Acta Pædiatrica ISSN 0803–5253

REVIEW ARTICLE

Understanding diagnostic tests 3: receiver operating characteristic curves

Anthony K Akobeng (tony.akobeng@cmmc.nhs.uk)

Department of Paediatric Gastroenterology, Central Manchester and Manchester Children's University Hospitals, Booth Hall Children's Hospital, Manchester, UK.

Keywords
Area under the curve, Optimal cut-off point, Receiver operating characteristic (ROC) curve, Sensitivity, Specificity

Correspondence
Dr A. K. Akobeng, Department of Paediatric Gastroenterology, Central Manchester and Manchester Children's University Hospitals, Booth Hall Children's Hospital, Charlestown Road, Blackley, Manchester M9 7AA, United Kingdom.
Tel: 0161 220 5458 | Fax: 0161 220 5072 | Email: tony.akobeng@cmmc.nhs.uk

Received
8 June 2006; revised 4 December 2006; accepted 8 December 2006.

DOI:10.1111/j.1651-2227.2006.00178.x

Abstract

The results of many clinical tests are quantitative and are provided on a continuous scale. To help decide the presence or absence of disease, a cut-off point for 'normal' or 'abnormal' is chosen. The sensitivity and specificity of a test vary according to the level that is chosen as the cut-off point. The receiver operating characteristic (ROC) curve, a graphical technique for describing and comparing the accuracy of diagnostic tests, is obtained by plotting the sensitivity of a test on the *y* axis against 1-specificity on the *x* axis. Two methods commonly used to establish the optimal cut-off point include the point on the ROC curve closest to (0, 1) and the Youden index. The area under the ROC curve provides a measure of the overall performance of a diagnostic test. In this paper, the author explains how the ROC curve can be used to select optimal cut-off points for a test result, to assess the diagnostic accuracy of a test, and to compare the usefulness of tests.

Conclusion: The ROC curve is obtained by calculating the sensitivity and specificity of a test at every possible cut-off point, and plotting sensitivity against 1-specificity. The curve may be used to select optimal cut-off values for a test result, to assess the diagnostic accuracy of a test, and to compare the usefulness of different tests.

INTRODUCTION

In the two previous articles of the series (1,2), the results of tests were simply considered as being positive ('abnormal') or negative ('normal'). However, the results of many clinical tests, for example blood glucose measurement or erythrocyte sedimentation rate are quantitative and are provided on a continuous scale. To help decide the presence or absence of disease, a cut-off point is chosen. Results which are on one side of this cut-off point, say above, may be considered abnormal while results which are below the cut-off point are regarded as normal. However, we know that in such a situation, not all patients whose results are above the cut-off point will necessarily have disease and not all those whose results fall below the cut-off will be free of disease.

The accuracy of a diagnostic test is characterized by its sensitivity and specificity. The sensitivity and specificity of a test, however, depends on the level that has been chosen as the cut-off point for normal or abnormal. The receiver operating characteristic (ROC) curve is widely accepted as a method for selecting an optimal cut-off point for a test and for comparing the accuracy of diagnostic tests (3,4). The curve is generated by plotting sensitivity of all possible cut-off points for the test on the *y* axis as a function of 1-specificity on the *x* axis.

SENSITIVITY, SPECIFICITY AND THE CUT-OFF POINT

As mentioned earlier, the sensitivity and specificity of a test depend on the level that has been chosen as the cut-off point for normal or abnormal, that is they depend on the definition of what constitutes an abnormal test. In a recent study, Brown et al. investigated the diagnostic accuracy of the peripheral white cell count (WCC) in identifying bacterial infections in febrile neonates (5). The sensitivity and specificity at various WCC count cut-offs is shown in Table 1.

You can see from this table that the sensitivity and specificity of the WCC in identifying bacterial infections in febrile neonates change according to what level was taken as the cut-off for normal or abnormal. For example, the sensitivity and specificity at a cut-off of 5×10^9 cells/l was 100% and 2%, respectively, while at a cut-off of 17×10^9 cells/l, sensitivity and specificity were, respectively, 38% and 89%. It is also clear from Table 1 that as the cut-off level of normal is increased, the sensitivity of the test decreases while specificity increases. This illustrates an important point: sensitivity and specificity are inversely related according to the choice of cut-off value. When increasing values of a test result are associated with disease, higher cut-off values are generally associated with lower sensitivities and higher specificities,

Table 1 Sensitivity and specificity of peripheral WCC for diagnosing bacterial infections in febrile neonates at different thresholds (5)

WCC threshold ($\times 10^9$ cells/l)	Sensitivity (%)	Specificity (%)
5	100	2
10	100	31
12	75	53
15	50	74
17	38	89
20	0	93
22	0	97
25	0	98

while lower cut-offs are associated with higher sensitivities and lower specificities (6). Thus changing the cut-off point to try and increase the sensitivity or specificity of a test will result in a reduction of the other.

This relationship between sensitivity and specificity has important implications. First, for any diagnostic test, we would like to select a cut-off value such that the optimal sensitivity and specificity are achieved, that is the cut-off point at which the test is most useful in helping to make the diagnosis. Second, it is obvious that sensitivity and specificity at a single cut-off value do not describe the test's performance at other potential cut-off values. Third, the selected cut-off value should be taken into account when comparing different diagnostic tests. One way of addressing all these issues is to use the ROC curve.

THE ROC CURVE

The ROC curve is a graphical technique for assessing the ability of a test to discriminate between those with disease and those without disease (3,7). ROC curves allow visual analyses of the trade-offs between the sensitivity and the specificity of a test with regard to the various cut-offs that may be used (8). The curve is obtained by calculating the sensitivity and specificity of the test at every possible cut-off point, and plotting sensitivity against 1-specificity.

Description of the ROC curve

A typical ROC curve is shown in Figure 1. By convention, sensitivity (the proportion of true positive results) is shown on the *y* axis, going from 0 to 1 (0–100%) and 1-specificity (the proportion of false positive results) is shown on the *x* axis, going from 0 to 1 (0–100%) (9,10).

As shown in Figure 1, the diagonal line on the graph going from the lower left-hand corner (0, 0) to the upper right-hand corner (1, 1) serves as a reference line, and represents the characteristics of a test which is completely useless at differentiating between those with disease and those without disease. Points along this line indicate that the test detects an equal number of true and false positives, that is it does not discriminate between those with disease and those without disease (9). A test that perfectly discriminates between diseased and non-diseased patients would yield a 'curve' that coincided with the left and top sides of the plot (3,11). In practice, however, it is unusual to have such a curve, and ROC curves usually lie between these extremes.

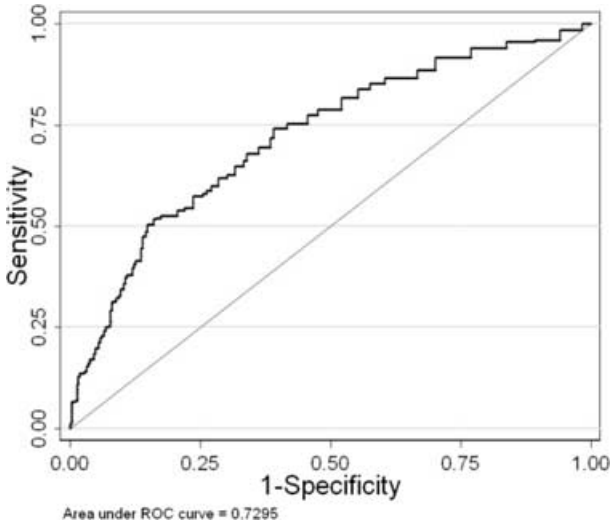

Figure 1 A typical ROC curve.

Uses of the ROC curve

The ROC curve may be used for three purposes:

1. it allows the determination of the cut-off point at which optimal sensitivity and specificity are achieved
2. it allows an assessment of the diagnostic accuracy of a test and
3. it allows the comparison of the usefulness of two or more diagnostic tests.

Determining the optimal cut-off point

A perfect medical test would have 100% sensitivity and 100% specificity, and such a test will identify all people with disease and all those without disease. The point on the ROC curve which corresponds to this perfect scenario (100% sensitivity and 100% specificity) would be at the upper left-hand corner (0, 1). In practice, however, few tests are perfect, and one has to strike a balance between sensitivity and specificity. Generally speaking, the closer the ROC curve gets to the upper left-hand corner (0, 1), the better the test is at discriminating between cases and non-cases (11).

Two methods for identifying optimal cut-off points using sensitivity, specificity and the ROC curve are commonly used (12,13). The first method assumes that the best cut-off point for balancing the sensitivity and specificity of a test is the point on the curve closest to the (0, 1) point. In this method, optimal sensitivity and specificity are defined as those yielding the minimal value for $(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2$. The cut-off point corresponding to these sensitivity and specificity values is the one closest to the (0, 1) point and is taken to be the cut-off point that best differentiates between people with disease and those without disease (12).

The second method that may be used to determine the optimal cut-off point for a test is the Youden index (*J*) (12,13).

J is defined as the maximum vertical distance between the ROC curve and the diagonal or chance line and is calculated as J = maximum {sensitivity + specificity −1}. Using this measure, the cut-off point on the ROC curve which corresponds to J, that is, at which (sensitivity + specificity − 1) is maximized, is taken to be the optimal cut-off point. An intuitive interpretation of J is that it corresponds to the point on the curve farthest from chance (12).

Assessing diagnostic accuracy

The ROC curve is also important because the area under the curve (AUC) is a reflection of how good the test is at distinguishing between patients with disease and those without disease. The AUC serves as a single measure, independent of prevalence, that summarizes the discriminative ability of a test across the full range of cut-offs (14). The greater the AUC, the better the test.

A perfect test (as described earlier) will have an AUC of 1.0, while a completely useless test (one whose curve falls on the diagonal line) has an AUC of 0.5. The AUC of many tests used in clinical practice fall between these two values. In general the closer the AUC is closer to 1, the better the overall diagnostic performance of the test, and the closer it is to 0.5, the poorer the test.

Figure 2 shows an ROC curve which is adapted from the afore-mentioned peripheral WCC count study by Brown et al. (5) The sensitivity and specificity values which were used to derive this curve at various cut-off points are shown in Table 1. As shown in the figure, the area under this curve was 0.723.

One way of interpreting the area under the ROC curve is that a test with an area greater than 0.9 has high accuracy, while 0.7–0.9 indicates moderate accuracy, 0.5–0.7, low accuracy and 0.5 a chance result (14).

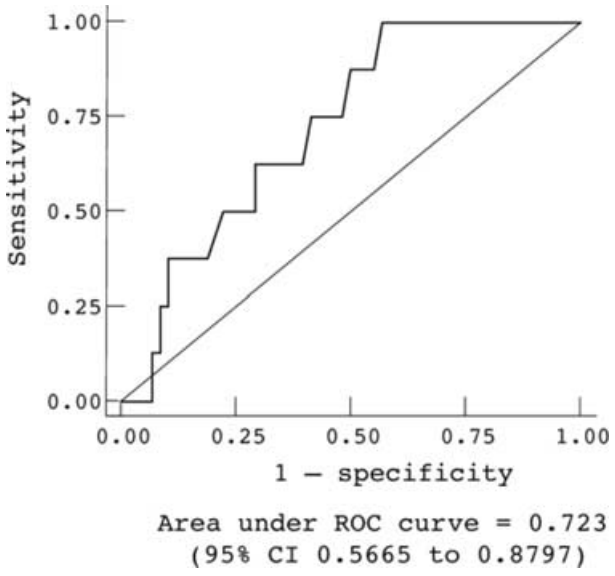

Figure 2 ROC curve for peripheral WCC for diagnosing bacterial infections in febrile neonates at different thresholds (5).

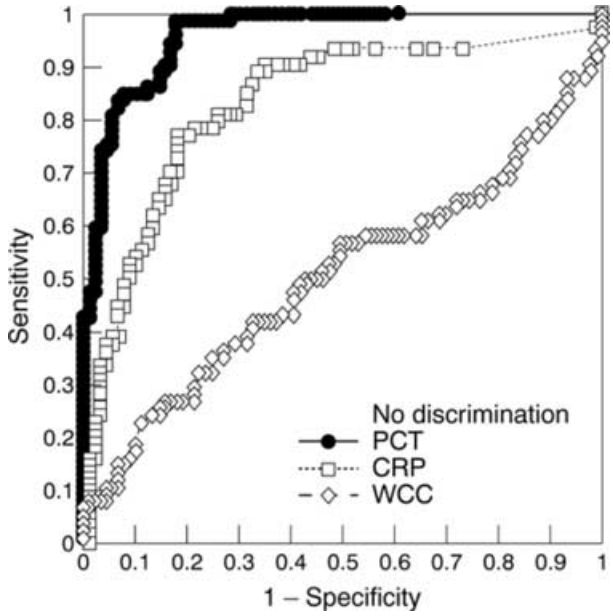

Figure 3 ROC curves comparing procalcitonin (PCT), C-reactive protein (CRP) and leucocyte count (WCC) for prediction of septic shock (15).

Comparing the usefulness of tests

The ROC curve is also very helpful when we want to compare the diagnostic accuracy of two or more tests. It helps in deciding which of the tests is better for the purpose for which they are being used. The optimal ROC curve is the one connecting the points highest and farthest to the left. The rationale for the optimal ROC curve is that one wants the highest true-positive rate (sensitivity) for the lowest false-positive rate (1-specificity).

In a study to evaluate diagnostic markers of infection in critically ill children, Hatherill and colleagues compared procalcitonin, C-reactive protein and WCC in a paediatric intensive care unit (15). ROC curves obtained for these three tests are shown in Figure 3. As you can see from this figure, the curve for procalcitonin is closer to the upper left-hand corner suggesting that it was a better test for predicting septic shock than C-reactive protein which in turn was a better test than WCC. You may also note from Figure 3 that the AUC for procalcitonin (0.96) was greater than the AUC for the other two tests (083 for C-reactive protein, and 0.51 for WCC).

WHAT CUT-OFF POINT SHOULD BE USED IN PRACTICE?

Ideally, one would want to have a test which is both highly sensitive and highly specific but this is not always possible. As earlier discussed, when the cut-off point between normal and abnormal is changed to increase either sensitivity or specificity, there is usually a concomitant decrease in the other.

In general, when it is very important not to miss a diagnosis (for instance, when there is a disease with high mortality but for which a cure is available), a test with high sensitivity is needed. On the other hand, when the consequences of

having a false positive test is very serious (e.g. the psychological problems associated with falsely diagnosing a person to have HIV), a test with a high specificity is important.

The cut-off point between normal and abnormal of a test may therefore be varied to increase sensitivity or specificity (with concomitant decrease in the other) according to what we are using the test for.

CONFIDENCE INTERVALS

It should be noted that all measures of diagnostic accuracy including the AUC are statistical estimates and should be reported with confidence intervals (14). For example, in the study by Brown et al., the AUC of the ROC curve for peripheral WCC for diagnosing bacterial infections in febrile neonates was reported to be 0.723 with a 95% confidence interval (95% CI) of 0.566–0.869 (Fig. 2) (5). The 95% CI inform the reader about the interval in which 95% of all estimates of AUC will fall if the study was repeated over and over again. In other words, one can be 95% certain that the true value of the AUC of the ROC curve for peripheral WCC for diagnosing bacterial infections in febrile neonates lies between 0.566 and 0.869. Various methods for estimating the confidence intervals of the AUC have been described (16,17).

CONCLUSION

The sensitivity and specificity of clinical tests whose results are quantitative vary according to what cut-off point is chosen to define normal or abnormal. The ROC curve allows analyses of the trade-offs between sensitivity and specificity at all possible cut-off points. The curve may be used to select optimal cut-off values for a test result, to assess the diagnostic accuracy of a test, and to compare the usefulness of tests.

References

1. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr* 96: 338–41.
2. Akobeng AK. Understanding diagnostic tests 2: using likelihood ratios to estimate probability of disease. *Acta Paediatr*, in press.
3. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994; 309: 188.
4. Obuchowski NA, Lieber ML, Wians FH. ROC Curves in clinical chemistry: uses, misuses, and possible solutions. *Clin Chem* 2004; 50: 1118–25.
5. Brown I, Shaw T, Wittlake WA. Does leucocytosis identify bacterial infections in febrile neonates presenting to the emergency department? *Emerg Med J* 2005; 22: 256–9.
6. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med* 2000; 45: 23–41.
7. Last JM. *A dictionary of epidemiology*. New York: Oxford University Press, 2001.
8. Vining DJ, Gladish GW. Receiver operating characteristic curves: a basic understanding. *Radiographics* 1992; 12: 1147–54.
9. Lang TA, Secic M. *How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers*. Philadelphia: American College of Physicians, 1997.
10. Mayer D. *Essential evidence based medicine*. Cambridge: Cambridge University Press, 2004.
11. Crichton N. Information point: receiver operating characteristic (ROC) curves. *J Cin Nurs* 2002; 11: 136.
12. Perkins NJ, Schisterman EF. The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristics curve. *Am J Epidemiol* 2006; 163: 670–5.
13. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005; 47: 458–72.
14. Fischer JE, Bachman LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* 2003; 29: 1043–51.
15. Hatherill M, Tibby SM, Sykes K. Diagnostic markers of infection: comparison of procalcitonin with C reactive protein and leucocyte count. *Arch Dis Child* 1999; 81: 417–21.
16. Schisterman EF, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. *Am J Epidemiol* 2001; 154: 174–9.
17. Tosteson TD, Buonaccorsi JP, Demidenko E, Wells WA. Measurement error and confidence intervals for ROC curves. *Biom J* 2005; 47: 409–16.