Clinical tests: sensitivity and specificity

Abdul Ghaaliq Lalkhen MB ChB FRCA
Anthony McCluskey BSc MB ChB FRCA

Key points

Sensitivity and specificity are terms used to evaluate a clinical test. They are independent of the population of interest subjected to the test.

Positive and negative predictive values are useful when considering the value of a test to a clinician. They are dependent on the prevalence of the disease in the population of interest.

The sensitivity and specificity of a quantitative test are dependent on the cut-off value above or below which the test is positive. In general, the higher the sensitivity, the lower the specificity, and vice versa.

Receiver operator characteristic curves are a plot of false positives against true positives for all cut-off values. The area under the curve of a perfect test is 1.0 and that of a useless test, no better than tossing a coin, is 0.5.

Many clinical tests are used to confirm or refute the presence of a disease or further the diagnostic process. Ideally such tests correctly identify all patients with the disease, and similarly correctly identify all patients who are disease free. In other words, a perfect test is never positive in a patient who is disease free and is never negative in a patient who is in fact diseased. Most clinical tests fall short of this ideal.

Sensitivity, specificity, and other terms

The following terms are fundamental to understanding the utility of clinical tests:

1. *True positive*: the patient has the disease and the test is positive.
2. *False positive*: the patient does not have the disease but the test is positive.
3. *True negative*: the patient does not have the disease and the test is negative
4. *False negative*: the patient has the disease but the test is negative.

When evaluating a clinical test, the terms sensitivity and specificity are used. They are independent of the population of interest subjected to the test. The terms positive predictive value (PPV) and negative predictive value (NPV) are used when considering the value of a test to a clinician and are dependent on the prevalence of the disease in the population of interest.

Sensitivity

The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease.

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

A test with 100% sensitivity correctly identifies all patients with the disease. A test with 80% sensitivity detects 80% of patients with the disease (true positives) but 20% with the disease go undetected (false negatives). A high sensitivity is clearly important where the test is used to identify a serious but treatable disease (e.g. cervical cancer). Screening the female population by cervical smear testing is a sensitive test. However, it is not very specific and a high proportion of women with a positive cervical smear who go on to have a colposcopy are ultimately found to have no underlying pathology.

Specificity

The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease.

$$\text{Specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}$$

Therefore, a test with 100% specificity correctly identifies all patients without the disease. A test with 80% specificity correctly reports 80% of patients without the disease as test negative (true negatives) but 20% patients without the disease are incorrectly identified as test positive (false positives).

As discussed above, a test with a high sensitivity but low specificity results in many patients who are disease free being told of the possibility that they have the disease and are then subject to further investigation. Although the ideal (but unrealistic) situation is for a 100% accurate test, a good alternative is to subject patients who are initially positive to a test with high sensitivity/low specificity, to a second test with low sensitivity/high specificity. In this way, nearly all of the false positives may be correctly identified as disease negative.

Positive predictive value

The PPV of a test is a proportion that is useful to clinicians since it answers the question: 'How likely is it that this patient has the

Abdul Ghaaliq Lalkhen MB ChB FRCA
Specialist Registrar
Salford Royal Hospitals NHS Trust
Hope Hospital
Salford M6 8HD
UK

Anthony McCluskey BSc MB ChB FRCA
Consultant
Department of Anaesthesia
Stockport NHS Foundation Trust
Stepping Hill Hospital
Stockport SK2 7JE
UK
Tel: +44 (0161) 419 5869
Fax: +44 (0161) 419 5045
E-mail: a.mccluskey4@ntlworld.com
(for correspondence)

doi:10.1093/bjaceaccp/mkn041
Continuing Education in Anaesthesia, Critical Care & Pain | Volume 8 Number 6 2008

disease given that the test result is positive?'

$$\text{Positive predictive value} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Negative predictive value

The NPV of a test answers the question: 'How likely is it that this patient does not have the disease given that the test result is negative?'

$$\text{Negative predictive value} = \frac{\text{True negatives}}{\text{True negatives} + \text{False negatives}}$$

Likelihood ratio

A final term sometimes used with reference to the utility of tests is the likelihood ratio. This is defined as how much more likely is it that a patient who tests positive has the disease compared with one who tests negative.

$$\text{Likelihood ratio} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

Dependence of PPV and NPV on disease prevalence

Unlike sensitivity and specificity, the PPV and NPV are dependent on the population being tested and are influenced by the prevalence of the disease. Consider the following example: screening for systemic lupus erythematosis (SLE) in a general population using the antinuclear antibody has a low PPV because of the high number of false positives it yields. However, if a patient has signs of SLE (e.g. malar flush and joint pain), the PPV of the test increases because the population from which the patient is drawn is different (from a general population with a low prevalence of SLE to a clinically suspicious population with a much higher prevalence).

We may also consider a woman who presents with breathlessness post-partum and where one of the differential diagnoses is pulmonary embolism. A D-dimer test would almost certainly be elevated in this patient population; therefore, the test has a low PPV for pulmonary embolism. However, it has a high NPV for pulmonary embolism since a low D-dimer is unlikely to be associated with pulmonary embolism.

The dependence of PPV and NPV on the prevalence of a disease can be illustrated numerically: consider a population of 4000 people who are divided equally into the ill and the well. A screening test to detect the condition has a sensitivity of 99% and a specificity of 99%. Screening this population would therefore yield 1980 true positives and 1980 true negatives with 20 patients being tested positive when they in fact are well and 20 patients testing negative when they are ill. Therefore, the PPV of this test is 99%. However, if the number of ill people in the population is only 200 and the number of well people is 3800, the number of false positives increases from 20 to 38 and the PPV falls to 84%.

This discussion highlights the fact that the ability to make a diagnosis or screen for a condition depends both on the discriminatory value of the test and on the prevalence of the disease in the population of interest. If the data relating to a test are inserted into a 2×2 contingency table, the Fisher's exact test of many statistical software packages may be used to calculate sensitivity, specificity, PPV, NPPV, and likelihood ratio.

Receiver operator characteristic curves

Consider the following hypothetical example: measurement of high endorphin levels in SpRs in Anaesthesia has been found to be associated with success in the final FRCA examination. A sample of SpRs is tested before the examination resulting in a range of endorphin values. The data are examined and an arbitrary cut-off point for endorphin levels is chosen above which most of the candidates passed with few failures. Despite choosing the cut-off value in such a way that the maximum possible number of SpRs is correctly classified, we may find that 10% of the cohort with endorphin levels above the cut-off level failed the exam (false positives) and 15% of the cohort with endorphin levels below the cut-off level passed the exam (false negatives).

The relatively crude measures of sensitivity and specificity discussed previously fail to take into account the cut-off point for a particular test. If the cut-off point is raised, there are fewer false positives but more false negatives—the test is highly specific but not very sensitive. Similarly, if the cut-off point is low, there are fewer false negatives but more false positives—the test is highly sensitive but not very specific.

Receiver operator characteristic curves (so called because they were originally devised by radio receiver operators after the attack on Pearl Harbour to determine how the US radar had failed to detect

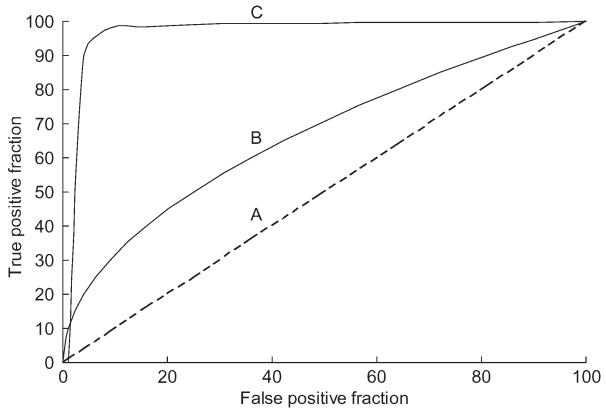

Fig 1 Receiver operator curves: (A) line of zero discrimination (AUC=0.5); (B) typical clinical test (AUC=0.5–1.0); perfect test (AUC=1.0).

the Japanese aircraft) are a plot of (1 − specificity) of a test on the *x*-axis against its sensitivity on the *y*-axis for all possible cut-off points. An identical plot is produced when the false positive rate of a test is shown on the *x*-axis against the true positive rate on the *y*-axis (Fig. 1). An ideal test is represented by the upper curve in the figure. The middle curve represents the characteristics of a test more typically seen in routine clinical use. The area under this curve (AUC) represents the overall accuracy of a test, with a value approaching 1.0 indicating a high sensitivity and specificity. The dotted line on the graph represents the line of zero discrimination with an AUC of 0.5 (the test is no better than tossing a coin).

Acknowledgements

The authors are grateful to Professor Rose Baker, Department of Statistics, Salford University for her valuable contribution in providing helpful comments and advice on this manuscript.

Bibliography

1. Bland M. *An Introduction to Medical Statistics*, 3rd Edn. Oxford: Oxford University Press, 2000
2. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1991
3. Rumsey D. *Statistics for Dummies*. New Jersey: Wiley Publishing Inc., 2003
4. Swinscow TDV. *Statistics at Square One*. Available from http://www.bmj.com/statsbk/ (accessed 20 October 2008)
5. Greenhalgh T. *How to Read a Paper*. London: BMJ Publishing, 1997
6. Elwood M. *Critical Appraisal of Epidemiological Studies and Clinical Trials*. 2nd Edn. Oxford: Oxford University Press, 1998

Please see multiple choice questions 19–22