Statistical note

Of bombers, radiologists, and cardiologists: time to ROC

Assessing the accuracy of any diagnostic procedure remains integral to method evaluation. Evaluation of a diagnostic procedure is assessed by its ability to categorise patients accurately into those with or without a disease state. The presence or absence of the disease state is defined according to some, often arbitrarily selected "gold standard". The nature of the gold standard can itself be a cause for debate. In the diagnosis of acute myocardial infarction (AMI), the gold standard has always been the criteria initially recommended by the WHO.[1 2] As newer tests for AMI have been developed they have been evaluated against the gold standard and progressively replaced the older gold standard tests. Thus, aspartate aminotransferase (AST) measurement has been replaced by creatine kinase (CK). CK has been superseded by measurement of the more cardiac specific MB isoenzyme (CK-MB). Even in this case, as newer methods are developed for measuring CK-MB, mass measurements have replaced activity measurements to produce as the new gold standard the triad of chest pain, ECG changes, and CK-MB mass measurement. This results in a creeping diagnostic classification, the "gold" becoming purer (or perhaps less tarnished).

Test accuracy is expressed in terms of sensitivity and specificity. The sensitivity of a test is defined as the ability to detect the diseased population. This equates to the number of real cases detected (referred to as true positives, TP) divided by the total number of cases in the population (the true positives plus those missed, the false negatives (FN)). Hence, sensitivity = TP/TP + FN. Specificity is the ability to exclude correctly the non-diseased population. This equates to the number of cases without disease (true negatives, TN) divided by the number of true negatives plus non-diseased patients giving a positive test result (false positives, FP). This yields the usual 2 × 2 contingency table (table 1), together with a set of permutations and combinations including positive and negative predictive values and likelihood ratios. Thus, Nirvana is achieved with 100% sensitivity, 100% specificity, and an infinite likelihood ratio but (like Nirvana) this is never achieved. Is this a reasonable method of assessing any one test, or of comparing tests?

When calculating sensitivity and specificity there will always be a trade off. As sensitivity increases, specificity will fall. Thus, a test may appear highly specific but closer examination will reveal that it does not detect the diseased population. A further confounding feature will be the prevalence of disease in the test set (defined as TP + FN/TN + FP). This will greatly affect the test performance. A test may appear highly specific in a low disease prevalence group but be clinically useless when translated to a more representative population. Hence, the level of cut off for the test must be critically selected, not only to balance sensitivity and specificity but also to take into account the underlying bias imposed by the population studied. Direct comparison of tests by comparing sensitivity and specificity is also fraught with hazard. If one test has a claimed sensitivity of 90% and another 65%, clearly bigger is better. Unfortunately, it will be influenced by the underlying sample size. This can be overcome by calculating confidence intervals (CI). Where confidence intervals overlap, tests may be equivalent despite apparent differences. Hence 90% when n = 20 (CI 68.3 to 98.9) is not better than 60% (CI 36.1 to 80.9)

This situation is not unique to medicine but is part of a larger problem of distinguishing signal from noise. This

Table 1 Standard 2 × 2 table used for calculating sensitivity and specificity

Gold standard test result	*New test result positive*	*New test result negative*
Diseased	True positive (TP)	False negative (FN)
Non-diseased	False positive (FP)	True negative (TN)

Sensitivity = TP/TP + FN.
Specificity = TN/TN + FP.
Positive predictive value = TP/TP + FP.
Negative predictive value = TN/TN + FN.
Likelihood ratio = sensitivity/1 − specificity = TP/(TP + FN)/1 − (TP/(TP + FP).

Table 2 Example data for ROC curve: CK-MB mass eight hours from admission
A Partial sample data set

MB mass (μg/l) 8 h from admission	*Myocardial infarction*
2.0	0
2.0	0
3.4	0
3.5	0
3.5	0
3.6	0
14.5	0
14.7	0
15.8	0
15.8	0
16.1	0
17.1	0
2.3	1
5.35	1
5.5	1
12.4	1
12.6	1
15.5	1
15.9	1
16.4	1
17.1	1
30.1	1
30.6	1
30.7	1
32.0	1
32.4	1
33.0	1
239.0	1
337.0	1

0 = no AMI; 1 = AMI.

B Partial dataset from ROC curve calculation

Cut off value	*Sensitivity*	*Specificity*	*TP*	*FN*	*FP*	*TN*
0.3	1.00	0.01	112	0	151	1
1	1.00	0.04	112	0	146	6
2	1.00	0.23	112	0	117	35
4	0.99	0.54	111	1	70	82
5.5	0.97	0.68	109	3	49	103
8	0.97	0.78	109	3	34	118
12	0.97	0.84	109	3	24	128
21	0.88	0.91	99	13	13	139
33	0.75	0.96	84	28	6	146
42	0.64	0.97	72	40	4	148
60	0.53	0.99	59	53	2	150
82	0.39	1.00	44	68	0	152
124	0.26	1.00	29	83	0	152
220	0.10	1.00	11	101	0	152

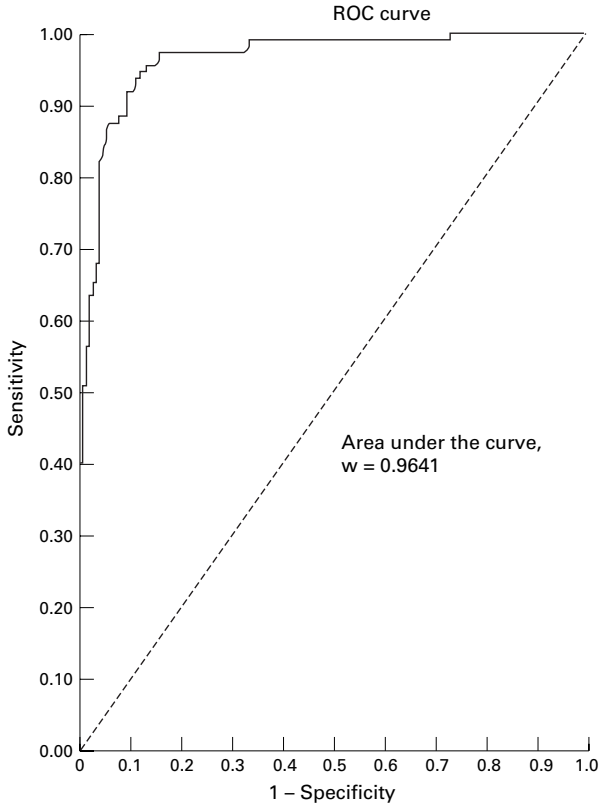

Figure 1 ROC curve from CK-MB data.

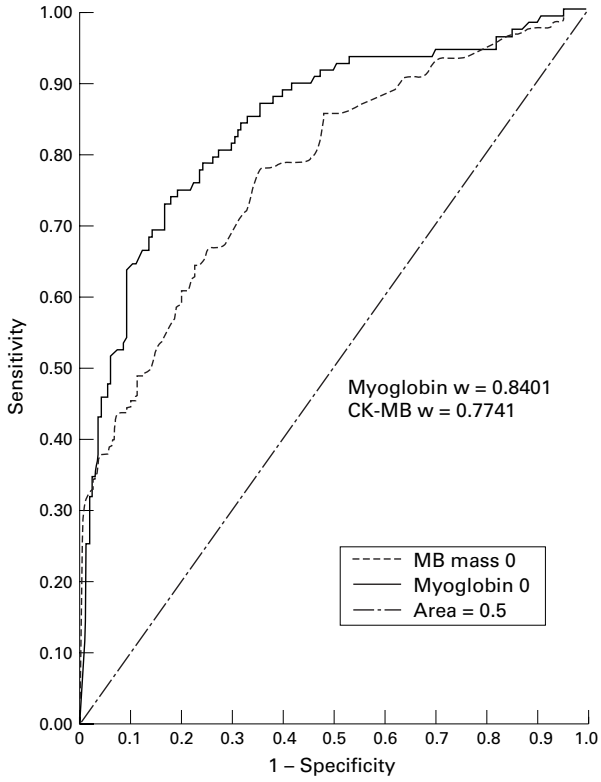

Figure 2 Comparative ROC curves for CK-MB and myoglobin for diagnosis of AMI on admission.

can be addressed by the technique of "receiver operating characteristic" or "relative operating characteristic" curves or ROC curves. This technique was developed initially to enhance early radar signals to detect bombers.[3] This technique has wide application and has been used for applications as varied as testing materials for flaws to checking for income tax evasion and in medicine for experimental psychology and psychophysics.[4] It was first extensively used in radiology to evaluate medical imaging devices,[5 6] especially computed topography phantoms, but is now considered the standard technique for test evaluation in clinical biochemistry.[7 8]

An ROC curve is conceptually very simple. It is a plot of sensitivity against specificity. The test set under consideration is initially classified into those with or without the feature under investigation. An example is the presence or absence of a disease, such as AMI *v* no AMI. A table is prepared of test result matched against diagnosis Table 2A illustrates this for CK-MB *v* diagnosis of AMI. The sensitivity and specificity is then calculated as the cut off level for the test is incrementally increased to produce a tabulated set of results of sensitivity and specificity corresponding to each test cut off level (table 2B). Sensitivity is then plotted against specificity, or more often 1 − specificity to produce the ROC curve (fig 1). This can be conveniently performed using a spreadsheet such as Excel, or more conveniently using an Excel add-in such as Analyse-It (Analyse-It, Leeds, UK).

Inspection of the curve, where the point of maximum curvature occurs, corresponds to the optimal trade off between sensitivity and specificity, and this is the optimal cut off value for the test. The ROC curve allows direct evaluation of the test power. The nearer the ROC curve is to a rectangle (the nearer to the top left hand corner of the graph) the better the test (a 45° line is a useless test). More quantitatively, the area under the curve can be calculated, allowing a direct assessment of the test ability. Hence, for the data illustrated, the area under the ROC curve is 0.9641. The greater the area under the curve, the better the test. Tests with an area under the curve of 0.5–0.7 have low accuracy, 0.7–0.9 moderate accuracy, and > 0.9 high accuracy.[9] An ideal test has an area under the ROC curve of 1. As expected, CK-MB measurement at eight hours is very accurate for diagnosis of AMI. The confidence intervals for the area under the curve can be calculated.[7 8]

The evaluation is not confined to disease *v* non-disease but can be produced for any feature that can be divided into a binary classification of the characteristic under investigation. This will include other characteristics such as survival *v* non-survival or, as we have used, ejection fraction above or below 40%. Although numeric data is often used, ROC curves can be constructed using interval (grouped data).[10] If two or more tests are to be compared, there are statistical techniques to compare the areas under the curve for significant differences (fig 2).[11] The likelihood ratio corresponds to the tangent to the curve at any point.[12]

ROC analysis is a robust and powerful methodology, which largely avoids the pitfalls of sensitivity and specificity described above. It cannot abolish the prevalence problem but serves to minimise it. It allows direct comparison of methods that use the same end point (whatever that may be providing it supports a binary categorisation of the test dataset). It should be used more widely.

P COLLINSON

Department of Chemical Pathology,
Mayday University Hospital,
London Road, Thornton Heath,
Surrey CR7 7YE, UK
email: poctrop@poctrop.demon.co.uk

1 Working Group on the Establishment of Ischaemic Heart Disease Registers. *Report of the fifth working group.* WHO, Eur 8201 (5), Copenhagen, 1971.
2 Nomenclature and criteria for diagnosis of ischaemic heart diseases. Report of the joint International Society and Federation of Cardiology/World Health Organisation task force on standardisation of clinical nomenclature. *Circulation* 1979;**59**:607–9.
3 Lusted LB. Signal delectability and medical decision making. *Science* 1971; **171**:1217–19.

4 Green DM, Swets JA. *Signal detection theory and psychophysics*. New York: John Wiley and Sons Inc, 1966.
5 Lusted LB. Decision making studies in patient management. *N Engl J Med* 1971;**284**:416–24.
6 Metz CE. ROC methodology in radiological imaging. *Invest Radiol* 1986;**21**: 720–33.
7 Zweig MH, Campbell G. Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;**39**:561–77.
8 Henderson AR. Assessing test accuracy and its clinical consequences. *Ann Clin Biochem* 1993;**30**:521–39.
9 Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;**240**: 1285–43.
10 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
11 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;**148**:839–43.
12 Pierce JC, Cornell RG. Integrating spectrum specific likelihood ratios with the analysis of ROC curves. *Med Decis Making* 1993;**13**:141–51.

STAMPS IN CARDIOLOGY

Congresses

The Asian-Pacific Society of Cardiology was formed in the Philippines in 1956 and the 8th Asian-Pacific Congress of Cardiology was held in Taipei in 1983. To celebrate this occasion the Directorate General of Posts, Republic of China issued two stamps that were released on 27 November, the opening day of the congress. The NTs 18 stamp was designed by Mr Lee Kuang-chi and printed by the China Color Printing Co, Inc, Republic of China. The stamps were printed in sheets of 100 and 1.2 million of this value were issued.

The 4.50 drachma stamp from Greece issued in 1968 commemorates the 5th European Congress of Cardiology held in Athens. It features the "Hand of Aesculapius" from a fragment of bas relief from Asclepios' Temple, Athens.

The stamp from the Philippines was issued to mark the 11th World Congress of Cardiology held in Manila in 1990.

M K DAVIES
A HOLLMAN