Archives of Disease in Childhood 1995; **72**: 358–361

STATISTICS FROM THE INSIDE

14. Diagnostic and screening tests and reference values

M J R Healy

Many studies in medical research are concerned with the properties of diagnostic and screening tests. These present a number of statistical problems, both in design and interpretation. I start by considering the simpler kind of test whose result can be described as being either positive or negative.

Tests with a positive/negative result

A clinician suspects that a patient has a certain disease condition and calls for a particular test to be carried out. The result of the test is positive. How much weight should be placed upon this result? To answer this question, suppose that a study has been done in which the presence or absence of the disease in a number of patients has been established for certain by some means or other, perhaps by an expensive gold standard test, and that the results for the new test were as in the table.

Results of a yes/no test

	Disease		
	Absent	*Present*	
Test negative	82	17	99
Test positive	12	73	85
	94	90	184

The numbers in this table can be summarised in various ways. To begin with, the proportion of patients with the disease who had a positive test was 73/90=0·81. This is called the *sensitivity* of the test. A high sensitivity means that only a few of the patients who actually have the disease are missed by the test. Similarly, the proportion of patients not having the disease who had a negative test was 82/94=0·87. This is the *specificity* of the test.* A test with high specificity gives only a small proportion of misleading positive results among patients who do not have the disease.

Before going further, a few remarks are in order. First of all, the fractions above are actually only estimates of the true properties of the test. The sensitivity in the example of the table has an approximate 95% confidence interval of 0·73 to 0·90, a fairly wide range, and the sensitivity is similarly imprecise. When such estimates are presented in a paper, it is important that confidence intervals should be given. We are also assuming that each column of the table constitutes a random sample of patients without and with the disease respectively, or more precisely of those patients who are liable to get the test. It is legitimate in this context for these two samples to be drawn separately if this is convenient.

It may be useful to point out, following the previous article in this series on 'probabilities and decisions',[1] that the true sensitivity and specificity are actually *conditional probabilities* – using the notation introduced there, we have

$$\begin{aligned}&\text{sensitivity}\\&\quad=\text{pr(test positive} \mid \text{disease present)}\\&\text{specificity}\\&\quad=\text{pr(test negative} \mid \text{disease absent)}\end{aligned}$$

where the vertical bars can be read as 'conditional upon' or 'assuming that'.

The sensitivity and specificity are inherent properties of the test. Provided the conduct of the test and the definition of the disease condition remain the same, they should be transferable from one situation to another. However, it will be seen that they are not the quantities that the practising clinician is commonly interested in. She is more likely to require the conditional probabilities the other way round, the probabilities that the disease is present (or absent) when the test comes out positive (or negative). These probabilities are called the positive and negative *predictive values*. Formally, written as conditional probabilities,

$$\begin{aligned}&\text{positive predictive value}\\&\quad=\text{pr(disease present} \mid \text{test positive)}\\&\text{negative predictive value}\\&\quad=\text{pr (disease absent} \mid \text{test negative)}\end{aligned}$$

To reverse the order of a conditional probability we need to invoke Bayes' theorem. For brevity let me write D+ and D− to denote disease present and absent and T+, T− for test positive and negative. Then the probability for *both* D+ *and* T+ can be written in two ways:

$$\begin{aligned}&\text{pr(T+ and D+)}\\&\quad=\text{pr(T+)}\times\text{pr(D+}\mid\text{T+)}\\&\quad=\text{pr(D+)}\times\text{pr(T+}\mid\text{D+)}\end{aligned}$$

*Thought will show that these two terms are reasonably appropriate to their meanings. Their close similarity is not considered a drawback by experts in the field, however tiresome it may be for others.

23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to:
Professor Healy.
No reprints available.

From this we find that

$$\mathrm{pr(D+\,|\,T+)=pr(T+\,|\,D+)\times\left\{\frac{pr(D+)}{pr(T+)}\right\}}$$

so that to reverse the order of the conditional probability we need to multiply by the term in the curly brackets. We can go a little farther by noting that either D+ or D− must be the case so that pr(T+), the overall probability of a positive test, is the sum of two terms covering the two situations:

$$\mathrm{pr(T+)=}\\ \mathrm{pr(T+\,|\,D+)pr(D+)+pr(T+\,|\,D-)pr(D-)}$$

Furthermore, pr(D+) has a meaningful interpretation – it is the probability of the disease being present when we do not know the test result. In a rather broad sense, this is what epidemiologists call the *prevalence* of the disease condition. Thus we find that

$$\mathrm{pr(D+\,|\,T+)}\\ \mathrm{=pr(T+\,|\,D+)}\\ \times\left\{\frac{\mathrm{pr(D+)}}{\mathrm{pr(T+\,|\,D+).pr(D+)+pr(T+\,|D-).(1-pr(D+))}}\right\}$$

All the quantities in this formula have been described in words above. In particular the multiplying factor can be written as

$$\mathrm{Prevalence/\{sensitivity\times prevalence}\\ \mathrm{+(1-specificity)\times(1-prevalence)\}}$$

The important feature is the intervention of both the specificity and the prevalence. The former is a property of the test but the magnitude of the latter will vary widely according to circumstances. In a specialist clinic it may be quite high, with most of the patients subjected to the test having the disease. In a screening environment it may be extremely low, one in 1000 or fewer. It is this latter circumstance that requires particular attention. Suppose that a superb test has both sensitivity and specificity equal to 0·99 and that this test is applied as a screen for a condition with a prevalence of one in 1000. Then the positive predictive value is

$$0{\cdot}99\times\frac{0{\cdot}001}{(0{\cdot}99\times0{\cdot}001+0{\cdot}01\times0{\cdot}999)}\\ =0{\cdot}99\times0{\cdot}0911=0{\cdot}0902$$

Given a positive test, the odds are 10 to 1 *against* the subject having the disease. In 10 000 subjects, approximately 10 will have the disease and these will all be expected to give a positive test result. But the remaining 9990 subjects will not have the disease, and these will give rise to around 100 false positive results. Of the 110 positive test results, 10 out of every 11 will on average be false positives.

The prevalence of a disease condition can be estimated from data such as those in the table if it can be assumed that the whole of the table (not just the separate columns) is a random sample of the relevant population of patients. This will be a useful method when the prevalence is fairly high but in a screening situation with a very low prevalence separate epidemiological studies will be needed. It must be remembered in either case that the measured prevalence will be no more than an estimate of the true value. When the data in the table are a proper random sample, the predictive values can be estimated directly from them in the same way as the specificity and sensitivity. Thus with this assumption, the predictive values from the table are 73/85=0·86 (positive) and 82/99=0·83 (negative).

Yet another pair of indices easily derived from the data in the table are the positive and negative *likelihood ratios*. The positive likelihood ratio, for example, is defined as the ratio pr(T+ | D+)/pr(T+ | D−). An estimate from the table is (73/90)/(12/94)=6·35. It is equal to the ratio sensitivity/(1−specificity). The interest of this ratio is connected with the direct use of Bayes' theorem in a diagnostic context. Suppose the prior odds on a patient having the disease is P (this is equal to prevalence/(1−prevalence)). Then the posterior odds after observing a positive test result is found by multiplying P by the positive likelihood ratio. This approach may be especially interesting when two alternative tests are to be compared.

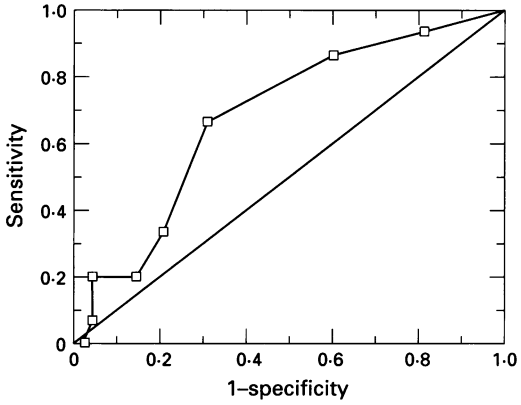
The various derived quantities that I have described come in pairs and it is tempting to try to find a single number to describe the performance of a test. One method that has been widely used involves what is known as *Youden's index*, which is simply equal to (sensitivity+specificity−1) and varies between 0 and 1 (unless the test is worse than useless in that it is negatively associated with the presence of disease). The use of any such single measure is not recommended in this context. In clinical practice (and indeed elsewhere) the consequences of false positive and false negative readings are commonly very different, and simply adding together the rates at which the two occur may have quite misleading implications.

I have used the phrase 'false positives' in the above discussion and the meaning is fairly clear – in my notation they are the (T+ | D−) cases where a patient without the disease gives rise to a positive test result. The table for example contains 12 of them. But the phrase 'false positive rate' should be avoided because of its ambiguity. In the table it might mean 12/94, which is 1 minus the sensitivity, or 12/85, which is 1 minus the positive predictive value.

Speaking from personal experience, it is not particularly easy for the non-specialist to keep in mind the interpretation of the various quantities that have been derived from the four entries in a table such as the table here. It would be helpful to readers of scientific papers if authors always quoted the table in full in addition to any indices that they may calculate from it. More extended discussion of the problems of test interpretation can be found in Galen and Gambino and Strike.[2 3]

Tests with a quantitative result

Many pathology tests nowadays do not produce a simple positive or negative result but rather a reading in numerical form. Doctors, in common with the rest of mankind, are not particularly comfortable with numerical results and adopt a number of practices to reduce



ROC of transport score as predictor of death in infants transferred to a neonatal unit.

them to a more manageable yes or no answer. The simplest tactic is to draw a line on the continuous scale and declare all results above the line to be positive, all below the line to be negative (this is especially common when the result is a titre). The problem of course is where to draw the line. Some light can be thrown on this by calculating the specificity and sensitivity for a range of different boundary values and plotting sensitivity against (1−specificity). The resulting curve, for rather obscure historical reasons, is called the *receiver operating characteristic* or ROC. An example is shown in the figure, where the data relate to infants transferred to a neonatal unit, the test is a transport score and the outcome is death (I am grateful to Dr T Stephenson and Mr A Leslie for letting me use the data as an illustration). The diagram illustrates how changing the boundary varies the pay off between sensitivity and specificity – we can increase either one of these but only at the expense of the other. A 'good' test will have an ROC which passes close to the upper left hand corner of the diagram which represents perfection in both respects. The diagonal line corresponds to the minimum performance as obtained by simple guessing. The ROC is a comprehensible summary of a good deal of data and may be a useful way of comparing informally the performance of alternative tests.

Reference values

Another strategy designed to render numerical results more comprehensible is the introduction of normal limits, or, as they are more commonly called today, reference values. The idea here is that the levels of some quantity, oral temperature, say, or a blood constituent, vary in the healthy population within a certain range of values. If a patient exhibits a value outside this range, this may be taken as evidence of abnormality or the presence of disease. Recognising that odd extreme values do turn up in perfectly healthy people, the range of values that do not arouse suspicion is commonly taken to be given by the mean plus and minus 2 standard deviations, these quantities relating to the frequency distribution of the values in the healthy population.

The derivation of reference values is a more complex business than appears at first sight. The most difficult question is a non-statistical one – how is the 'healthy population' to be defined? Many authors have suggested that results obtained from hospital patients could be used for defining reference values, but this is almost a contradiction in terms; people attending hospital cannot by definition be assumed to be healthy. The problem is more acute when the values depend upon factors such as age, ethnic origin, and gender. Separate ranges may have to be specified for males and females, for example. The construction of age specific reference ranges such as growth standards is a topic on its own I hope to return to in a future article in this series.

The purpose of reference values must also be borne in mind. The 'mean±2 SD' rule is clearly intended to give a sensitivity of 95% – only 5% of the healthy population may be expected to have values outside the reference range. This is closely analogous to a statistical significance test controlling the rate of type I errors. Like a significance test, it has little to say about specificity or type II errors; a value lying inside the reference range is far from being a guarantee of lack of abnormality. It should, however, be stated that reference values are technically to be distinguished from confidence limits, which apply to parameter values – in the jargon of statistics, reference values are better referred to as *tolerance limits*.

From a statistical point of view, the first thing to remember about reference values is that they are necessarily estimates based upon sample data. This has two immediate consequences. First, the quality of the sample is of prime importance. We may perhaps define the healthy population as consisting of anybody without overt disease. To assume that the collection of laboratory staff and medical students who are most immediately available for venepuncture constitutes a representative sample from such a population is rash in the extreme. When reporting reference values, a specification of the nature of the sample on which they are based is essential.

Secondly, since they are estimates, the reference values must be subject to sampling error in the sense that, were the study to be repeated in identical fashion, non-identical values would be found. The size of the sampling errors is much larger than is sometimes appreciated. Suppose that the distribution of values in the healthy population is Gaussian (the usual term 'Normal', even with a capital, is best avoided in this context, and the more historically correct 'de Moivrean' is unlikely to catch on) and that the upper reference value is calculated as mean+2 SD using estimates obtained from a sample of size n. Then the 95% confidence limits on the value obtained are $1{\cdot}71\ \sigma/\sqrt{n}$ where σ is the population standard deviation. For a sample size of 50 this amounts to $0{\cdot}24\ \sigma$, a far from negligible amount relative to the distance between the reference value and the mean which will be about 2σ. Looked at another way, while the proportion of the population to be included between the limits may be set at 5%, the actual proportion between the

estimated limits will also be subject to error. The standard error of this proportion is approximately $0{\cdot}15/\sqrt{n}$ and when n=50 this amounts to 0·0212, suggesting that the actual proportion included may be somewhere between 91% and 99%. When reference values are to be determined, large samples cannot be avoided.

The use of plus and minus two standard deviations to define a 95% reference range is based upon a tacit assumption that the underlying population is at least approximately Gaussian. The multiplier 2 is derived from this assumption (it is an approximation to the exact value of 1·96), but the mean and standard deviation will in practice both be estimates and it appears that the t distribution ought to be involved to allow for this, just as in an ordinary significance test. In fact, to ensure an *average* coverage of 95%, the correct multiplier for the standard deviation with a sample of size n is $t\sqrt{(n+1)/n}$ where t is the 2½% point of the t distribution with $(n-1)$ degrees of freedom, and this is very close to 2·0 for realistic sample sizes. However, the Gaussian assumption is very often false. What is to be done then? One possibility is to make no distributional assumption at all and simply to sort the observed sample values, setting the reference values so as to cut off just 2½% at each end. (This requires a little care. With 95% coverage the lower reference value corresponds to the 2½th centile. The estimate of this from a sample of size 100 (say) is equal to the 3rd smallest observation, not half way between this and the one below it.) However, this non-parametric technique has its disadvantages. In particular, the reference values estimated in this way are very much less precise when the distribution is actually Gaussian than those based upon the sample mean and standard deviation. To achieve the same degree of precision as is afforded by the parametric estimates, the sample size must be practically doubled.

As stated in a previous article in this series,[4] there are two sorts of departure from Gaussianity that commonly occur in practice. One of these is the presence of outlying values, all too often due to mistakes in recording or in laboratory technique. Detecting and eliminating these is particularly important when it is the tails of the distribution that are of interest, and also particularly difficult. One method that can be adopted when the uncontaminated distribution is approximately Gaussian is to estimate the mean and standard deviation from only the central part of the sample, trimming off and ignoring for purposes of calculation the outermost 5% or so of readings at each end where the outliers, if there are any, will be located. The mean of the trimmed sample can be used directly but the standard deviation estimate needs to allow for the trimming – I have described a possible method for doing this.[5] Given these estimates, sample values that are exceptionally extreme – perhaps those beyond 3 standard deviations from the mean, corresponding to a frequency of around 1 in 1000 – may be regarded as probably erroneous and omitted from the rest of the calculations.

The other common form of non-Gaussian distribution is one that is noticeably skew to the right. Here transforming the data values to logarithms very often succeeds in bringing them to Gaussian form. The use of a Normal plot is the most straightforward way of checking whether the transformation has been successful.[4] If some degree of skewness remains, or if the transformation has overcorrected, it may be helpful to subtract or add a constant quantity to each value before taking logarithms (the article by Flynn *et al* provides some examples[6]). Provided that the resulting distribution is reasonably symmetrical, more subtle departures from the Gaussian form are unlikely to cause trouble in this context.

The analogy between the use of a reference range and that of a significance test suggests that it would be useful to go beyond the simplistic split into high/normal/low by providing an estimate of the amount of the departure of a reading from the normal mean. A major advantage of an underlying distribution which is at least approximately Gaussian (possibly after transformation) is that it allows this to be done. Given the mean and standard deviation (or at least estimates of them) it is easy to express any result as some multiple of the standard deviation away from the mean. This is sometimes called an *SD score*. The usual limits of the 95% reference range correspond to SD scores of ±2, but lesser values should arouse suspicion, while scores of 3 and above are beyond plausible limits of normality.

1 Healy MJR. Statistics from the inside 13. Probability and decisions. *Arch Dis Child* 1994; **71:** 90–4.
2 Galen RS, Gambino SR. *Beyond normality: the predictive value and efficiency of medical diagnosis.* New York: Wiley, 1975.
3 Strike PW. *Statistical methods in laboratory medicine.* Oxford: Butterworth-Heinemann, 1991.
4 Healy MJR. Statistics from the inside 12. Non-normal data. *Arch Dis Child* 1994; **70:** 158–63.
5 Healy MJR. A linear estimator of standard deviation in symmetrically trimmed normal samples. *Applied Statistics* 1982; **31:** 174–5.
6 Flynn FV, Piper KAJ, Garcia-Webb P, McPherson K, Healy MJR. The frequency distributions of commonly determined blood constituents in healthy blood donors. *Clin Chim Acta* 1974; **52:** 163–71.