Emergency Medicine (1999) **11**, 175–181

Emergency Medicine

REVIEW ARTICLE: STATISTICS

An introduction to sensitivity, specificity, predictive values and likelihood ratios

Kevin Chu
Department of Emergency Medicine, Royal Brisbane Hospital, Herston Road, Brisbane, Queensland, Australia

Abstract

Objective: To assist clinicians and other health-care providers to understand the terms used to describe the accuracy of diagnostic tests.

Methodology: This paper reviews the calculation and interpretation of sensitivity, specificity, predictive values, receiver operating characteristic curves and likelihood ratios.

Results: Sensitivity and specificity are measures of the accuracy of a diagnostic test. There is a trade-off between sensitivity and specificity that is dependent on the cut-off level chosen for a positive diagnosis. Predictive values are measures of the usefulness of a test once the test results are known. Predictive values depend on the prevalence of the disease, as well as on the sensitivity and specificity of the test. Sensitivity, specificity and predictive values are easily calculated by the construction of a two-by-two table. Multiple testing, either in parallel or in series, can alter the sensitivity, specificity and predictive values. Receiver operating characteristic curves plot the relationship between sensitivity and specificity at various cut-offs, facilitating the comparison of accuracy among tests. Likelihood ratios are an alternative measure of accuracy and have the advantage of being able to assess multiple test outcomes, rather than simply assessing dichotomized positive or negative results.

Conclusion: An understanding of how the measurements that are used to describe the accuracy of diagnostic tests are calculated is essential to the interpretation of test results.

Key words: *likelihood ratios, multiple testing, negative predictive values, positive predictive values, receiver operating characteristic curves, sensitivity, specificity.*

Introduction

In the practice of emergency medicine, the emergency physician relies on his or her expertise in history taking, physical examination and selective investigation to make a diagnosis, while often providing treatment simultaneously. The history, examination and investigations can be thought of each, or in combination, as a diagnostic test. The accuracy of the test is judged by its sensitivity, specificity and predictive values, which are quoted at journal clubs and clinical meetings, sometimes without a clear

Correspondence: Dr Kevin Chu, Department of Emergency Medicine, Royal Brisbane Hospital, Herston Road, Brisbane Qld 4029, Australia. Email: <chuk@health.qld.gov.au>

Kevin Chu, MB, BS, MS, FACEM, Staff Specialist in Emergency Medicine.

understanding of how these elements are calculated. This paper reviews the fundamental principles behind sensitivity, specificity, predictive values and related concepts, including multiple testing, receiver operating characteristic curves and likelihood ratios. Further discussion on the interpretation of diagnostic studies may be found in the evidence-based medical literature.[1–3]

Tests and diagnosis

A diagnosis is made using a test, a composite of history, examination and investigations. The most accurate test is referred to commonly as the *gold standard*. The gold standard test may, however, be scarce, lengthy, hazardous or expensive. An example of a gold standard test is an angiogram that is used to diagnose traumatic aortic transection in a patient presenting to the emergency department (ED) with a widened mediastinum on chest X-ray. Alternative tests are developed to overcome problems associated with a current gold standard; for example, a transoesophageal echocardiogram is an alternative test to angiogram for making the diagnosis of aortic transection.[4] The accuracy of alternative tests is judged by comparing the test with the gold standard.

Sensitivity and specificity

Test results have four possible interpretations: two correct (or true) and two incorrect (or false) as depicted by the two-by-two table (Fig. 1).

Sensitivity is defined as the proportion of patients with the disease who have a positive test [1]. Sensitivity is calculated by the formula:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad [1]$$

Specificity is defined as the proportion of patients without the disease who have a negative test [2]. Specificity is calculated by the formula:

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \quad [2]$$

For a test to be accurate, it must be both highly sensitive and highly specific. Decreasing the number of false negatives increases the sensitivity. Using a highly sensitive test is essential when it is important not to miss a particular diagnosis, such as when acute myocardial infarction (AMI) is suspected. Decreasing the number of false positives increases the specificity. Using a highly specific test is important when mislabelling a non-diseased patient as diseased could be detrimental, resulting in unnecessary invasive and expensive investigations and treatments for the patient.

	Disease present	Disease absent
Test positive	True positive	False positive
Test negative	False negative	True negative

Figure 1. Four possible interpretations of test results.

While a highly sensitive and highly specific test is desirable there is ususally a trade-off between sensitivity, and specificity: as one increases, the other decreases. For example, a particular plasma concentration of a biochemical marker is selected for the diagnosis of AMI. The test is considered positive if the patient's level is above the cut-off level and negative if the patient's level is below the cut-off. Decreasing the cut-off decreases the number of false negatives and, thus, increases the sensitivity of the test. However, decreasing the cut-off will also increase the number of false positives and, thus, decrease the specificity of the test. Increasing the cut-off will have the opposite effect, that is, decreasing the sensitivity while increasing the specificity.

Even after balancing sensitivity and specificity, an alternative test can never be more sensitive or specific than the gold standard, when the gold standard is used to define whether a disease is present or absent. This is true even if the alternative test is apparently more accurate. A more definitive means of diagnosis, for example, an autopsy, is required to demonstrate the improved accuracy of an alternative test compared with the current gold standard.

In the interpretation of sensitivity and specificity, it should be noted that percentages are frequently estimated from a small number of patients. Given that the number of patients is small, the sensitivity and specificity can fall within a range of values. The range or precision of the estimate of sensitivity and specificity is given by the 95% confidence interval.[5,6] The precision increases, that is, the 95% confidence interval shortens, as the sample size increases. In addition, sensitivity and specificity are known as the

true positive rate and the true negative rate, respectively. Two other proportions, namely, the false positive rate and the false negative rate, are sometimes reported and may be calculated (Fig. 2.).

$$\text{True positive rate} = \frac{TP}{TP + FN} = \text{sensitivity}$$

$$\text{True negative rate} = \frac{TN}{FP + TN} = \text{specificity}$$

$$\text{False positive rate} = \frac{FP}{FP + TN} = 1 - \text{specificity}$$

$$\text{False negative rate} = \frac{FN}{TP + FN}$$

$$\text{Positive predictive value} = \frac{TP}{TP + FP}$$

$$\text{Negative predictive value} = \frac{TN}{TN + FN}$$

Figure 2. Summary of terms used to describe the accuracy of a diagnostic test. TP, true positive; FP, false positive; TN, true negative; FN, false negative.

Predictive values

Sensitivity and specificity convey information as to whether a test is useful in making a diagnosis. Once the test has been performed, sensitivity and specificity do not indicate whether a positive result truly means the presence of disease. That information is given by the predictive values.

The positive predictive value (PPV) is defined as the proportion of patients with a positive test who have the disease [3]. Positive predictive value is calculated by the formula:

$$\text{Positive predictive value} = \frac{\text{True positive}}{\text{True positive + False positive}} \quad [3]$$

Negative predictive value (NPV) is defined as the proportion of patients with a negative test who do not have the disease [4]. Negative predictive value is calculated by the formula:

$$\text{Negative predictive value} = \frac{\text{True negative}}{\text{True negative + False negative}} \quad [4]$$

It is vital to note that predictive values vary with the prevalence of the disease. The prevalence of a disease is defined as the proportion of the population that has the disease at a given time. This is best illustrated by working through the following examples [5] [6]:

Example 1:
Intermediate disease prevalence

	Disease present	Disease absent	Total
Test positive	90	20	110
Test negative	10	80	90
Total	100	100	200

$$\text{Sensitivity} = \frac{90}{90 + 10} = 0.90$$

$$\text{Specificity} = \frac{80}{80 + 20} = 0.80$$

$$\text{Prevalence} = \frac{100}{200} = 0.50$$

$$\text{PPV} = \frac{90}{90 + 20} = 0.82$$

$$\text{NPV} = \frac{80}{10 + 80} = 0.89 \quad [5]$$

Example 2:
Low disease prevalence

	Disease present	Disease absent	Total
Test positive	9	38	47
Test negative	1	152	153
Total	10	190	200

$$\text{Sensitivity} = \frac{9}{9 + 1} = 0.90$$

$$\text{Specificity} = \frac{152}{38 + 152} = 0.80$$

$$\text{Prevalence} = \frac{10}{200} = 0.05$$

$$\text{PPV} = \frac{9}{9 + 38} = 0.19$$

$$\text{NPV} = \frac{152}{1 + 152} = 0.99 \quad [6]$$

The sensitivity and specificity in the two examples are the same; however, changing the prevalence drastically alters the positive predictive value. In fact, both positive and negative predictive values depend on the prevalence of the disease in the population studied. As the prevalence falls, PPV decreases and NPV increases. As the prevalence rises, PPV increases and NPV decreases. A practical implication of this is the specialist who argues that a given test has a higher PPV in his or her hands. However, because the prevalence of the disease is much higher in the population of patients referred to him or her by other doctors (who have already screened out some patients prior to referral), this is hardly surprising. Similarly, a test with a constant sensitivity and specificity may have different predictive values when performed in the ED compared with general practice, or performed among different demographic groups. If the disease being assessed is uncommon, the PPV will be low (Table 1), even when a test is highly sensitive and specific; hence, most positive tests will be false. Thus, an exercise stress test is more meaningful in older men who smoke than in younger women with atypical chest pain.[7]

Examples 1 and 2 revealed that the predictive values estimated from a study are only of value if the prevalence in the study population is the same as the population to which the results will be applied. If the local prevalence is different from that reported in the literature, the reported predictive values cannot be applied locally. If the local prevalence (prev) is known, however, local predictive values can be calculated using the sensitivity (sens) and specificity (spec) for the test reported in the literature. The calculation is performed using a formula derived from conditional probability and is known as the Bayes's Theorem [7]:[8]

$$\text{PPV} = \frac{\text{Sens} \times \text{Prev}}{\text{Sens} \times \text{Prev} + (1 - \text{Spec}) \times (1 - \text{Prev})}$$

$$1 - \text{NPV} = \frac{(1 - \text{Sens}) \times \text{Prev}}{(1 - \text{Sens}) \times \text{Prev} + \text{Spec} \times (1 - \text{Prev})} \quad [7]$$

Thus, using Bayes's Theorem, if the reported sensitivity and specificity are 0.90 and 0.80, respectively, and it was determined that the local prevalence was 0.25 (rather than 0.50), the calculated local PPV and NPV are 0.60 and 0.97 (rather than 0.82 and 0.89), respectively.

Table 1. Prevalence and predictive values for a test with 75% sensitivity and specificity

Prevalence	Positive predictive value	Negative predictive value
0.1	0.250	0.964
0.01	0.029	0.997
0.001	0.003	1.000

Multiple testing

A clinician usually requires more than one test to make a diagnosis. These tests may be performed in parallel or in a series. Parallels tests, such as an electrocardiogram and biochemical markers for the diagnosis of AMI, are performed simultaneously. Parallel testing is used commonly in ED when rapid assessment is necessary. Any one of the tests can be positive to make the diagnosis. Parallel testing decreases the number of false negatives, thus increasing the sensitivity and NPV compared with each individual test. Parallel testing also increases the number of false positives, thus decreasing specificity and PPV. Serial tests are performed in sequence and all must be positive to make the diagnosis. Serial testing decreases the number of false positives, thus maximizing specificity and PPV.

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves are used to compare the accuracy of two or more tests over a range of cut-offs. A ROC curve for a given test is constructed by plotting sensitivity (true positive rate) against 1 – specificity (false positive rate). The points on a curve represent pairs of sensitivity and specificity at the various cut-off levels selected for a given positive diagnosis. A different ROC curve is constructed for each test (Fig. 3). The accuracy of a test is judged qualitatively by the position of its ROC curve. The diagonal line (Fig. 3) represents a test that does not

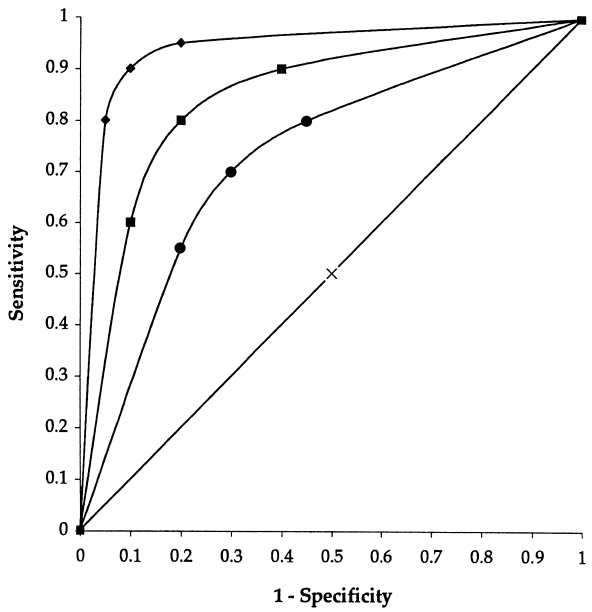

Figure 3. Receiver operating characteristic curves for the diagnostic tests A (♦), B (■) and C (●).

discriminate diseased from non-diseased patients. Receiver operating characteristic curves that lie further to the left of the diagonal line represent the more accurate tests. For example, test A is more accurate than test B if test A's ROC curve lies to the left of test B's. Conversely, test C is less accurate than test B if test C's ROC curve lies to the right of test B's. The accuracy of a test is judged quantitatively by the area under the ROC curve. The greater the area under the curve, the more accurate the test. More detailed discussions of ROC curves are available.[9]

The selection of a cut-off level for clinical use means a trade-off between sensitivity and specificity. An appropriate cut-off for a particular test is usually that value associated with the pair of sensitivity and specificity on the 'shoulder' of the ROC curve;[10] that is, closest to the left upper corner where sensitivity and specificity approach 100%. This cut-off must balance the cost of missing a diagnosis (false negative) against the cost of mislabelling a non-diseased patient as diseased (false positive).

Likelihood ratios

Likelihood ratios (LR) are an alternative, newer method of judging the accuracy of a test. Some authors argue that LR are more useful than sensitivity and specificity.[2] Likelihood ratios compare the proportion of patients with the disease that has the positive (or negative) test result with the proportion of patients without the disease that has the positive (or negative) test result. The likelihood ratio is the ratio of these two proportions or *likelihoods* [8].

Example 3:

	Disease present		Disease absent		
	No.	Proportion	No.	Proportion	Total
Test positive	90	90/100	20	20/100	110
Test negative	10		80		90
Total	100		100		200

$$\text{LR for a positive test} = \frac{90/100}{20/100} = 4.5 \quad [8]$$

A LR of 4.5 indicates that the test is 4.5 times more likely to be positive in patients with the disease compared with patients without the disease. The advantage of using LR is that more detailed clinical information is not ignored because the test results may be divided into two or more outcomes, rather than simply dichotomized into positive or negative [9].

Example 4:

	Disease present		Disease absent		
	No.	Proportion	No.	Proportion	Total
Test strongly positive	60	60/100	5	5/100	75
Test weakly positive	30	30/100	15	15/100	35
Test negative	10		80		90
Total	100		100		200

$$\text{LR for a strongly positive test} = \frac{60/100}{5/100} = 12$$

$$\text{LR for a weakly positive test} = \frac{30/100}{15/100} = 2 \quad [9]$$

Strongly and weakly positive tests are 12 and two times, respectively, more likely to be positive in patients with the disease compared with patients without the disease. This provides more clinical information than stating that a positive test is 4.5 times more likely to be positive in patients with the disease

compared with patients without the disease. The assessment of the accuracy of a test is more detailed when test results can be divided into more than two outcomes. An example of where a test result is divided into two or more outcomes is ventilation-perfusion scans for the diagnosis of pulmonary embolism.

Another advantage of using LR is that it can be used to convert the pretest probability of disease into the post-test probability of disease. The pretest probability is the prevalence or the probability of disease, given a clinical presentation before the test result is known. An experienced clinician will assess a patient's probability of a positive diagnosis. A pretest probability is converted to a post-test probability using the following three steps:

1. Convert the pretest probability into pretest odds. Odds are commonly used in daily conversation; for example, a team having an 80% probability of winning is translated into four-to-one odds in favour of winning.

$$\text{Pretest odds} = \frac{\text{Pretest probability}}{1 - \text{pretest probability}}$$

2. Multiply the pretest odds by the LR to get the post-test odds

$$\text{Post-test odds} = \text{Pretest odds} \times \text{likelihood ratio}$$

3. Convert the post-test odds into post-test probability

$$\text{Post-test probability} = \frac{\text{Post-test odds}}{1 + \text{post-test odds}}$$

A test is considered useful if it significantly increases or decreases the pretest to post-test probability of disease.

Example 5:

Given:

1. Prevalence or pretest probability of disease = 0.4
2. Likelihood ratio = 12

Calculation of post-test probability:

1. Pretest odds = 0.4/(1–0.4) = 0.67
2. Post-test odds = 0.67 × 12 = 8
3. Post-test probability = 8/(1 + 8) = 0.89

Increasing the pretest to post-test probability from 40% to 89% makes the test useful. A test is not considered useful if it cannot significantly increase or decrease the pretest to post-test probability.

Example 6:

Given:

1. Prevalence or pretest probability of disease = 0.4
2. Likelihood ratio = 1.2

Calculation of post-test probability:

1. Pretest odds = 0.4/(1–0.4) = 0.67
2. Post-test odds = 0.67 × 1.2 = 0.80
3. Post-test probability = 0.8/(1 + 0.8) = 0.44

Increasing the pretest to post-test probability from 40% to only 44% does not make the test clinically useful. When the pretest probability is already high, for example, over 90%, a test will not increase the pretest to post-test probability significantly no matter how accurate is that test. Thus, when an experienced clinician's intuition or clinical judgement about the diagnosis of acute appendicitis is 90% or higher, further investigations will not significantly improve the post-test probability; therefore, a full blood count clearly is unhelpful. An accurate test is most useful when the pretest probability is between 40% and 60%.[3] When the pretest probability cannot be estimated accurately, a range may be specified to allow calculation of a range of post-test probabilities. Finally, the size of the LR most likely to significantly increase or decrease the pre-test to post-test probability is given in Table 2.

Conclusion

An understanding of the methods used to calculate sensitivity, specificity and predictive values is essential to enable interpretation of the accuracy of a diagnostic test. These calculations require the construction of a simple two-by-two table. More involved calculations are aided by a computer spreadsheet. Likelihood ratios and ROC curves build upon the concepts used to assess the accuracy of a diagnostic test. Ultimately, the most powerful diagnostic process will always remain careful history taking and physical examination, followed by

Table 2. Interpretation of likelihood ratios

Likelihood ratio	Changes from pretest to post-test probabilities
> 10 or < 0.01	Large, often conclusive
5–10 and 0.1–0.2	Moderate
2–5 and 0.5–0.2	Small but sometimes important
1–2 and 0.5–1	Small, rarely important

From the text of Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Med. Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for patients? JAMA *1994;* ***271:*** *703–7.*

selective discretionary testing aimed at ruling in or out a putative diagnosis. Indiscriminate, multiple testing will never compensate for a poor clinician.

Acknowledgement

The author gratefully acknowledges Clinical Associate Professor Anthony Brown for critically reading the manuscript.

Accepted 19 April 1999

References

1. Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine Working Group. 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.
2. Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for patients? *JAMA* 1994; **271:** 703–7.
3. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston: Little Brown and Company, 1991.
4. Cameron PA, Dziukas L, Hadi A, Hooper S, Tatoulis J. Aortic Transection. *Aust. N.Z. J. Surg.* 1998; **68:** 264–7.
5. Young KJ, Lewis RJ. What is confidence? Part 1. The use and interpretation of confidence intervals. *Ann. Emerg. Med.* 1997; **30:** 307–10.
6. Young KD, Lewis RJ. What is confidence? Part 2: Detailed definition and determination of confidence intervals. *Ann. Emerg. Med.* 1997; **30:** 311– 18.
7. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *N. Eng. J. Med.* 1979; **300:** 1350–8.
8. Woolson RF. *Statistical Methods for the Analysis of Biomedical Data*. New York: Wiley & Sons, 1987.
9. Grzybowski M, Younger JC. Statistical methodology III. Receiver operating characteristic (ROC) curves. *Acad. Emerg. Med.* 1997; **4:** 818–26.
10. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials*, 3rd edn. Baltimore: Williams and Wilkins, 1996.