intervals is complex.[2]

Tests that yield results on a continuous scale require specification of a test threshold to define positive and negative results. Changing the threshold alters the proportion of false positive and false negative diagnoses. Figure 2 shows how the sensitivity of CA19-9 for diagnosis of pancreatic cancer increases as the threshold value is lowered from 1000 to 15 U/ml, while specificity decreases.

Presenting results at multiple thresholds

For many diagnostic tests, however, there are multiple potential thresholds at which different clinical decisions could be made, often reflecting diagnostic uncertainty. For example, the mammographic BI-RADS classification for breast screening has six categories: 0=additional imaging evaluation required; 1=negative; 2=benign findings;

Diagnostic accuracy measures
Diagnostic test results are expressed in comparison with reference standard diagnosis of disease

	Reference standard test	
	Disease positive	Disease negative
Test positive	True positive (TP)	False positive (FP)
Test negative	False negative (FN)	True negative (TN)

Paired diagnostic measures at specific thresholds (such as sensitivity and specificity)
Sensitivity is the proportion of patients (or test results) with disease correctly diagnosed using a specific threshold to define a positive test result

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Specificity is the proportion of patients (or test results) without disease correctly diagnosed

$$Specificity = \frac{TN}{(TN + FP)}$$

Other paired diagnostic measures include: positive predictive value (PPV) and negative predictive value (NPV) or positive (LR+) and negative (LR–) likelihood ratios

Single diagnostic measure across multiple test thresholds: for example, area under the ROC curve (ROC AUC)

ROC AUC is the area under the ROC curve of sensitivity v 1–specificity

The ROC AUC for each test corresponds to the probability that of two randomly chosen people, one with and one without disease, the diagnostic test will rank the person with disease with a higher suspicion of disease than the one without disease. For example, an ROC AUC of 0.7 means that of two randomly chosen people, there is a 70% chance (equivalent to a 0.7 probability) that the person with disease will be ranked with higher suspicion than the person without disease. An alternative interpretation of the ROC AUC is the average sensitivity given that all values of specificity are equally likely[27]
Other single diagnostic measures used include the H-measure[21] (where misclassification costs can be fixed) and the diagnostic odds ratio (DOR)[28]

Fig 1 | Diagnostic accuracy measures

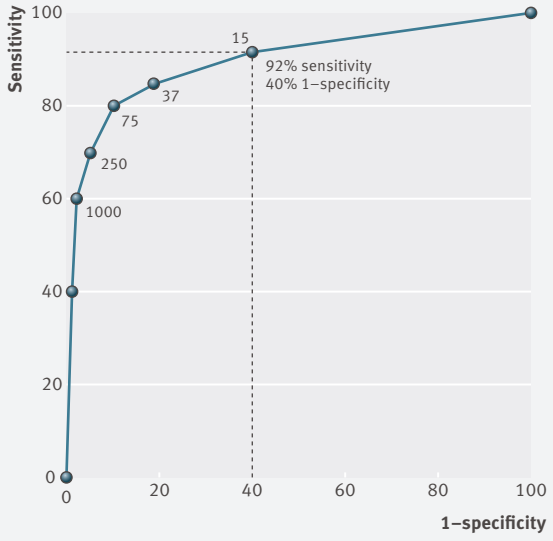

Fig 2 | ROC plot of test accuracy at different thresholds. Data from systematic review of CA19-9.[29] Threshold values are shown in U/mL. At 15 U/mL, test accuracy is 92% sensitivity and 60% specificity (1–specificity=40%)

3=probably benign finding; 4=suspicious abnormality; and 5=highly suggestive of malignancy.[3]

For many diagnostic tests there is no consensus regarding the clinically optimal threshold that separates a positive from a negative result as it is difficult to agree at which threshold it is acceptable to risk missing disease. With measures such as sensitivity and specificity, diagnostic accuracy can be reported for each test threshold relevant to the management of patients. Even then, it is important to understand that not all thresholds are equally important. For the diagnosis of breast cancer with the BI-RADS scale, the threshold between "highly suggestive of malignancy" and "suspicious abnormality" is clearly more clinically important to a patient and her doctor than the threshold between "benign" and "probably benign."

Presenting a performance measure combined across thresholds

Alternatively, diagnostic accuracy can be summarised by combining accuracy across a range of thresholds with a measure such as ROC AUC (fig 1).[4] This, however, can be a disadvantage if thresholds that are clinically relevant are combined with those that are clinically nonsensical.[5] Clinically, information is needed on how a test performs in patients at a clinically relevant threshold rather than a summary of how the test might perform across all possible thresholds.

Are false positive and false negative diagnoses equally important?

No diagnostic test is perfect and almost all tests will sometimes miss disease or indicate disease in normal patients (see FN and FP, respectively, in fig 1). False negative and false positive diagnoses, however, are rarely equally important. Missing a life threatening disease will probably be regarded by a patient (and his or her doctor) as much more important than a false positive diagnosis in a healthy patient. For example, a study of attitudes and knowledge of mammography for screening for breast cancer found that 63% of women thought that 500 or more women receiving false positive results was reasonable for each life saved.[6]

The relative importance of a false negative versus a false positive diagnosis (also called relative misclassification cost) varies according to where the test fits in the patient pathway and who is making the assessment. For example, funders or commissioners of healthcare might have a different perspective from patients or clinicians as additional false positive diagnoses will increase costs. The relative importance of additional false negative versus additional false positive diagnoses is particularly important in decisions about which of two tests is "better"—which is more important, an increase in sensitivity or an increase in specificity? To evaluate which test is better, performance needs to incorporate clinical costs.

Presenting diagnostic accuracy for patients

For diagnostic accuracy studies to usefully inform clinical practice, their results should be related to decisions regarding management of patients. Presentation in terms of individual patients is often best,[1] and formats such as animations with smiley faces have been successful.[7]

Interpretation in terms of patients is straightforward and direct for paired measures such as sensitivity and specificity, PPV and NPV, or positive and negative likelihood ratios. Sensitivity and specificity provide test accuracy in terms of patients in a population, although interpretation for an individual patient with unknown disease status is less obvious. PPV and NPV are useful to understand the probability that a patient with a given positive or negative test result has a diagnosis of disease. Positive and negative likelihood ratios are useful to understand the role of a test result in changing a clinician's estimate of the probability of disease in a patient. These paired measures can be combined into a single measure (for example, "net benefit" measure; see below), which is also easily understood, particularly when it is reported with the component paired measures.

By contrast, interpretation of a single numerical value of the ROC AUC is problematic because the summary across all thresholds is difficult to reconcile with a specific threshold for the individual patient. Also ROC AUC is hard to interpret in practice, as it is the probability that randomly selected pairs of patients, one with and one without disease, would be ordered correctly for probability of disease (see fig 1). However, patients do not walk into the clinician's room in pairs,[8] and patients want their results and diagnosis, rather than the order of their results compared with another patient.

Net benefit methods to measure diagnostic performance

Net benefit measures can provide an overall impact across changes in paired measures. For example, the weighted comparison (WC) measure[13] is an index weighting the difference in sensitivity and difference in specificity of two tests, taking into account the relative clinical cost (misclassification costs) of a false positive compared with a false negative diagnosis and disease prevalence. We note that the WC measure is similar to the net reclassification index (NRI),[14] if the latter is adapted to account for disease prevalence and relative misclassification costs.

$$WC=\Delta\text{sensitivity}+[(1-\text{prevalence}/\text{prevalence})\times\text{relative cost (FP/TP)}\times\Delta\text{specificity}]$$

What do weighted comparison values mean?

- Positive WC values indicate a net benefit
- Zero WC values show no net benefit
- Negative WC values show a net loss
- 95% confidence intervals and thresholds for clinical benefit are used to indicate significance of results. To aid interpretation, WC can be converted into an equivalent increase in true positive patients per 1000.

Example calculating WC for two biomarker tests of pancreatic cancer

Comparing two tumour marker tests for diagnosis of pancreatic cancer, CA 19-9 with 83% sensitivity and 81% specificity to CA 242 with 74% sensitivity and 91% specificity,[9] the difference in sensitivity (Δsensitivity) is 9% (equivalent to 0.09) and the difference in specificity (Δspecificity) is −10% (or −0.10). So in a population with estimated disease prevalence of 33%, and a 10-fold higher relative weighting for true positive diagnoses compared with false positive diagnoses, the WC is obtained as:

$$WC=0.09-(2\times0.1\times0.10)=0.07$$

As WC is positive there is an increased net benefit favouring CA 19-9.

To aid interpretation, WC can be converted into an equivalent increase in true positive patients per 1000, if all the benefit was focused into TP patients by calculating WC×prevalence×1000.

A WC of 0.07 converts to a benefit equivalent to 23 more true positive patient results per 1000 patients, based on actual values of 30 more patients receiving a true positive result and 66 more patients receiving a FP diagnosis, at prevalence and relative weighting as specified.

Other single diagnostic measures include: other net effect measures[10-12 15-17 30] and net reclassification index.[14]

Comparing the performance of two diagnostic tests

Three main approaches can be used to compare the diagnostic accuracy of two tests that differ depending on whether a specific test threshold is used or performance is averaged across multiple thresholds. They also vary in whether they can be interpreted in terms of patients and whether they can incorporate clinical context, such as relative weightings of false negative and false positive diagnoses and also disease prevalence. Ideally, diagnostic tests should be compared within the same patients or, if this is not practical, on randomised groups from the same population of patients. This ensures that differences in observed test results are because of the tests rather than differences in characteristics of patients or study methods.

Paired measures at specific thresholds

The first method compares two tests according to differences in paired measures such as sensitivity and specificity. For example, of two biomarker tests for pancreatic cancer, CA 19-9 with 83% sensitivity and 81% specificity can be compared with CA 242 with a sensitivity of 74% and specificity of 91%: CA 19-9 has 9% higher sensitivity, but 10% lower specificity.[9] The clinical context of these differences in sensitivity and specificity would be enhanced by using clinically relevant disease prevalence to report the difference in the actual number of patients with true and false positive diagnoses. For a given increase in sensitivity, if the prevalence of disease is twice as high, then the number of patients who receive a true positive diagnosis is doubled. Nevertheless, paired measures are transparent enough for healthcare providers or patients to incorporate their own relevant contextual information.

Summary measure at specific thresholds: net benefit methods

In the second approach, a single overall measure of diagnostic performance can be presented by using net benefit or net utility methods, calculated from test performance at a specific clinically relevant threshold.[10-17] Several of these measures are based directly on the difference in sensitivity and specificity between the two tests being compared at one[10 13 18] or more than one clinical threshold.[16 19] A single overall measure of diagnostic performance is often preferred for simplicity when guiding healthcare spending or regulatory approval decisions. These methods directly incorporate the contextual information regarding prevalence and relative importance of false negative and false positive diagnoses.

The weighted comparison (WC) net benefit measure[13] method weights differences in sensitivity and specificity between two tests by the relative clinical costs and disease prevalence (see box). With the previous example of CA 19-9 and CA 242, the net benefit is positive (weighted comparison=0.07) if CA 19-9 is used instead of CA 242, at a disease prevalence of 33%, and a 10-fold higher relative weighting of true positive diagnoses over false positive diagnoses (box). To aid interpretation, the weighted comparison can be converted to a net benefit equivalent to 23 more true positive test results per 1000 patients, based on actual values of 30 more patients receiving a true positive result and 66 more patients receiving a false positive diagnosis.

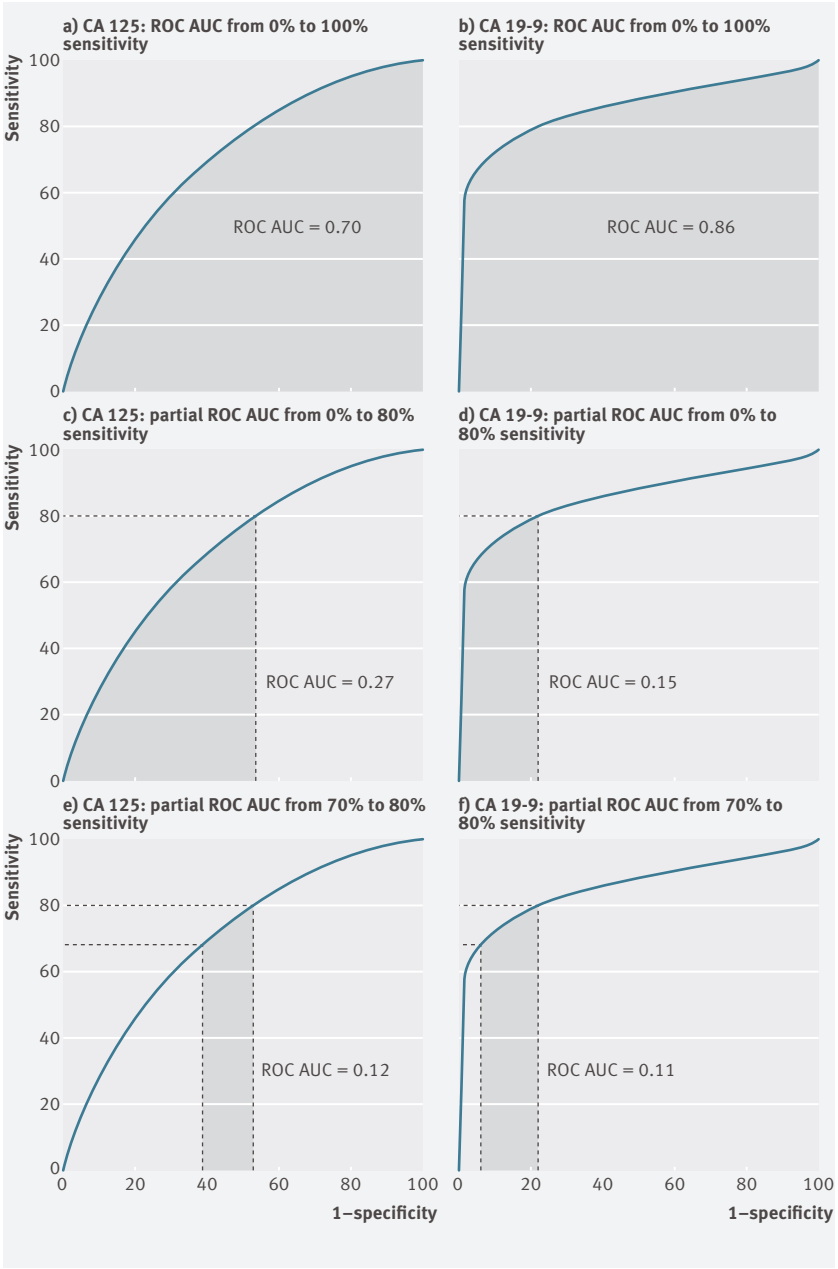

Fig 3 | Use of ROC AUC to compare two tests: CA 19-9 and CA 125. Shaded areas indicate ROC AUC for regions of interest. Blood samples from 51 control patients with pancreatitis and 90 patients with pancreatic cancer were analysed for CA 125 and CA 19-9[31]

Single measure averaged across multiple thresholds

A third approach calculates a single overall measure of diagnostic accuracy but averaged across multiple test thresholds—for example, ROC AUC[20] (fig 1) and the newer H-measure.[21] We illustrate ROC AUC with two tumour markers measured on the same patients[9]; CA 19-9 seems to be the superior test as it has an AUC of 0.86, which is greater than 0.70 for CA 125 (fig 3).

Problems with ROC AUC for diagnostic performance

The use and interpretation of ROC AUC as a measure of diagnostic performance highlights several advantages[6] and disadvantages.[4 22] Somewhat surprisingly, ROC AUC remains the recommended measure of effectiveness for some evaluations of devices submitted to the US Food and Drug Administration, for example in imaging and computer aided detection.[23]

AUC or partial AUC?

The standard ROC AUC averages across all possible thresholds. Not all test thresholds, however, are clinically relevant.[5] For many tests, thresholds offering high sensitivity (such as greater than 80%) are not clinically useful because specificity is too low (see fig 3a and b); patients with false positive results would overwhelm diagnostic services. One way to deal with this is to calculate a partial ROC AUC (pAUC), thus restricting comparisons to sensible thresholds.[24] For example, by excluding sensitivity above 80%, the partial ROC AUC is 0.27 for CA 125 and 0.15 for CA 19-9, suggesting that CA 125 is the superior test (see fig 3c and d). It could also be argued that a sensitivity of less than 70% is unlikely to be clinically useful (too little disease would be detected). A pAUC therefore restricted to the range between 70% and 80% sensitivity produces values of 0.12 for CA 125 and 0.11 for CA 19-9, suggesting the tests are equally effective (fig 3e and f).

This example illustrates a dilemma in ROC AUC interpretation. Should the AUC be calculated across all test thresholds (including those that are clinically illogical[5 25]) or should a pAUC be calculated, restricted to clinically sensible thresholds? If a partial AUC approach is taken, as illustrated in figure 3, even small changes in the choice of threshold can affect which test has the greater AUC and is classified as superior.[26]

Extrapolation beyond available data

The choice between standard AUC or pAUC needs particular consideration when available data are restricted to a small region of the ROC plot space. To calculate a standard AUC the ROC curve must be extrapolated beyond the available data so that the whole AUC encompassing 0% to 100% sensitivity can be calculated. This is a key issue in systematic reviews in which data from included studies are often limited to a small region of ROC space.

Moreover, the extrapolated region of the curve dominates the AUC as it includes the right hand side of the plot, which dominates the ROC AUC. This region lacks clinical importance because it is based on thresholds where over half the patients receive false positive results. The poor utility of the full AUC has been noted in breast screening, where high specificity is important to avoid large numbers of false positive results leading to unnecessary biopsies in a population with a low prevalence.[25]

Incorporating relative misclassification costs

ROC AUC does not allow incorporation of the relative clinical consequences of false negative and false positive diagnoses. It is often believed that ROC AUC uses equally balanced misclassification costs for these diagnoses, but this applies only at one point on the ROC curve, where the gradient equals one. In reality, the misclassification costs for false negative and false positive diagnoses vary along the ROC curve and are dictated by its shape[14] and therefore do not relate to any clinically meaningful information. This has been described as nonsensical and fundamentally incoherent.[27] If ROC AUC is used as a performance measure, then

when we compare two ROC curves with different shapes, different balances of misclassification costs of false negative and false positive diagnoses underlie each curve.[21] This is analogous to comparing the height of two people by using only the numerical output from two rulers, regardless that one ruler measures in inches and the other in centimetres.[27]

Incorporating disease prevalence

To be useful as a performance measure, ROC AUC needs to use realistic disease prevalence. For a given ROC curve, the calculated AUC is the same regardless of the underlying prevalence of the study data, given the same disease severity. When ROC AUC is used to compare two tests, this is sometimes wrongly perceived as evaluation at 50% prevalence. As with misclassification costs, unless the ROC curve corresponds to a straight line, it is not possible to fix a single disease prevalence with ROC AUC, as the gradient changes along the curve. To our knowledge this issue has not been previously highlighted. This is problematic when ROC AUC is used to compare tests because the absolute benefit of the difference in sensitivity and specificity is clearly dependent on disease prevalence.

Summary

Diagnostic test accuracy studies need to provide evidence in a comprehensible and intuitive format that facilitates choice of test for clinicians, their patients, and healthcare providers. Results should be reported in the context of clinical management decisions made at clinically sensible and important thresholds, preferably in terms of patients. For comparisons of tests, differences in true positive and false positive diagnoses should be reported, and it is important that any overall measures of diagnostic accuracy should incorporate relative misclassification costs to account for the fact that false negative and false positive diagnoses are rarely clinically equivalent. Measures need to be interpreted at a disease prevalence that reflects the real clinical situation. Analyses based on net benefit measures achieve these aims. In contrast, methods based on ROC AUC often incorporate thresholds that are clinically nonsensical, do not account for disease prevalence, and cannot account for the differing clinical implications of false negative and false positive diagnoses. We therefore caution researchers against solely reporting ROC AUC measures when summarising diagnostic performance, and caution healthcare providers against using ROC AUC alone to inform decisions regarding diagnostic performance. We recommend that diagnostic accuracy is presented by using paired measures with clinical context or using net benefit measures with their associated paired measures.

Contributors: SM initiated the manuscript based on discussions of the topic with the other authors, who all contributed to manuscript drafting. SM is guarantor.

Funding: This work was funded by UK Department of Health via a National Institute for Health Research (NIHR) programme grant (RP-PG-0407-10338) and Cancer Research UK programme grant (C5529). A proportion of this work was undertaken at UCLH/UCL, which receives a proportion of funding from the NIHR Comprehensive Biomedical Research Centre funding scheme. The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare that Medicsight plc (Hammersmith, London) funded research relating to computer assisted detection that precipitated some of the views expressed in this article.

Provenance and peer review: Not commissioned; peer reviewed.

1 Gigerenzer G, Gaissmaier W, Kurz-Mileke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychol Sci Public Interest* 2008;8:53-96.
2 Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 2000;56:345-51.
3 D'Orsi CJ, Mendelson EB, Ikeda DM. Breast imaging reporting and data system: ACR BI-RADS-Breast Imaging Atlas. In: D'Orsi CJ, Bassett LW, Berg WA, eds. BI-RADS: Mammography. American College of Radiology, 2003.
4 Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Acad Radiol* 2007;14:723-48.
5 Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by Pencina MJ, et al. Statistics in Medicine (doi:10.1002/sim.2929). *Stat Med* 2008;27:199-206.
6 Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false-positive mammography results and detection of ductal carcinoma in situ: cross-sectional survey. *West J Med* 2000;173:307-12.
7 Speigelhalter D. Understanding uncertainty. 2011. http://understandinguncertainty.org/view/animations.
8 Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by Pencina MJ, et al. Statistics in Medicine (doi:10.1002/sim.2929). *Stat Med* 2008;27:173-81.
9 Haglund C, Lundin J, Kuusela P, Roberts PJ. CA 242, a new tumour marker for pancreatic cancer: a comparison with CA 19-9, CA 50 and CEA. *Br J Cancer* 1994;70:487-92.
10 Adams NM, Hand DJ. Comparing classifers when the misallocation costs are uncertain. *Pattern Recognition* 1999;32:1139-47.
11 DeNeef P, Kent DL. Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. *Med Decis Making* 1993;13:126-32.
12 Lusted LB. Decision-making studies in patient management. *N Engl J Med* 1971;284:416-24.
13 Moons KG, Stijnen T, Michel BC, Buller HR, Van Es GA, Grobbee DE, et al. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Med Decis Making* 1997;17:447-54.
14 Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157-72.
15 Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Validity of prognostic models: when is a model clinically useful? *Semin Urol Oncol* 2002;20:96-107.
16 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-74.
17 Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population of readers and cases. *Med Decis Making* 2004;24:561-72.
18 Peirce CS. The numerical measure of the success of predictions. *Science* 1884;IV:453-4.
19 Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* 2008;8:53.
20 Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Acad Radiol* 2002;9:1264-77.
21 Hand DJ. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Stat Med* 2010;29:1502-10.
22 Hilden J. The area under the ROC curve and its competitors. *Med Decis Making* 1991;11:95-101.
23 Gallas BD, Chan HP, D'Orsi CJ, Dodd LE, Giger ML, Gur D,et al. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol* 2012;19:463-77.
24 Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat Med* 1989;8:1277-90.
25 Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst* 2003;95:511-5.
26 Walter SD. The partial area under the summary ROC curve. *Stat Med* 2005;24:2025-40.
27 Hand D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn* 2009;77:103-23.
28 Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129-35.
29 Niederau C, Grendell JH. Diagnosis of pancreatic carcinoma. Imaging techniques and tumor markers. *Pancreas* 1992;7:66-86.
30 Bandos AI, Rockette HE, Gur D. Incorporating utility-weights when comparing two diagnostic systems: a preliminary assessment. *Acad Radiol* 2005;12:1293-300.
31 Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989;76:585-92.