LETTERS TO THE EDITOR

ROC curves and confidence intervals: getting them right

EDITOR,—It is encouraging to see an increasing recognition of the desirability of quoting confidence intervals (CI) in association with indices of the performance of diagnostic tests. It is unfortunate, therefore, when errors are made in the calculation and interpretation of these, as happened in two recent papers in *Heart*.

A few errors spoil Collinson's[1] otherwise useful introduction to ROC curves. First he has defined the prevalence of disease wrongly. Using the notation in his table 1, the prevalence is not (TP+FN)/(TN+FP) but (TP+FN)/(TP+FN+TN+FP), where the denominator is in fact just the total study size. Brackets are essential here to clarify numerator and denominator—these are omitted from several expressions in Collinson's paper. His expression for the likelihood ratio is also incorrect: it should be [TP/(TP+FN)]/[1−TN/(TN+FN)].

Second, Collinson[1] notes that variation in the prevalence of disease "will greatly affect the test performance". While the sensitivity and specificity *may* vary according to setting (and hence disease prevalence)[2] more often the opposite is true. Indeed it is often noted as a characteristic of these measures that they are not affected by disease prevalence.[3] For example, the sensitivity of a test is unaffected by how many disease negative patients are included in the study.

Third, he makes a common error in relation to CIs. He suggests that one should compare two sensitivities by seeing whether their CIs overlap. In fact the difference between two sensitivities may be statistically significant even when the CI overlap. The correct procedure when comparing two groups is always to calculate the CI for the contrast of interest,[4] here the difference in sensitivities.

Collinson suggests that ROC curves avoid the "pitfalls of sensitivity and specificity". While ROC curves are certainly more informative, it is only a plot of sensitivity *v* 1 − specificity for different cut points and thus cannot avoid sharing the characteristics of those measures. Indeed critics observe that the use of ROC curves does not take disease prevalence into account.

In the same issue, Rao *et al* examined the relation between troponin T and left ventricular ejection < 40% in which they presented measures of test performance, including the area under the ROC curve, with CIs.[5] Here, too, there are some statistical problems.

While some statistical methods are equally valid in small samples (such as the *t* test), others are valid only when the sample or samples are quite large. These so called "asymptotic" methods include the χ^2 test for comparing two proportions, and the conventional associated method for constructing CI for a single proportion or the difference between two proportions. Large sample methods rely on the fact that the sampling distribution of a statistic is Normal.[6] When a proportion is near to 0 or 1, or when the sample size is small, and especially when both of these occur, the assumptions of these methods are not met and they give unreasonable results. For example, the CI for sensitivity of a diagnostic test may exceed 100%.[7] In such circumstances alternative methods should be used for the construction of CIs.[8–10]

One way of recognising that use of a particular method is not appropriate is when it gives impossible answers. For example, the CI for a proportion should lie wholly within the range 0 to 1 (or 0% to 100%). Confidence intervals should never be quoted that include impossible values. This applies to the sensitivity and specificity of diagnostic tests, and to the area under the ROC curve.

Rao *et al* quote a sensitivity of 100% for a cut off of 2.8 μg/l with a CI of 84.6% to 100%.[5] The observed proportion seems to be 24/24. They do not say which method they used to obtain the CI, but it would seem to be an exact method.

In contrast, they quote the area under the ROC curve of 0.9773 with a 95% CI from 0.9409 to 1.0136. (Incidentally, four decimal places is certainly excessive for such measures; two or three at most would be reasonable.) Again they do not say which method they used to obtain the CI, but the upper limit exceeds 1 and thus the interval includes impossible (and hence absurd) values. It seems that they used a method that relied on a Normal sampling distribution. Obuchowski and Lieber[11] note that for methods of high accuracy (ROC area > 0.95) use of the asymptotic method for the area under a single ROC curve may require a sample size of 200. For smaller samples, such as the 50 of Rao *et al*,[5] a bootstrap approach is recommended.

In addition, it seems to me that either the quoted value of the area under the ROC curve of 0.9773 or the graph is incorrect. Inspection of the authors' fig 2 indicates that the correct value for the plotted ROC curve is < 0.95. In addition, there is no point on the plotted curve corresponding to the quoted sensitivity of 100% and specificity of 92% for the troponin T cut point of 2.8 μg/l.

Finally, in neither paper is there recognition that the use of a data derived "best" cut point leads to overoptimistic assessment of test performance characteristics, especially in small samples. It is very likely that using the cut point of 2.8 μg/l in a further group of patients would give worse sensitivity and specificity than the values quoted.

DOUGLAS G ALTMAN
ICRF Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Old Road, Headington, Oxford OX3 7LF, UK

1 Collinson P. Of bombers, radiologists, and cardiologists: time to ROC. *Heart* 1998;**80**:215–17.
2 Lachs MS, Nachamkin I, Edelstein PH, *et al*. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;**117**:135–40.
3 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental tool in clinical medicine. *Clin Chem* 1993;**39**:561–77.
4 Altman DG, Gore SM, Gardner MJ, *et al*. Statistical guidelines for contributors to medical journals. In: Gardner MJ, Altman DG, eds. *Statistics with confidence*. London: BMJ Publishing Group, 1989:83–100.
5 Rao ACR, Collinson PO, Canepa-Anson R, *et al*. Troponin T measurement after myocardial infarction can identify left ventricular ejection of less than 40%. *Heart* 1998;**80**:223–5.
6 Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991: 153–9.
7 Deeks JJ, Altman DG. Sensitivity and specificity and their confidence intervals cannot exceed 100%. *BMJ* 1999;**318**:193–4.
8 Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;**17**:857–72.
9 Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med* 1998;**17**:873–90.
10 Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med* 1998;**17**: 2635–50.
11 Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Acad Radiol* 1998;**5**:561–71.

This letter was shown to the author, who replies as follows:

Dr Altman makes his usual excellent and insightful comments on statistical matters. First, my thanks for his correction of the error that crept in on prevalence and likelihood ratios.

In regard to his second point the clinical (as opposed to statistical) performance of a test is critically dependent on disease prevalence. In situation of low prevalence, such as patients with chest pain in the emergency department, the use of creatine kinase (specificity 90%) as cardiac enzyme rather than cardiac troponin T (specificity 100%) results in inappropriate admissions. The corollary is that when a test is examined for apparent sensitivity and specificity, the study of an inappropriate population biases test evaluation. Examination of cardiac markers in patients with Q wave acute myocardial infarction patients in the coronary care unit shows almost all markers perform with high sensitivity and specificity.

The comparison of CIs provides a rapid means of assessing if tests are different. I certainly agree that if CIs overlap, a more rigorous approach is required. I do not like to compare test sensitivity alone as it represents one single point on the ROC curve and can be subject to selection bias. I prefer to compare areas under the ROC curve. If CIs overlap, I compare the area under the curve statistically.

Finally, the article on ROC curves serves as an introduction rather than a comprehensive account. Discussion of the alternative methods of comparison of areas under the ROC curve is covered in the referenced articles. The question of the merits of confidence intervals for ROC curves would require an article in its own right; perhaps Dr Altman would oblige?

In respect of his comments on the paper by Rao *et al*, I will respond to his points. The CIs for sensitivity and specificity were calculated by a program from a reputable source—Dr Altman. The data for the ROC curve are the product of the statistical program and are reproduced as they were generated. Beyond this I cannot comment. Finally, in a large prospective study of ejection fraction we have been able to confirm the findings (unpublished data, 1999)

CORRECTION

Of bombers, radiologists, and cardiologists: time to ROC. P Collinson. *Heart* 1998;**80**:215–17.

The calculation of prevalence and likelihood ratio should been (TP, true positive; TN, true negative; FP, false positive; FN, false negative):

Prevalence: (TP+FN)/(TP+FN+TN+FP)

Likelihood ratio: [TP/(TP+FN)]/[1 – TN/(TN+FN)]