Education and debate

How to read a paper

Papers that report diagnostic or screening tests

Trisha Greenhalgh

This is the seventh in a series of 10 articles introducing non-experts to finding medical articles and assessing their value

Unit for Evidence-Based Practice and Policy, Department of Primary Care and Population Sciences, University College London Medical School/ Royal Free Hospital School of Medicine, Whittington Hospital, London N19 5NF
Trisha Greenhalgh, *senior lecturer*

p.greenhalgh@ucl.ac.uk

BMJ 1997;315:540–3

Ten men in the dock

If you are new to the concept of validating diagnostic tests, the following example may help you. Ten men are awaiting trial for murder. Only three of them actually committed a murder; the seven others are innocent of any crime. A jury hears each case and finds six of the men guilty of murder. Two of the convicted are true murderers. Four men are wrongly imprisoned. One murderer walks free.

This information can be expressed in what is known as a two by two table (table 1). Note that the "truth" (whether or not the men really committed a murder) is expressed along the horizontal title row, whereas the jury's verdict (which may or may not reflect the truth) is expressed down the vertical row.

These figures, if they are typical, reflect several features of this particular jury:

- the jury correctly identifies two in every three true murderers;
- it correctly acquits three out of every seven innocent people;
- if this jury has found a person guilty, there is still only a one in three chance that they are actually a murderer;
- if this jury found a person innocent, he or she has a three in four chance of actually being innocent; and
- in five cases out of every 10 the jury gets it right.

These five features constitute, respectively, the sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of this jury's performance. The rest of this article considers these five features applied to diagnostic (or screening) tests when compared with a "true" diagnosis or gold standard. A sixth feature—the likelihood ratio—is introduced at the end of the article.

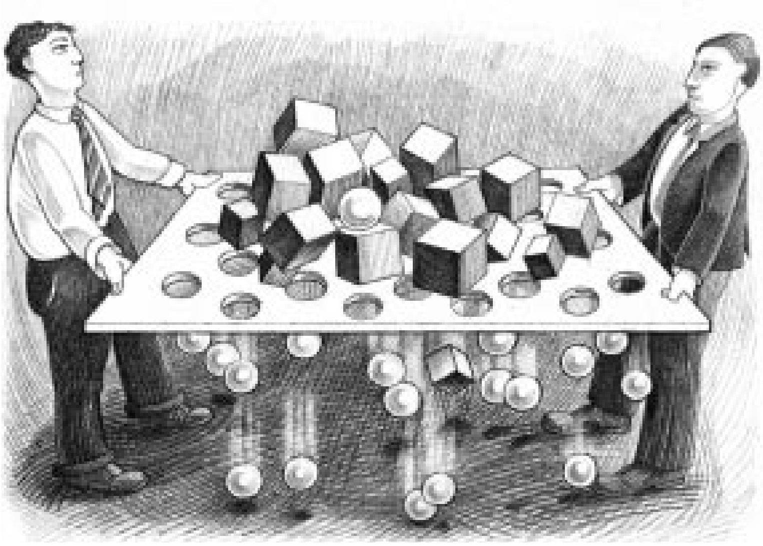
Summary points

New tests should be validated by comparison against an established gold standard in an appropriate spectrum of subjects

Diagnostic tests are seldom 100% accurate (false positives and false negatives will occur)

A test is valid if it detects most people with the target disorder (high sensitivity) and excludes most people without the disorder (high specificity), and if a positive test usually indicates that the disorder is present (high positive predictive value)

The best measure of the usefulness of a test is probably the likelihood ratio—how much more likely a positive test is to be found in someone with, as opposed to without, the disorder



PETER BROWN

Validating tests against a gold standard

Our window cleaner told me that he had been feeling thirsty recently and had asked his general practitioner to be tested for diabetes, which runs in his family. The nurse in his surgery had asked him to produce a urine specimen and dipped a stick in it. The stick stayed green, which meant, apparently, that there was no sugar in his urine. This, the nurse had said, meant that he did not have diabetes.

I had trouble explaining that the result did not necessarily mean this, any more than a guilty verdict necessarily makes someone a murderer. The definition of diabetes, according to the World Health Organisation, is a blood glucose level above 8 mmol/l in the fasting state, or above 11 mmol/l two hours after a 100 g oral glucose load, on one occasion if the patient has symptoms and on two occasions if he or she does not.[1] These stringent criteria can be termed

Table 1 Two by two table showing outcome of trial for 10 men accused of murder

	True criminal status	
Jury verdict	**Murderer**	**Not murderer**
Guilty	Rightly convicted (2 men)	Wrongly convicted (4 men)
Innocent	Wrongly acquitted (1 man)	Rightly acquitted (3 men)

Table 2 Two by two table notation for expressing the results of validation study for diagnostic or screening test

	Result of gold standard test	
Result of screening test	**Disease positive** (a+c)	**Disease negative** (b+d)
Test positive (a+b)	True positive (a)	False positive (b)
Test negative (c+d)	False negative (c)	True negative (d)

Table 3 Two by two table showing results of validation study of urine glucose testing for diabetes against gold standard[3]

	Result of glucose tolerance test	
Result of urine test for glucose	**Diabetes positive (n=27)**	**Diabetes negative (n=973)**
Glucose present (n=13)	True positive (n=6)	False positive (n=7)
Glucose absent (n=987)	False negative (n=21)	True negative (n=966)

the gold standard for diagnosing diabetes (although purists have challenged this notion[2]).

The dipstick test, however, has some distinct practical advantages over the fullblown glucose tolerance test. To assess objectively just how useful the dipstick test for diabetes is, we would need to select a sample of people (say 100) and do two tests on each of them: the urine test (screening test) and a standard glucose tolerance test (gold standard). We could then see, for each person, whether the result of the screening test matched the gold standard (see table 2). Such an exercise is known as a validation study.

The validity of urine testing for glucose in diagnosing diabetes has been looked at by Andersson and colleagues,[3] whose data I have adapted for use (expressed as a proportion of 1000 subjects tested) in table 3.

From the calculations of important features of the urine dipstick test for diabetes (box), you can see why I did not share the window cleaner's assurance that he did not have diabetes. A positive urine glucose test is only 22% sensitive, which means that the test misses nearly four fifths of people who have diabetes. In the presence of classical symptoms and a family history, the window cleaner's baseline chances (pretest likelihood) of having the condition are pretty high and is reduced to only about four fifths of this (the negative likelihood ratio, 0.78; see below) after a single negative urine test. This man clearly needs to undergo a more definitive test.

Does the paper validate the test?

The 10 questions below can be asked about a paper that claims to validate a diagnostic or screening test. In preparing these tips, I have drawn on several sources.[4-8]

Question 1: Is this test potentially relevant to my practice?
Sackett and colleagues call this the utility of the test.[6] Even if this test were 100% valid, accurate, and reliable, would it help me? Would it identify a treatable disorder? If so, would I use it in preference to the test I use now? Could I (or my patients or the taxpayer) afford it? Would my patients consent to it? Would it change the probabilities for competing diagnoses sufficiently for me to alter my treatment plan?

Question 2: Has the test been compared with a true gold standard?
You need to ask, firstly, whether the test has been compared with anything at all. Assuming that a "gold standard" test has been used, you should verify that it merits the description, perhaps by using the questions listed in question 1. For many conditions, there is no gold standard diagnostic test. Unsurprisingly, these tend to be the conditions for which new tests are most actively sought. Hence, the authors of such papers may need to develop and justify a combination of criteria against which the new test is to be assessed. One specific point to check is that the test being validated in the paper is not being used to define the gold standard.

Question 3: Did this validation study include an appropriate spectrum of subjects?
Although few investigators would be naive enough to select only, say, healthy male medical students for their validation study, only 27% of published studies explicitly define the spectrum of subjects tested in terms of age, sex, symptoms or disease severity, and specific eli-

Features of diagnostic test that can be calculated by comparison with gold standard in validation study

Feature of the test	**Alternative name**	**Question addressed**	**Formula (see table 2)**
Sensitivity	True positive rate (positive in disease)	How good is this test at picking up people who have the condition?	a/(a + c)
Specificity	True negative rate (negative in health)	How good is this test at correctly excluding people without the condition?	d/(b + d)
Positive predictive value	Post-test probability of a positive test	If a person tests positive, what is the probability that he or she has the condition?	a/(a + b)
Negative predictive value	Post-test probability of a negative test	If a person tests negative, what is the probability that he or she does not have the condition?	d/(c + d)
Accuracy	—	What proportion of all tests have given the correct result? (true positives and true negatives as a proportion of all results)	(a + d)/(a + b + c + d)
Likelihood ratio of a positive test	—	How much more likely is a positive test to be found in a person with the condition than in a person without it?	sensitivity/(l—specificity)
Likelihood ratio of a negative test	—	How much more likely is a negative test to be found in a person without the condition than in a person with it?	(l—sensitivity)/specificity

Calculating the important features of screening test

Feature	Formula	Data (see table 3)	Value
Sensitivity	a/(a + c)	6/27	22.2%
Specificity	d/(b + d)	966/973	99.3%
Positive predictive value	a/(a + b)	6/13	46.2%
Negative predictive value	d/(c + d)	966/973	97.8%
Accuracy	(a + d)/(a + b + c + d)	972/1000	97.2%
Likelihood ratio:			
Positive test	Sensitivity/(1–specificity)	22.2/0.7	32
Negative test	(1–sensitivity)/specificity	77.8/99.3	0.78

gibility criteria.[7] Importantly, the test should be verified on a population which includes mild and severe disease, treated and untreated subjects, and those with different but commonly confused conditions.[6]

Although the sensitivity and specificity of a test are virtually constant whatever the prevalence of the condition, the positive and negative predictive values depend crucially on prevalence. This is why general practitioners are sceptical of the utility of tests developed exclusively in a secondary care population, and why a good diagnostic test is not necessarily a good screening test.

Question 4: Has workup bias been avoided?
This is easy to check. It simply means, “Did everyone who got the new diagnostic test also get the gold standard, and vice versa?” There is clearly a potential bias in studies where the gold standard test is performed only on people who have already tested positive for the test being validated.[7]

Question 5: Has expectation bias been avoided?
Expectation bias occurs when pathologists and others who interpret diagnostic specimens are subconsciously influenced by the knowledge of the particular features of the case—for example, the presence of chest pain when interpreting an electrocardiogram. In the context of validating diagnostic tests against a gold standard, all such assessments should be “blind.”

Question 6: Was the test shown to be reproducible?
If the same observer performs the same test on two occasions on a subject whose characteristics have not changed, they will get different results in a proportion of cases. Similarly, it is important to confirm that reproducibility between different observers is at an acceptable level.[9]

Question 7: What are the features of the test as derived from this validation study?
All the above standards could have been met, but the test might still be worthless because the sensitivity, specificity, and other crucial features of the test are too low—that is, the test is not valid. What counts as acceptable depends on the condition being screened for. Few of us would quibble about a test for colour blindness that was 95% sensitive and 80% specific, but nobody ever died of colour blindness. The Guthrie heel-prick screening test for congenital hypothyroidism, performed on all babies in Britain soon after birth, is over 99% sensitive but has a positive predictive value of only 6% (it picks up almost all babies with the condition at the expense of a high false positive rate),[10] and rightly so. It is more important to pick up every baby with this treatable condition who would otherwise develop severe mental handicap than to save hundreds the minor stress of a repeat blood test.

Question 8: Were confidence intervals given?
A confidence interval, which can be calculated for virtually every numerical aspect of a set of results, expresses the possible range of results within which the true value will probably lie. If the jury in the first example had found just one more murderer not guilty, the sensitivity of its verdict would have gone down from 67% to 33%, and the positive predictive value of the verdict from 33% to 20%. This enormous (and quite unacceptable) sensitivity to a single case decision is, of course, because we validated the jury’s performance on only 10 cases. The larger the sample, the narrower the confidence interval, so it is particularly important to look for confidence intervals if the paper you are reading reports a study on a relatively small sample.[11]

Question 9: Has a sensible “normal range” been derived?
If the test gives non-dichotomous (continuous) results—that is, if it gives a numerical value rather than a yes/no result—someone will have to say what values count as abnormal. Defining relative and absolute danger zones for a continuous variable (such as blood pressure) is a complex science, which should take into account the actual likelihood of the adverse outcome which the proposed treatment aims to prevent. This process is made considerably more objective by the use of likelihood ratios (see below).

Question 10: Has this test been placed in the context of other potential tests in the diagnostic sequence?
In general, we treat high blood pressure simply on the basis of a series of resting blood pressure readings. Compare this with the sequence we use to diagnose coronary artery stenosis. Firstly, we select patients with a typical history of effort angina. Next, we usually do a resting electrocardiogram, an exercise electrocardiogram, and, in some cases, a radionuclide scan of the heart. Most patients come to a coronary angiogram only after they have produced an abnormal result on these preliminary tests.

If you sent 100 ordinary people for a coronary angiogram, the test might show very different positive and negative predictive values (and even different sensitivity and specificity) than it did in the ill population on which it was originally validated. This means that the various aspects of validity of the coronary angiogram as a diagnostic test are virtually meaningless unless these figures are expressed in terms of what they contribute to the overall diagnostic work up.

A note on likelihood ratios

Question 9 above described the problem of defining a normal range for a continuous variable. In such circumstances, it can be preferable to express the test result not as “normal” or “abnormal” but in terms of the actual chances of a patient having the target disorder if the test result reaches a particular level. Take, for example, the use of the prostate specific antigen (PSA) test to screen for prostate cancer. Most men will have some detectable

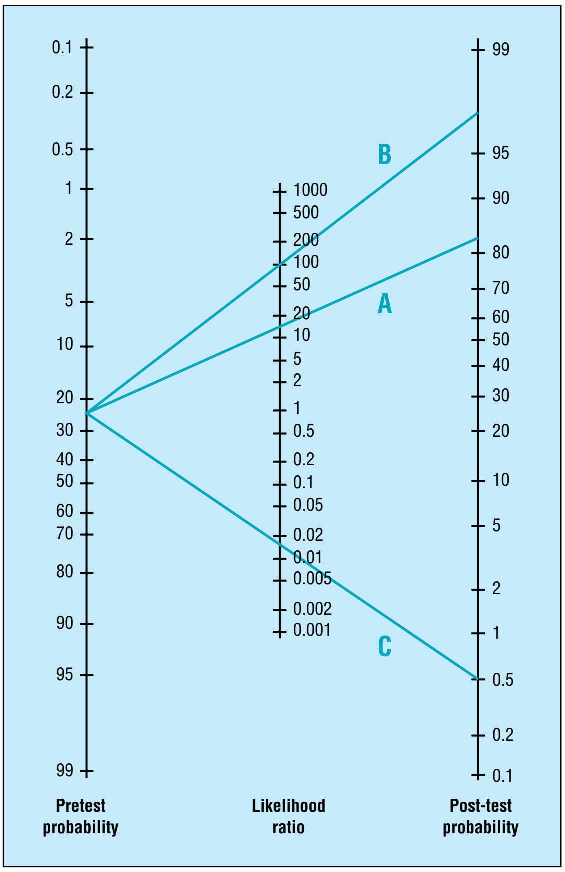


Fig 1 Use of likelihood ratios to calculate post-test probability of someone being a smoker[6]

antigen in their blood (say, 0.5 ng/ml), and most of those with advanced prostate cancer will have high concentrations (above about 20 ng/ml). But a concentration of, say, 7.4 ng/ml may be found either in a perfectly normal man or in someone with early cancer. There simply is not a clean cutoff between normal and abnormal.[12]

We can, however, use the results of a validation study of this test against a gold standard for prostate cancer (say a biopsy of the prostate gland) to draw up a whole series of two by two tables. Each table would use a different definition of an abnormal test result to classify patients as "normal" or "abnormal." From these tables, we could generate different likelihood ratios associated with an antigen concentration above each different cutoff point. When faced with a test result in the "grey zone" we would at least be able to say, "This test has not proved that the patient has prostate cancer, but it has increased [or decreased] the odds of that diagnosis by a factor of *x*."

The likelihood ratio thus has enormous practical value, and it is becoming the preferred way of expressing and comparing the usefulness of different tests.[6] For example, if a person enters my consulting room with no symptoms at all, I know that they have a 5% chance of having iron deficiency anaemia, since I know that one person in 20 in the population has this condition (in the language of diagnostic tests, the pretest probability of anaemia is 0.05).[13]

Now, if I do a diagnostic test for anaemia, the serum ferritin concentration, the result will usually make the diagnosis of anaemia either more or less likely. A moderately reduced serum ferritin concentration (between 18 and 45 µg/l) has a likelihood ratio of 3, so the chances of a patient with this result having iron deficiency anaemia is 0.05 × 3—or 0.15 (15%). This value is known as the post-test probability of the serum ferritin test. The likelihood ratio of a very low serum ferritin concentration (below 18 µg/l) is 41, making the chances of iron deficiency anaemia in a patient with this result greater than unity. On the other hand, a very high concentration (above 100 µg/l; likelihood ratio 0.13) would reduce the chances of the patient being anaemic from 5% to less than 1%.[13]

Figure 1 shows a nomogram, adapted by Sackett and colleagues from an original paper by Fagan,[14] for working out post-test probabilities when the pretest probability (prevalence) and likelihood ratio for the test are known. The lines A, B, and C, drawn from a pretest probability of 25% (the prevalence of smoking among British adults), are the trajectories through likelihood ratios of 15, 100, and 0.015, respectively—three different tests for detecting whether someone is a smoker.[15] Actually, test C detects whether the person is a non-smoker, since a positive result in this test leads to a post-test probability of only 0.5%.

Thanks to Dr Sarah Walters and Dr Jonathan Elford for advice, and in particular to Dr Walters for the jury example.

The articles in this series are excerpts from *How to read a paper: the basics of evidence based medicine.* The book includes chapters on searching the literature and implementing evidence based findings. It can be ordered from the BMJ Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

1 WHO Study Group. Diabetes mellitus. *WHO Tech Report Ser* 1985:No 727.

2 McCance DR, Hanson RL, Charles M-A, Jacobsson LTH, Pettitt DJ, Bennett PH, et al. Comparison of tests for glycated haemoglobin and fasting and two-hour plasma glucose concentrations as diagnostic measures for diabetes. *BMJ* 1994;308:1323-8.

3 Andersson DKG, Lundblad E, Svardsudd K. A model for early diagnosis of type 2 diabetes mellitus in primary health care. *Diabet Med* 1993;10:167-73.

4 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 1994;271:389-91.

5 Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What were the results and will they help me in caring for my patients? *JAMA* 1994;271:703-7.

6 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology—a basic science for clinical medicine.* London: Little, Brown, 1991:51-68.

7 Read MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274:645-51.

8 Mant D. Testing a test: three critical steps. In: Jones R, Kinmonth A-L, eds. *Critical reading for primary care.* Oxford: Oxford University Press, 1995:183-90.

9 Bush B, Shaw S, Cleary P, Delbanco TL, Aronson MD. Screening for alcohol abuse using the CAGE questionnaire. *Am J Med* 1987;82:231-6.

10 Verkerk PH, Derksen-Lubsen G, Vulsma T, Loeber JG, de Vijlder JJ, Verbrugge HP. Evaluation of a decade of neonatal screening for congenital hypothyroidism in the Netherlands. *Ned Tijdschr Geneesk* 1993;137: 2199-205.

11 Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines.* London: BMJ Books, 1989.

12 Catalona WJ, Hudson MA, Scardino PT, Richie JP, Ahmann FR, Flanigan RC, et al. Selection of optimal prostate specific antigen cutoffs for early diagnosis of prostate cancer: receiver operator characteristic curves. *J Urol* 1994;152:2037-42.

13 Guyatt GH, Patterson C, Ali M, Singer J, Levine M, Turpie I, Meyer R. Diagnosis of iron deficiency anaemia in the elderly. *Am J Med* 1990; 88:205-9.

14 Fagan TJ. Nomogram for Bayes' theorem. *N Engl J Med* 1975;293:257-61.

15 How good is that test—using the result. *Bandolier* 1996;3:6-8.