TESTING A TEST

Rashmi Kumar
P.K. Misra
S. Kumar

Diagnostic testing is an extremely important aspect of medical care and forms a large chunk of inpatients and outpatients expenditure. Newer diagnostic tests are continually coming into clinical use but till lately not much was said about assessment of the test itself, or analysis of its efficacy. A consequence of this is the wide variability in the results of similar studies and inability to draw valid conclusions.

This question is addressed in the emerging science of "Clinical Epidemiology". The principles governing the interpretation of diagnostic tests (which includes clinical, laboratory, pathological, radiological and nuclear medicine investigations) can be applied equally well to clinical data, symptoms and signs or even a set of symptoms and signs. Thus we could talk about the sensitivity and specificity of say, meningeal signs for making a diagnosis of bacterial meningitis or that of systolic or diastolic murmurs in the diagnosis of congenital heart disease and so on. Therefore this section is really about interpretation of diagnostic 'data' rather than tests alone. Although the examples cited may be for laboratory data, the wider applicability of the same principles to clinical data must be borne in mind. In fact the clinical data are usually far more powerful (and cheaper) tools for diagnosis than laboratory tests.

From the Departments of Pediatrics and Clinical Epidemiology Unit, King George's Medical College, Lucknow.

Reprint requests: Dr. Rashmi Kumar, Assistant Professor in Pediatrics, King George's Medical College, Lucknow.

Received for publication: March 12, 1992;
Accepted : March 22, 1992

Measures of Test Efficacy

Most diagnostic testing is to do with measurements. One of the attributes of a good test is its repeatability. Measurement of biological functions are very likely to vary in a day or hour manner (biological variation). This variation can be minimized by ensuring that measurements are taken at the same time of day and under similar conditions. Also, different people are likely to interpret the same event or measurement in different ways (interobserver variation) and the same observer will also exercise his judgement in a slightly different manner on different occasions (intraobserver variation). All these sources of error increase the 'variation' around any measurement and should be minimized. Another source of error is when, say, one observer always reads high or low. Such 'systematic' error should be carefully eliminated(1).

Most measurements in medicine are indirect indices of something we are really interested in. For example, doctors are not really interested in the height of a mercury

column except as it indicates blood pressure. In fact, to the practising doctor, the blood pressure is mainly of interest in so far as it indicates a risk of its complication (stroke, coronary artery disease). Validity of a test refers to how well these indirect indices relate to the attribute we are really interested in. Thus, validity of a test is an expression of the degree to which a measurement measures what it is supposed to measure.

Whenever we consider a diagnosis, we talk about probabilities. On the basis of our history and clinical examination we already have some estimate of the probability of the diagnosis (prior probability). Above a certain level of probability, we would treat the patient and below a certain level we would do nothing (*Fig. 1*). In between these two levels we would order investigations. The test/tests would be expected to move us along the scale of probability into either the 'treat' or 'do nothing' zone (post test probability).

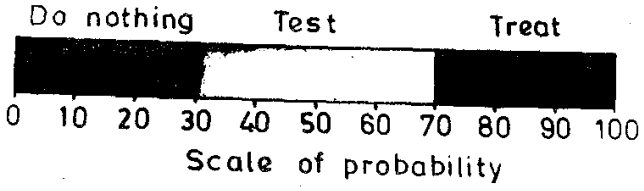

Fig. 1. Scale of probability of diagnosis showing the 'do nothing' and 'treat' zones in black and the intermediate 'test' zone in grey.

How much a test affects the probability of a diagnosis will depend on how good test is. Two of the attributes of a test are its sensitivity and specificity. Any test must be considered in these terms, which are easy enough to understand. Sensitivity is the ability of the test to pick out those patients who really have the disease. It is synonymous with PiD rate or positivity in disease rate or true positives. Specificity, on the other hand, is the ability of a test to say that a disease is absent when it is truly absent and is synonymous with NiH or negative in health rate or true negative(2).

Sensitivity and specificity for any test can be derived by comparing the test with the 'real answer' about whether the disease is present or absent in a simple 2 × 2 table. At the top of the table is 'truth'-the real answer as to whether the disease is present or absent. A test which is closest to the real answer is called the 'gold standard'. Along the side is the positive or negative result of test used. Thus, we have 4 groups of Patients (*Table I*):

(*a*) Patients with disease and a positive test.
(*b*) Patients without disease but a positive test.
(*c*) Patients with disease but a negative test.
(*d*) Patients without disease and with a negative test.

Sensitivity, therefore, equals a/(a+c) × 100 and specificity equals d/(b+d) × 100.

TABLE I–*The 2 × 2 Table*

Result of test	Truth about the disease	
	Disease present	Disease absent
Positive	a	b
Negative	c	d
	a+b	b+d

To illustrate this, let us take the example of acid fast bacilli (AFB) in gastric aspirate as a test of pulmonary tuberculosis. True presence or absence of disease is established by chest radiographs which serve as the 'gold standard' in this instance. Let us imagine, we screened 200 children of which only 100 actually had the disease. The test was positive in 60 patients of which 58 actually had the disease and it was

negative in 140 children of which 98 were actually disease free. The 2 × 2 table then looks something like *Table II* .

This test, because of its high specificity

TABLE II—*The 2 × 2 Table for Acid fast Bacilli Example*

AFB in gastric aspirate	Pulmonary tuberculosis	
	Yes	No
Positive	58	2
Negative	42	98
	100	100

Thus the sensitivity of the test is

$$58/[58+42] \times 100 = 58\%$$

and specificity $98/[98+2] \times 100 = 98\%$.

is more useful to 'rule in' pulmonary tuberculosis than to 'rule out' this diagnosis. That is to say, if it is positive the diagnosis is almost certain but if it is negative, the diagnosis of a pulmonary tuberculosis is by no means ruled out.

In the clinical situation we need to know the significance of a positive or negative test result. For this we require a knowledge of its predictive values. Positive predictive value (PPV) is the proportion of patients with a positive test who actually have the target disorder, *i.e.*, PPV = true positive/total positive × 100. For the example cited above, the PPV becomes a/[a+b] × 100 or 58/[58+2] 100 = 96.6%. In other words, in about 97% of occasions a positive test signifies the presence of disease. Similarly, a negative predictive value (NPV) means the proportion of patients with a negative test who actually did not have the disease in question, *i.e.*, NPV = true negative/total negative 100 and in this example it is d/[c+d] or 98/[98+42] × 100 = 70%, *i.e.*, in 70% of occasions a negative test signifies the absence of disease.

It is important to understand here that predictive values are not constant but must change with the prevalence of the disease in the setting in which the test is used. In the example above, out of 200 patients (a+b+c) tested, 100 (a+c) had the disease and therefore the prevalence was 50%. Now let us suppose that we used this test to screen for pulmonary tuberculosis in the population (rather than in patients with symptoms) and we know that prevalence of pulmonary tuberculosis in the population is say, 1% or 100/10,000. The sensitivity and specificity remain the same. The comparison of test with reality in the free living population is shown in *Table III*.

TABLE III—*The 2 × 2 Table —Relation to Prevalence*

Result of test	Pulmonary tuberculosis	
	Present	Absent
Positive	58	198
Negative	42	9702
	100	9900

Here, the PPV changes to 58/[58+198] × 100 *i.e.*, 22.6% and the negative predictive value changes to 9702/[9702+42] × 100 = 99.5%.

Thus, we see that predictive values are not constant but must change with the prevalence of the disease in the target population. As the prevalence falls, PPV also falls and NPV rises. Prevalence of the disease in the tested population is the same as pretest probability or prior probability that have already been spoken about. In some cases this can be obtained from actual prevalence data. In the absence of such data one may make an 'educated guess' based on clinical experience. It may also be increased (or decreased) by

suitably choosing the patients to be tested, *e.g.*, pretest probability of newborns screened for congenital heart lesions by ECHO cardiogram could be raised if only babies with a murmur were subjected to the ECHO. Similarly, we can increase the predictive value of amniotic fluid acetylcholinesterase levels for the prenatal diagnosis of neural tube defects by doing it only in mothers with raised ∝-fetoprotein test result rather than in all pregnancies.

Post test probability or posterior probability of a disease when the test is positive is synonymous with the positive predictive value (PPV) just discussed and similarly the post test probability of a negative test is equivalent to the negative predictive value.

Thus, having some idea about the prevalence of the disease in a particular situation and knowing the sensitivity and specificity of the test in question one can reconstruct the 2 × 2 table and find the post test probability or negative test result. Of course, the pretext probability figure could be a stumbling block. We know that clinicians disagree with one another considerably when asked to estimate the probability of a disease from a patient's signs or symptoms(3). Moreover, one may not be quite sure of one's own estimate but may at best be able to put it into a range. This problem becomes easier to handle as one gains experience with a disease and the pretest probabilities become more accurate. Secondly, tables and logarithms on pretest probabilities of common disorder are becoming increasingly available. Lastly, even if one can only specify a range, one stands to benefit because we could calculate predictive values (or post-test probabilities) for the centre point of this range and both the extremes to see the effect of pretest on post-test probability. Surprisingly, often one would find this information also quite useful.

The next stumbling block is that one must decide at what level of post-test probability one would discard or accept a diagnosis. Ordinarily, one might discard at 40% and accept at 60%. However, this decision must also depend on how harmful a misdiagnosis either as diseased or non-diseased would be in that situation. It would also depend on how the diagnosis in question compared with the post test probabilities of other diagnoses on our 'short list'.

Another index of how good a test (or sign or symptom) is the likelihood ratio. It expresses the odds that a positive or negative test would be expected in a patient with (as opposed to one without) the target disorder, *e.g.*, if the likelihood ratio for a test is 7.6, then it means that a positive test is 7.6 times as likely to come from a patient with the disease as from one without the disease. In other words it compares the chance of positive (or negative) test result in those with disease and in those without disease. From the same table one can deduce that

$$\text{likelihood ratio of a positive test} = \frac{\text{proportion of patients with disease having a positive test}}{\text{proportion of patients without disease having a positive test}}$$

$$= a/a + c \div b/b + d$$

$$= \text{sensitivity}/1-\text{specificity}.$$

Similarly,

$$\text{likelihood ratio of a negative test} = c/c \div d/b + d$$

$$= 1\text{-sensitivity/specificity}.$$

Applying this to the above example of pulmonary tuberculosis, likelihood ratio of positive gastric aspirate test $= 0.58/1-0.98 = 28$

which means that a positive test is 28 times as likely to come from patients with the disease than from patients without disease.

The likelihood ratio for a negative test = 1–0.58/0.98 = 0.43. Thus, the chances of a negative test in patients with disease are less than ½ those in patients without disease.

Likelihood ratios (LR) are a useful index of diagnostic data. Because the proportions that make up the likelihood ratios are calculated vertically, they do not change with changes in prevalence of the target disorder. There is also the option of calculating these for several levels of the test results. We can thus make the most of the entire range of our diagnostic test results and by keeping track of the likelihood of the patient having the disorder at each point, we can carry patients to extremely high or low likelihoods. This reduces the number of false positives and false negatives. Finally, likelihood ratios can be used as a powerful way to shorten the diagnostic process as pre test odds × LR = post test odds.

To use the above formula, one needs to shift back and forth between pretext and post test odds and probabilities.This problem can be obviated by using nomograms based on pretest probability, likelihood ratio and post-test probability(2). It is also simple enough to calculate these values.

Pretest odds = Pretest probability/ 1-pretest probability.

Post test probability = Post test odds/ Post test odds + 1.

Thus, we can directly derive the post test probability if the pretest odds and likelihood ratios are known. The LR strategy is also very useful when we are planning a sequence of tests, as the post test odds of the 1st test become the pretest odds for the 2nd test.

As more and more clinicians recognize the power of this index, more and more demand is being created for information about likelihood ratios for various levels of different laboratory tests. It is likely that these will soon replace sensitivity, specificity and predictive values.

Receiver Operator Characteristic Curves

Diagnostic test results are often not expressed simply as positive or negative but as actual values. The probability of the patient having the disease would depend on 'how' positive the test is and in fact, sensitivity and specificity figures can be obtained for different values of test results. It will usually be found that as sensitivity increases, specificity decreases, *i.e.,* there is a trade off between the two. The receiver operator characteristic (ROC) curve is a way of displaying the test characteristic at different test values and is charted between sensitivity and 1-specificity (also the false positive rate). It can also be used for comparing 2 or more different tests for the same target disorder and for choosing the cut off point at which the test becomes most useful. The upper left hand corner denotes a perfect diagnostic test. It follows that the point on the ROC curve which is closest to the left upper corner is usually the best cut off. *Figure 2* shows the ROC curve of different respiratory rates as indicators of lower respiratory infections in infants. The point closest to the top left corner, *i.e.,* a respiratory rate of above 50/min was the best cut off value to serve as an indicator of lower respiratory infection in infants.

Having gone through the above discussion most readers would have realized that all this talk about odds and probabilities is just putting figures to something that they are doing intuitively any way. We deal with uncertainty in almost everything we do–the patient's history, our physical examination, the laboratory test results, the

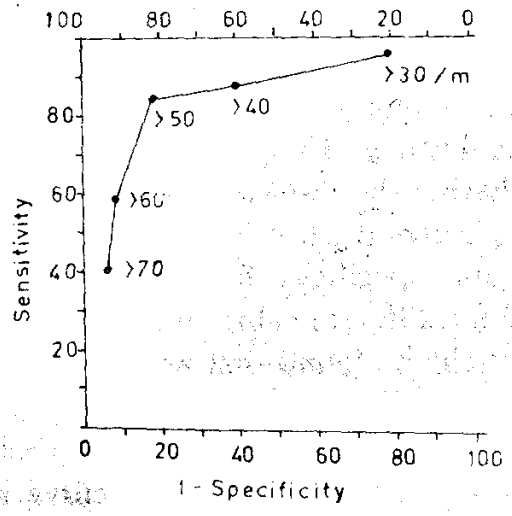

Fig. 2. ROC curve indicating the cut off point for respiratory rate/min which would serve as the best indicator of lower respiratory tract infection in infants.

diagnosis, prognosis, and response to treatment. Is there any need then to quantify our diagnostic uncertainty? The answer is yes. Firstly, this acknowledges that uncertainty does exist. Secondly, words like rarely, doubtful, often, occasionally and the like may mean a lot to the person who uses them but other physicians may have widely different estimates (in terms of percentages) of the meaning of these words. This was revealed in an interesting communication by Bryant and Norman(4) who collected 30 such terms from diagnostic reports and then asked different clinicians to estimate to the nearest 5% the probability of disease corresponding to each of these words. Rightly, these authors recommend that such words be abandoned from medical literature. Finally, as we go along in our clinical practice applying these principles both to clinical data and laboratory tests, we are bound to improve as clinicians as the significance of different signs, symptoms and tests emerges more clearly. This also form the base of a new 'science' to the art of medicine–'Clinicial Decision Analysis'.

REFERENCES

1. Christie D, Gordon I, Heller R. Epidemiology-An Introductory Text for Medical and Other Health Science Students. Newcastle, New South Wales University Press, 1990, pp 45-47.
2. Sackett DL, Haynes RB, Tugwell P. Clinical Epidemiology–A Basic Science for Clinical Medicine. Boston, Little Brown and Co, 1985, pp 59-138.
3. Freightner JW, Norman GR, Haynes RB. The reliability of likelihood estimates by physicians. Clin Res 1982, 30: 298
4. Bryant GD, Norman GR. Expression of probability: words and numbers. N Engl J Med (letter) 1980, 302 : 411.