How to read a paper

The basics of evidence based medicine

Trisha Greenhalgh

BMJ Publishing Group

First published in 1997
Second impression 1997
Third impression 1998
Fourth impression 1998
Fifth impression 1999
Sixth impression 2000
by the BMJ Publishing Group, BMA House, Tavistock Square, London WC1H 9JR

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-7279-1139-2

Typeset and printed by Derry, Nottingham

[5] Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Health Technol Assess* 1996; **12**: 209-37.
[6] Kleijnen J, de Craen AJ, van Everdingen J, *et al.* Placebo effect in double-blind clinical trials: a review of interactions with medications. *Lancet* 1994; **344**: 1347-9.
[7] Joyce CR. Placebo and complementary medicine. *Lancet* 1994; **344**: 1279-81.
[8] Laporte JR, Figueras A. Placebo effects in psychiatry. *Lancet* 1994; **344**: 1206-9.
[9] Johnson AG. Surgery as a placebo. *Lancet* 1994; **344**: 1140-2.
[10] Thomas KB. The placebo in general practice. *Lancet* 1994; **344**: 1066-7.
[11] Chaput de Saintonge DM, Herxheimer A. Harnessing placebo effects in health care. *Lancet* 1994; **344**: 995-8.
[12] Gotzsche PC. Is there logic in the placebo? *Lancet* 1994; **344**: 925-6.
[13] Sackett DL, Haynes RB, Guyatt GH, *et al. Clinical epidemiology - a basic science for clinical medicine.* London: Little, Brown, 1991; 187-248.
[14] Bostwick DG, Burke HB, Wheeler TM, *et al.* The most promising surrogate endpoint biomarkers for screening candidate chemopreventive compounds for prostatic adenocarcinoma in short-term Phase II clinical trials. *J Cell Biochem* 1994; suppl 19; 283-9.
[15] Gøtzsche P, Liberati A, Torri V, *et al.* Beware of surrogate outcome measures. *Int J Health Technol Assess* 1996; **12**: 238-46.
[16] Lipkin M. Summary of recommendations for colonic biomarker studies of candidate chemopreventive compounds in phase II clinical trials. *J Cell Biochem* 1994; suppl 19: 94-8.
[17] Kimbrough RD. Determining acceptable risks: experimental and epidemiological issues. *Clin Chem* 1994; **40**: 1448-53.
[18] CONCORDE Coordinating Committee. CONCORDE MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet* 1994; **343**: 871-81.
[19] Jacobson MA, Bacchetti P, Kolokathis A, *et al.* Surrogate markers for survival in patients with AIDS and AIDS related complex treated with zidovudine. *BMJ* 1991; **302**: 73-8.
[20] Blatt SP, McCarthy WF, Bucko-Krasnicka B, *et al.* Multivariate models for predicting progression to AIDS and survival in HIV-infected patients. *J Infect Dis* 1995; **171**: 837-44.
[21] Tsoukas CM, Bernard NF. Markers predicting progression of HIV-related disease. *Clin Microbiol Rev* 1994; 7: 14-28.
[22] Epstein AE, Hallstrom AO, Rogers WJ, *et al.* Mortality following ventricular arrhythmia suppression by encainide, flecainide and moricizine after myocardial infarction. *JAMA* 1993; **270**: 2451-5.
[23] Lipicky RJ, Packer M. Role of surrogate endpoints in the evaluation of drugs for heart failure. *J Am Coll Cardiol* 1993; **22** (suppl A); 179-84.
[24] Hyatt JM, McKinnon PS, Zimmer GS, *et al.* The importance of pharmacokinetic/pharmacodynamic surrogate markers to outcome. Focus on antibacterial agents. *Clin Pharmacokinetics* 1995; **28**: 143-60.
[25] Anonymous. Interferon beta-1b—hope or hype? *Drug Ther Bull* 1996; **34**: 9-11.
[26] Entire issue of *J Cell Biochem* 1994; suppl 19.
[27] Aicken M. If there is gold in the labelling index hills, are we digging in the right place? *J Cell Biochem* 1994; suppl 19; 91-3.
[28] Anonymous. Getting good value from drug reps. *Drug Ther Bull* 1983; **21**: 13-5.
[29] Ferner RE. Newly licensed drugs. *BMJ* 1996; **313**: 1157-8.

Chapter 7: Papers that report diagnostic or screening tests

7.1 Ten men in the dock

If you are new to the concept of validating diagnostic tests and if algebraic explanations ("let's call this value *x*...") leave you cold, the following example may help you. Ten men are awaiting trial for murder. Only three of them actually committed a murder; the other seven 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 7.1). Note that the "truth" (that is, 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 title row.

You should be able to see that these figures, if they are typical, reflect several features of this particular jury:

Table 7.1 2x2 Table showing outcome of trial for ten 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

- 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 has a three in four chance of actually being innocent
- In five cases out of every ten the jury gets the verdict 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 chapter considers these five features applied to diagnostic (or screening) tests when compared with a "true" diagnosis or gold standard. Section 7.4 also introduces a sixth, slightly more complicated (but very useful), feature of a diagnostic test—the likelihood ratio. (After you have read the rest of this chapter, look back at this section. You should, by then, be able to work out that the likelihood ratio of a positive jury verdict in the above example is 1.17 and that of a negative one 0.78. If you can't, don't worry—many eminent clinicians have no idea what a likelihood ratio is.)

7.2 Validating diagnostic tests against a gold standard

Our window cleaner told me the other day 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 general practitioner's surgery had asked him to produce a urine specimen and dipped a special stick in it. The stick stayed blue, which meant, apparently, that there was no sugar (glucose) in his urine. This, the nurse had said, meant that he did not have diabetes.

I had trouble explaining to the window cleaner that the test result did not necessarily mean this at all, 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 concentration above 8 mmol/L in the fasting state or above 11 mmol/L two hours after a 100 g oral glucose load (the much dreaded "glucose tolerance test", in which the subject has to glug down every last drop of a sickly glucose drink and wait two hours

for a blood test). These values must be achieved on two separate occasions if the person has no symptoms but on only one occasion if they have typical symptoms of diabetes (thirst, passing large amounts of urine, and so on)[1].

These stringent criteria can be termed the *gold standard* for diagnosing diabetes. In other words, if you fulfil the WHO criteria you can call yourself diabetic, and if you don't, you can't (although some critics have, in fact, challenged this notion[2].) The same cannot be said for dipping a stick into a random urine specimen. For one thing, you might be a true diabetic but have a high renal threshold—that is, your kidneys conserve glucose much better than most people's so your blood glucose concentration would have to be much higher than most people's for any glucose to appear in your urine. Alternatively, you may be an otherwise normal individual with a *low* renal threshold, so glucose leaks into your urine even when there isn't any excess in your blood. In fact, as anyone with diabetes will tell you, diabetes is very often associated with a negative test result for urine glucose.

There are, however, many advantages in using a urine dipstick rather than the full blown glucose tolerance test to "screen" people for diabetes. The test is cheap, convenient, easy to perform and interpret, acceptable to patients, and gives an instant yes or no result. In real life, people like my window cleaner may decline to take an oral glucose tolerance test. Even if he was prepared to go ahead with it, his general practitioner might decide that the window cleaner's symptoms did not merit the expense of this relatively sophisticated investigation. I hope you can see that even though the urine test cannot say for sure if someone is diabetic, it has a definite practical edge over the gold standard. That, of course, is why we use it!

To assess objectively just how useful the urine glucose 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. Such an exercise is known as a *validation study*. We could express the results of the validation study in a two by two table (also known as a two by two matrix) as in table 7.2 and calculate various features of the test as in table 7.4 just as we did for the features of the jury in section 7.1.

If the values for the various features of a test (such as sensitivity

Table 7.2 2x2 Table notation for expressing the results of a validation study for a 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**

and specificity) fell within reasonable limits, we would be able to say that the test was *valid* (see question 7 below). The validity of testing urine for glucose in the diagnosis of diabetes has been looked at by Andersson and colleagues[3], whose data I have used in the example in table 7.3. The original study was performed on 3268 subjects, of whom 67 either refused to produce a specimen or, for some other reason, were not adequately tested. For simplicity's sake, I have ignored these irregularities and expressed the results in terms of a denominator (total number tested) of 1000 subjects.

Table 7.3 2x2 Table showing results of validation study of urine glucose testing for diabetes against gold standard of glucose tolerance test

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

In actual fact these data came from an epidemiological survey to detect the prevalence of diabetes in a population; the validation of urine testing was a side issue to the main study. If the validation had been the main aim of the study, the subjects selected would have included far more people with diabetes, as question 2 in section 7.3 below will show. If you look up the original paper, you will also find that the gold standard for diagnosing true diabetes was not the oral glucose tolerance test but a more unconventional series of observations. Nevertheless, the example serves its purpose

Feature of the test	**Alternative name**	**Question that the feature examines**	**Formula (see tables 7.1, 7.2 and 7.4)**
Sensitivity	True positive rate (**P**ositive in **D**isease)	How good is this test at picking up people who have the condition?	a/a+c
Specificity	True negative rate (**N**egative in **H**ealth)	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 (s)he 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 (s)he does not have the condition?	d/c+d
Accuracy		What proportion of all tests have given the correct result (that is, 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 result to be found in a person with, as opposed to without, the condition?	sensitivity/ (1-specificity)

Table 7.4 Features of a diagnostic test that can be calculated by comparing it with a gold standard in a validation study

as it provides us with some figures to put through the equations listed in the last column of table 7.4. We can calculate the important feature of the urine test for diabetes as follows:

- 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 = 966/987 = 97.8%
- Accuracy = (a+d)/(a+b+c+d) = 972/1000 = 97.2%
- Likelihood ratio of a positive result = sensitivity/(1 - specificity) = 22.2/0.7 = 32
- Likelihood ratio of a negative result = (1 - sensitivity) / specificity = 77.8/99.3 = 0.78.

From these features, you can probably 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 true diabetics. 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 it is reduced to only about four fifths of this (the negative likelihood ratio, 0.78; see section 7.4) after a single negative urine test. In view of his symptoms, this man clearly needs to undergo a more definitive test for diabetes.

7.3 Ten questions to ask about a paper that claims to validate a diagnostic or screening test

In preparing the tips below, I have drawn on three main published sources: the "Users' guides to the medical literature"[4 5] and the book by the same authors[6]; a more recent article in the *Journal of the American Medical Association*[7], and David Mant's simple and pragmatic guidelines for "testing a test"[8].

Question 1—Is this test potentially relevant to my practice?

This is the "so what?" question that Sackett and colleagues call 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? If the answers to these questions are all "no", you may be able to reject the paper (and the test) without reading further than the abstract or introduction.

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! Papers have occasionally been written (and, in the past, published) in which nothing has been done except perform the new test on a few dozen subjects. This exercise may give a range of possible results for the test, but it certainly does not confirm that the "high" results indicate that target disorder (the disease you are looking for) is present or that the "low" results indicate that it isn't.

Next, you should verify that the "gold standard" test used in the survey merits the term. A good way of assessing a gold standard is to use the "so what?" questions listed above. For many conditions, there is no absolute gold standard diagnostic test that will say for certain if it is present or not. Unsurprisingly, these tend to be the very 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 here (or a variant of it) is not being used to contribute to the definition of the gold standard.

Question 3—Did this validation study include an appropriate spectrum of subjects?

If you validated a new test for cholesterol in 100 healthy male medical students, you would not be able to say how the test would perform in women, children, older people, those with diseases that seriously raise the cholesterol concentration, or even those who had never been to medical school! Although few people would be naive enough to select quite such a biased sample for the validation study, only 27% of published studies explicity define the spectrum

of subjects tested in terms of age, sex, symptoms or severity of disease or both, and specific eligibility criteria[7].

Definition of both the range of participants and the spectrum of disease to be included is essential if the values for the different features of the test are to be worth quoting—that is, if they are to be transferable to other settings. A particular diagnostic test may, conceivably, be more sensitive in women than men or in younger rather than older subjects. For the same reasons, as Sackett and colleagues stipulate, the subjects on which any test is verified should include those with both mild and severe disease, treated and untreated, and those with different but commonly confused conditions[6].

While the sensitivity and specificity of a test are virtually constant whatever the prevalence of the condition, the positive and negative predictive values are crucially dependent on prevalence. This why general practitioners are, often rightly, sceptical of the utility of tests developed exclusively in a secondary care population, where the severity of disease tends to be greater (see section 4.2) and why a good *diagnostic* test (generally used when the patient has some symptoms suggestive of the disease in question) is not necessarily a good *screening* test (generally used in people without symptoms, who are drawn from a population with a much lower prevalence of the disease).

Question 4—Has work up bias been avoided?

This is easy to check. It simply means, "Did everyone who got the new diagnostic test also get the gold standard test and vice versa?" I hope you have no problem spotting the potential bias in studies in which the gold standard test is performed only on people who have already tested positive for the test being validated. There are, in addition, a number of more subtle aspects of work up bias that are beyond the scope of this book. If you are interested, you could follow the discussion on this subject in Read and colleagues' paper[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 they are interpreting an

electrocardiogram. In the context of validating diagnostic tests against a gold standard, the question means, "Did the people who interpreted one of the tests know what result the other test had shown on each particular subject?" As I explained in section 4.5, all assessments should be "blind"—that is, the person interpreting the test should not be given any inkling of what the result is expected to be in any particular case.

Question 6—Was the test shown to be reproducible both within and between observers?

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. All tests show this feature to some extent, but a test with a reproducibility of 99% is clearly in a different league from one with a reproducibility of 50%. Several factors may contribute to the poor reproducibility of a diagnostic test: the technical precision of the equipment, observer variability (for example, in comparing a colour with a reference chart), arithmetical errors, and so on.

Look back again at page 62 to remind yourself of the problem of interobserver agreement. Given the same result to interpret, two people will agree in only a proportion of cases, generally expressed as the κ score. If the test in question gives results in terms of numbers (such as the blood cholesterol concentration in mmol/L), interobserver agreement is hardly an issue. If, however, the test involves reading x rays (such as the mammogram example in section 4.5) or asking people questions about their drinking habits[9], it is important to confirm that reproducibility between observers is at an acceptable level.

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 test itself is not valid—that is, its sensitivity, specificity, and other crucial features are too low. That is arguably the case for using urine glucose as a screening test for diabetes (see section 7.2 above). After all, if a test has a false negative rate of nearly 80%, it is more likely to mislead the clinician than assist the diagnosis if the target disorder is actually present.

There are no absolutes for the validity of a screening test, as

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 the UK soon after birth, is over 99% sensitive but has a positive predictive value of only 6% (in other words, it picks up almost all babies with the condition at the expense of a high false positive rate)[10], and rightly so. It is far more important to pick up every single baby with this treatable condition, who would otherwise develop severe mental handicap, than to save hundreds of parents the relatively minor stress of a repeat blood test on their baby.

Question 8—Were confidence intervals given for sensitivity, specificity, and other features of the test?

As section 5.5 explained, 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 lies. Go back to the jury example in section 7.1. If they had found just one more murderer not guilty, the sensitivity of their 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 ten cases. The confidence intervals for the features of this jury are so wide that my computer programme refuses to calculate them! Remember, the larger the sample size, 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. If you would like the formula for calculating confidence intervals for diagnostic test features, see Gardner and Altman's textbook *Statistics with Confidence*[11].

Question 9—Has a sensible "normal range" been derived from these results?

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 at what value the test result will count as abnormal. Many of us have been there with our own blood pressure reading.

We want to know if our result is "okay" or not, but the doctor insists on giving us a value such as "142/92". If 140/90 were chosen as the cutoff for high blood pressure, we would be placed in the "abnormal" category, even though our risk of adverse outcome is hardly different from that of a person with a blood pressure of 138/88. Quite sensibly, many practising doctors advise their patients, "Your blood pressure isn't quite right, but it doesn't fall into the danger zone. Come back in three months for another check". Nevertheless, the doctor must at some stage make the decision that *this* blood pressure needs treating with tablets but *that* one does not.

Defining relative and absolute danger zones for a continuous physiological or pathological variable is a complex science, which should take into account the actual likelihood of the adverse outcome that the proposed treatment aims to prevent. This process is made considerably more objective by the use of likelihood ratios (see section 7.4). For an entertaining discussion on the different possible meanings of the word "normal" in diagnostic investigations, see Sackett and colleagues' textbook[6], page 59.

Question 10—Has this test been placed in the context of other potential tests in the diagnostic sequence for the condition?

In general, we treat high blood pressure simply on the basis of the blood pressure reading alone (although we tend to rely on a series of readings rather than a single value). Compare this with the sequence we use to diagnose stenosis ("hardening") of the coronary arteries. Firstly, we select patients with a typical history of effort angina (chest pain on exercise). Next, we usually do a resting electrocardiogram, an exercise electrocardiogram, and, in some cases, a radionucleide scan of the heart to look for areas short of oxygen. Most patients come to a coronary angiogram (the definitive investigation for coronary artery stenosis) only *after* they have produced an abnormal result on these preliminary tests.

If you took 100 people off the street and sent them straight for a coronary angiogram, the test might display very different positive and negative predictive values (and even different sensitivity and specificity) than it did in the sicker 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.

7.4 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 PSA in their blood (say, 0.5 ng/ml), and most of those with advanced prostate cancer will have very high concentrations of PSA (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 cut off between normal and abnormal[12].

We can, however, use the results of a validation study of the PSA test against a gold standard for prostate cancer (say a biopsy) to draw up a whole series of two by two tables. Each table would use a different definition of an abnormal PSA result to classify patients as "normal" or "abnormal". From these tables, we could generate different likelihood ratios associated with a concentration above each different cut off point. Then, when faced with a 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 likelihood of that diagnosis by a factor of *x*".

Although the likelihood ratio is one of the more complicated aspects of a diagnostic test to calculate, it has enormous practical value and it is becoming the preferred way of expressing and comparing the usefulness of different tests. As Sackett and colleagues explain at great length in their textbook[6], the likelihood ratio can be used directly in ruling a particular diagnosis in or out. 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, this means that the pretest probability of anaemia, equivalent to the prevalence of the condition, 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 chance of a patient with this result having iron deficiency anaemia is generally calculated to be 0.05 x 3—or 0.15 (15%). This value is known as the post-test probability of the serum ferritin test. Strictly speaking, likelihood ratios should be used on odds rather than probabilities, but the simpler method shown here gives a good approximation when the pre-test probability is low. In this example, a pre-test probability of 5% is equal to a pre-test odds of 0.5/0.95 or 0.053; a positive test with a likelihood ratio of 3 gives a post-test odds of 0.158, which is equal to a post-test probability of 14%.

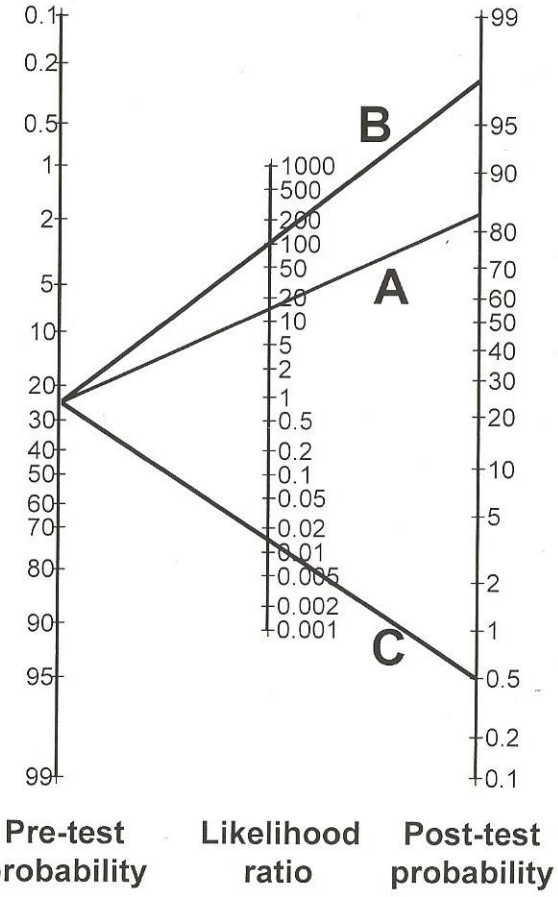

Fig 7.1 Using likelihood ratios to calculate the post-test probability of someone being a smoker

Figure 7.1 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, respectively, the trajectories through likelihood ratios of 15, 100, and 0.015—three different tests for detecting whether someone is a smoker[15]. Actually, test C detects whether the person is a *non-smoker* as a positive result in this test leads to a post-test probability of only 0.5%.

In summary, as I said at the beginning of this chapter, you can get a long way with diagnostic tests without referring to likelihood ratios. I avoided them myself for years. But if you put aside an afternoon to get to grips with this aspect of clinical epidemiology, I predict that your time will have been well spent.

1 WHO Study Group. Diabetes mellitus. *WHO Tech Rep Ser* 1985; **727**.

2 McCance DR, Hanson RL, Charles M-A, *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, Lundbld 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, *et al. 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, et *al.* Screening for alcohol abuse using the CAGE questionnaire. *Am J Med* 1987; **82**: 231-6.

10 Verkerk PH, Derksen-Lubsen G, Vulsma T, *et al.* Evaluation of a decade of neonatal screening for congenital hypothyroidism in the Netherlands. *Nederlands Tijdschrift voor Geneeskunde* 1993; **137**: 2199-205.

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

12 Catalona WJ, Hudson MA, Scardino PT, *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, *et al.* Diagnosis of iron deficiency anemia 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 Anonymous. How good is that test—using the result. *Bandolier* 1996; **3**: 6-8.

Chapter 8: Papers that summarise other papers (systematic reviews and meta-analyses)

8.1 When is a review systematic?

Remember the essays you used to write when you first started college? You would mooch round the library, browsing through the indexes of books and journals. When you came across a paragraph that looked relevant you copied it out, and if anything you found did not fit in with the theory you were proposing, you left it out. This, more or less, constitutes the methodology of the *narrative* review—an overview of primary studies that have not been identified or analysed in a systematic (that is, standardised and objective) way. Journalists, who get paid according to how much they write rather than how much they read or how critically they process it, take the narrative review to its most selective extreme, which explains why most of the "new scientific breakthroughs" you read in your newspaper today will probably be discredited before the year is out. In contrast, a *systematic review* is an overview of primary studies that

- Contains an explicit statement of objectives, materials, and methods
- Has been conducted according to explicit and reproducible methodology (see figure 8.1)[1]

The most enduring and useful systematic reviews, notably those undertaken by the Cochrane Collaboration (see section 2.10), are regularly updated to incorporate new evidence.