THE EPIDEMIOLOGY OF CLINICAL TESTS

Adrian Bauman, Department of Public Health, University of Sydney

Diagnostic and screening tests are used every day by clinicians, but the epidemiological principles underlying them are not well understood. Tests are used to (i) make diagnoses (or increase the probability of a diagnosis); (ii) judge the severity of an illness; and (iii) predict its clinical course or likely response to treatment. Tests should be chosen based on the useful information that they provide to clinicians, and an explanation of the principles behind diagnostic tests is the theme of this paper.

First, it is important to distinguish between *diagnostic* and *screening* tests. The latter are tests applied to asymptomatic healthy subjects from the general population. An example might be population screening for cholesterol. Diagnostic tests are used in patients with specific symptoms to explain or investigate their most likely cause. Diagnostic tests are particularly useful when clinical uncertainty exists — to 'rule in' or exclude the likelihood of a particular diagnosis. However, sound clinical judgement about suspicious signs and symptoms remains the best indicator for the use of diagnostic tests.

The 'best' test to use is a pre-determined 'gold standard'. The gold standard reflects the 'true' diagnosis in that patient. In many situations, establishing the true diagnosis may be invasive, expensive and potentially unrealistic — thus we often use diagnostic tests rather than gold standard measurements. Examples of some screening and diagnostic tests and possible 'gold standards' are listed in Table 1.

There is not always a gold standard which provides 100% certainty about diagnosis. For example, angina pectoris is still a clinical diagnosis. In the assessment of diabetes, haemoglobin A_1C may not be the 'gold standard' marker of control that was initially thought. For some conditions such as childhood asthma, it may be difficult to determine which of several markers is the gold standard — symptoms, lung function variability or the results of a bronchial provocation challenge test.

Epidemiological approaches are useful in assisting clinicians in determining the quantitative value of a test in confirming a diagnosis where we have a gold standard for comparison. Epidemiology does not estimate the economic and social costs of missing a case, or the costs of over-investigation. Sometimes more expensive tests such as abdominal CT scans may be more sensitive than abdominal ultrasound in detecting pancreatic lesions. Otherwise simple tests such as a rectal examination for indurated nodules in the prostate may be more sensitive than a serum acid phosphatase level.

Table 1

Examples of screening and diagnostic tests and possible gold standards

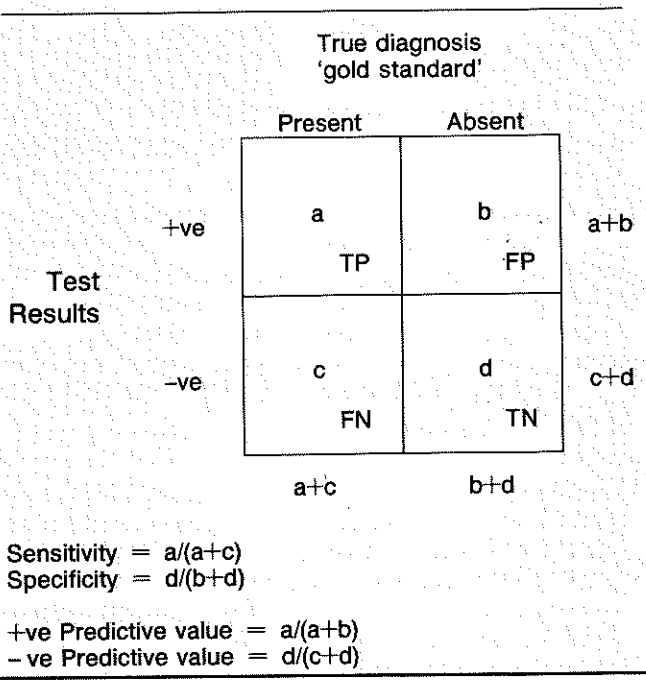
Disease	*Tests*	*Gold standard*
Urinary tract infection	urine microscopy	urine culture
Congenital heart disease	exercise ECG	coronary angiography
Myocardial infarction	ECG or cardiac enzymes	cardiac biopsy (only available at autopsy)
Breast cancer	mammography	biopsy result
Bowel cancer	Hemoccult test of stool	colonoscopy +/- biopsy
Hypertension	blood pressure (Korotkoff sounds)	intra-arterial measurement of pressures

The sensitivity and specificity of tests

The concepts of sensitivity and specificity help us to explore the relationship between a diagnostic test and the (true) presence or absence of disease. The principles are outlined in Fig. 1 and discussed in Sackett and Haynes.[1] Note that the

Fig. 1

Estimating the sensitivity and specificity of diagnostic tests



total number of cases who truly have disease is (a+c), and truly without disease is (b + d). However, (a+b) are test positive and (c+d) are test negative.

The sensitivity is defined as the proportion of people with disease who have a positive test — a/(a + c). A test which is very sensitive will rarely miss people with the disease. It is important to choose a sensitive test if there are serious consequences for missing the disease. Treatable malignancies (*in situ* cancers or Hodgkin's disease) should be found early — thus sensitive tests should be used in their diagnostic work-up.

Specificity of a test is defined as the proportion of people without the disease who have a negative test result — d/(b+d). A specific test will have few false positive results — it will rarely misclassify people without the disease as being diseased. If a test is not specific, it may be necessary to order additional tests to rule in a diagnosis.

In Example A (Fig. 2), the test is exercise ECG and the gold standard is angiographically defined coronary artery stenosis. Data from 100 fictitious clinic patients are presented, of whom 50 were subsequently found to have coronary stenosis. Of the 50 with the disease, the test recorded 30 as positive (true positives) and 20 as test negative (false negatives). The sensitivity of the test was 0.6 (i.e. 30/50). Of the 50 without disease, the test identified 45 as test negative, giving a specificity of 45/50 or 0.9.

One of the issues here is that of defining normal levels for continuous physiological variables, such as cholesterol, blood pressure and serum chemistry tests. We usually dichotomise such parameters into 'abnormal' and 'normal', based on an arbitrary cut-point. The cut-point is determined based on a trade-off between sensitivity and specificity — an increase in the sensitivity of a test is associated with a reduction in its specificity (and vice versa). Where we define the cut-point depends on what Se/Sp trade-off we wish to make.

An example of this trade-off is the cut-point for abnormal blood sugar levels (BSL) above which levels the diagnosis of diabetes becomes likely. Usually, we set a BSL of 8 mmol/L (fasting) or 11 mmol/L (post-prandial), above which we suspect diabetes. At these cut-points, the sensitivity is about 57% and specificity 99%.

If we chose a cut-point of 5 mmol/L, the sensitivity would be 98%, but the specificity would be less than 25% — very few people would be missed, but many normal people would be falsely labelled as positive (diabetic). Similarly, if we set our BSL cut-point at > 13 mmol/L, the test would have a perfect specificity (100%), but many true diabetics would be missed by the test (a high false negative rate).

In reality, we set diagnostic cut-points based on the trade-off between sensitivity and specificity. We also use epidemiological evidence for risk e.g. recent evidence has led to a lowering of the suggested 'abnormal' cholesterol value from 6.5 to 5.5 mmol/L.

The use of predictive values

Clinicians are principally interested in the interpretation of a test result. Thus, the clinical utility of sensitivity and specificity are limited by the fact that you need outcome (gold standard) data to calculate them. Clinicians need to know how likely disease is, given the result of their test. Predictive values are useful here.

The positive predictive value (+PV) of a test is defined as the proportion of patients with positive test results who truly have the disease, or algebraically from Fig. 1, a/(a+b). From Example A in Fig. 2, the +PV 30/35 (0.86) is very good. The likelihood of coronary artery disease is very high given a positive exercise ECG test. The negative predictive value (–PV) is the proportion of patients with a negative test who do not have the disease, calculated as d/(c+d). In Example A, the –PV is 45/65 (0.69).

One important clinical issue is to be aware of the approximate prevalence of the problem in your practice population. The predictive values of a test vary with the underlying prevalence of the disease in the target population. This is illustrated in Fig. 2 by comparing Example B with Example A. In Example B, a community sample of 1000 was screened using an exercise ECG and the true disease status also determined by angiography. The sensitivity and specificity of the test are identical in both Examples A and B (Se = 0.6, Sp = 0.9). However, the positive predictive value, which was diagnostically useful at 0.86 in Example A, has fallen to 0.38 in the screened population in Example B (60/160). Thus, in

Fig. 2
The use of epidemiological tests in clinic and community populations

Example A
CLINIC Sample

Exercise ECG 'Test'	Angiographic Coronary Artery Stenosis 'gold standard' Present	Absent
+ve	30	5
–ve	20	45
	50	50

Example B
COMMUNITY Sample

Exercise ECG 'Test'	Angiographic Coronary Artery Stenosis 'gold standard' Present	Absent
+ve	60	100
–ve	40	800
	100	900

the unselected community sample where the underlying prevalence of disease was only 100/1000 (or 10%) the proportion of patients with a positive test who truly had the disorder had fallen to 38% — in other words, in this sample, conducting the test did not contribute to diagnostic certainty about the presence of atherosclerotic disease.

Even if the test were very specific, a low prevalence of disease in the underlying population would produce a low positive predictive value. Test positive results in this setting will be largely false positives (100/160 in Example B).

Clinical judgement and examination increases the 'prevalence of disease'. Thus, exercise ECGs applied to the unselected community would imply a low prevalence and poor +PV. Clinical judgements can increase the prevalence — if one targeted middle aged males who smoked and were hypertensive, then the yield from an exercise ECG test would be much higher (a +PV approaching that in Example A).

This process of 'increasing the likelihood of disease' (prevalence) selects subjects so that diagnostic tests are more useful. Clinical signs and a history of post-prandial pain will be needed before gastric endoscopy is recommended. More detailed test results will be essential before submitting patients to a liver biopsy to diagnose chronic active hepatitis — the preliminary tests, including LFTs and ultrasound results increase the likelihood of hepatitis and make the (potentially hazardous) biopsy test worthwhile.

Conclusion

It is useful for clinicians to know the sensitivity and specificity of common tests to help in deciding which tests to use to 'rule in' or 'rule out' disease. However, predictive values are of more direct clinical usefulness, enabling the clinician to estimate the probability of disease *given* the test result. One problem is that predictive values are prevalence dependent, but the prevalence (likelihood) of disease can be increased by clinical signs, other tests and even clinical 'intuition'.

Finally, clinical signs and judgement should never be ignored in the face of a technological test result. As Langlands has recently pointed out[2], a negative mammogram should be ignored if a clinical breast lump remains palpable. In such circumstances, clinical judgement should suggest definitive biopsy, even though the test result was negative. Tests are to be used to assist clinicians, not to rule clinical decision-making.

REFERENCES

1. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown, 1985.
2. Walker QJ, Langlands AO. The misuse of mammography in the management of breast cancer. Med J Aust 1986;145:185–7.