Mathematics and Medicine

Test reduction: II—Bayes's theorem and the evaluation of tests

C C SPICER

The most useful statistical procedure for assessing diagnostic tests is based on a theorem named after the Rev Thomas Bayes, who discovered it in about 1760. Its use in the more general field of testing statistical hypotheses has caused a good deal of controversy, and it suffered for many years from the influential criticisms of Sir Ronald Fraser. In the applications to be discussed here the use of the theorem is not controversial and has much the same status as any other theorem, such as that of Pythagoras.

The ideas incorporated in it are those familiar to all clinicians making a diagnosis, where the likelihood of disease being present depends not only on the signs and symptoms but also on the frequency of the disease in the community. The latter probability is called the a priori or prior probability, and is familiar to every medical student through the maxim that "common things commonly occur."

Attributes and pattern frequencies

It is easiest to explain the use of Bayes's theorem by referring to a specific example, and the one chosen is based on data collected at Northwick Park Hospital on criteria for admission of patients with acute abdominal pain. For simplicity, three of the most useful signs and symptoms for making such a decision have been selected—namely, patient's assessment of severity of pain; patient's statement about whether the pain is getting better or worse; and presence or absence of guarding. It is convenient to use the general term "indicant" proposed by Card and Good[1] to describe all such signs, symptoms, and other tests. The three indicants each take two states, and therefore give rise to eight possible patterns, which are shown in table I together with their frequency in patients whose follow-up showed them to require or not to require admission. The data are open to a number of criticisms, as is their interpretation, but these are not relevant to the present discussion, which is concerned only with the method.

Table I shows, not unexpectedly, considerable differences in the frequencies of the patterns in the two groups, and the obvious selection rule would be to assign a patient to the group for which the corresponding pattern is most common. For example, severe pain, getting worse, with guarding present, occurs in 14% of those who require admission and only 1% of those who do not, and the decision is obviously to admit. But if the prior probabilities were grossly disparate this decision might become less clear-cut—for example, if the proportion needing admission in the population studied was only 1%, instead of about 60%. Bayes's theorem provides the method for combining the two probabilities to obtain an overall estimate. Table II gives the arithmetical details of its application.

Clinical Research Centre, Division of Medical Computing, Harrow, Middlesex HA1 3UJ

C C SPICER, MRCP (present address: Department of Mathematical Statistics, University of Exeter, Exeter EX4 4PU)

Firstly, the product of the pattern frequency and the prior probability in the two admission groups is calculated for each pattern. The prior probabilities in the present example are about 0·6 that a patient requires admission and 0·4 that he does not, so that for the first pattern the probabilities in the two groups are $0{\cdot}6 \times 0{\cdot}1376 = 0{\cdot}0826$ and $0{\cdot}4 \times 0{\cdot}0096 = 0{\cdot}0038$ respectively.

TABLE I—*Frequencies of indicant patterns in patients with acute abdominal pain requiring and not requiring admission to hospital*

Attribute patterns	Needing admission	Not needing admission
Severe; worse; guarding	0·1376	0·0096
Severe; worse; no guarding	0·0183	0·0192
Severe; better; guarding	0·1101	0·0096
Severe; better; no guarding	0·0734	0·0577
Moderate; worse; guarding	0·1834	0·0288
Moderate; worse; no guarding	0·0551	0·1154
Moderate; better; guarding	0·3578	0·0769
Moderate; better; no guarding	0·0642	0·6827
Sums	1·0	1·0

The sum of these two numbers, 0·0864, is the overall frequency of the pattern in the whole population observed. Bayes's theorem states that the probability of requiring admission, given the pattern, is:

$$0{\cdot}6 \times 0{\cdot}1376 \ / \ (0{\cdot}6 \times 0{\cdot}1376 + 0{\cdot}4 \times 0{\cdot}0096) = \frac{0{\cdot}0826}{0{\cdot}0864} = 0{\cdot}9560$$

and of not needing admission:

$$0{\cdot}4 \times 0{\cdot}0096 \ / \ (0{\cdot}6 \times 0{\cdot}1376 + 0{\cdot}4 \times 0{\cdot}0096) = \frac{0{\cdot}0038}{0{\cdot}0864} = 0{\cdot}0440$$

or, in general terms:

(probability of disease given the pattern) = (prior probability of disease)
× (probability of observed pattern in that disease)
÷ (probability of pattern in population).

This is written symbolically as:

$$P(D|S) = \frac{P(D) \times P(S|D)}{P(S)}$$

where S represents the pattern of symptoms and signs observed and the upright bar "|" is read as "given." The probability P(D|S) is called the "posterior" probability in contrast to the prior. Table II shows these probabilities for each pattern. Doubtful decisions, where the chances of being right are about 50:50 are shown in brackets; wrongly assigned patterns are asterisked in columns 1 and 2.

The usefulness of the set of attributes can now be assessed by adding up the frequencies in table II of the four possible categories: admitted correctly, admitted incorrectly, discharged correctly, and discharged incorrectly. For example, the

TABLE II—*Calculation of probabilities of occurrence of the two admission classes for each pattern of indicants. Cols (1) and (2) are derived from table I by multiplying by the prior probabilities of the two classes, 0·6 and 0·4 respectively. The decisions indicated by the probabilities are given in col (6), where doubtful decisions are in brackets. Wrong decisions are asterisked in cols (1) and (2)*

	(1) Admissions	(2) Discharges	(3) (1)+(2)	(4) (1)÷(3)	(5) (2)+(3)	(6) Decision
Severe; worse; guarding..	0·0826	0·0038*	0·0864	0·9560	0·0440	Admit
Severe; worse; no guarding	0·0110	0·0077*	0·0187	0·5882	0·4118	(Admit)
Severe; better; guarding	0·0661	0·0038*	0·0699	0·9456	0·0544	Admit
Severe; better; no guarding	0·0440	0·0231*	0·0671	0·6557	0·3443	(Admit)
Moderate; worse; guarding	0·1100	0·0115*	0·1215	0·9053	0·0947	Admit
Moderate; worse; no guarding ..	0·0331*	0·0462	0·0793	0·4174	0·5826	Do not admit
Moderate; better; guarding	0.2147	0·0308*	0·2455	0·8745	0·1255	Admit
Moderate; better; no guarding..	0·0385*	0·2731	0·3116	0·1235	0·8765	Do not admit
Sums	0·60	0·40	1·0			

*Wrong decisions.

probability of a case being incorrectly discharged is the sum of the asterisked numbers in column 1:

$$0{\cdot}0331+0{\cdot}0385=0{\cdot}0716$$

The frequencies of the other three categories are calculated primarily and table III shows all four frequencies. Such a table is sometimes referred to technically, and perhaps aptly, as a confusion matrix.

TABLE III—*Proportions of correct and incorrect decisions derived from table II (confusion matrix) using all three indicants. Proportion misclassified = 0·1523*

		Assigned class		
		A	D	Prior
True class	Admission	0·5284	0·0716	0·6
	No admission	0·0807	0·3193	0·4

Evaluation of the set of indicants

The value of the set of three indicants can now, in principle, be estimated if some figure of cost can be put on the consequences of a wrong decision by multiplying the cost of each decision by the probability of its occurrence and adding the products. To this would be added the cost of the test where this is appreciable. Probably most clinicians would sooner admit a patient wrongly than discharge one wrongly. Administrators or competitors for beds in the hospital might place a lower cost on this type of misclassification. It is instructive to repeat the calculations of table II using different prior probabilities and to examine the effects of varying costs, using the resulting confusion matrix.

Evaluation of a single indicant

If one wishes to assess the value of a single test then the confusion matrix can be recalculated omitting this test from the pattern and comparing the costs calculated from the new confusion matrix. Alternatively, a set of indicants can be built up sequentially, adding on the most valuable test at each step. Table IV shows the confusion matrix for the two symptoms only, without guarding. Leaving out guarding greatly increases the number of those incorrectly discharged but makes little difference to the numbers wrongly admitted. The advantage of this approach is that it evaluates a test *in relation to the other evidence available*, which is not the common practice, especially with biochemical tests, where the specificity and sensitivity are usually given in splendid isolation from the context in which they are to be used.

TABLE IV—*Confusion matrix using information on severity and progress of pain. Overall proportion of misclassification is 0·3493 (= 0·2532 + 0·0961)*

		Assigned class		
		Admit	Do not admit	Prior
True class	Admit	0·3467	0·2532	0·6
	Do not admit	0·0961	0·3038	0·4

Problems in the application of Bayes's theorem

The main difficulty in applying the Bayes procedure arises when the number of indicants is large and when they have many categories. It is not uncommon in this kind of work for the clinician to suggest 100 or more indicants of possible diagnostic significance, some with several categories such as absent, mild, moderate, severe, very severe. In these circumstances the number of possible patterns rises astronomically. de Dombal's[2] very moderate set of attributes for diagnosis of the acute abdomen is capable of generating about 10^{17} patterns (compare world population 4×10^9). It is possible that the uniqueness that arises from this profusion of patterns is connected with the recognition by the doctor that each patient is an individual. Several simplifications are possible, of which the commonest is to assume that the states of the attributes are independent. In the example discussed the three used are statistically independent, but if abdominal rigidity and rebound tenderness were included the presence or absence of either of them would not be independent of that in the other or with guarding. Indeed, some clinicians might not regard rigidity as distinct from guarding at all.

TABLE V—*Frequencies of manifestation of possible states of severity of pain, progress of pain, and presence or absence of guarding in the two admission groups*

	Severity of pain		Progress of pain		Guarding	
	Moderate	Severe	Better	Worse	Absent	Present
Admit	0·661	0·339	0·605	0·395	0·211	0·789
Discharge	0·904	0·096	0·827	0·173	0·875	0·125

The assumption of independence is often described as "the Bayes method" but the example discussed should show that the use of Bayes's theorem is quite unaffected by the existence of correlations, provided that the pattern frequencies can be estimated. If independence is assumed then the pattern frequencies are given by the product of the probabilities of the states of the indicants. These probabilities for the example are given in table V and, for example, the frequency of the pattern (moderate; worse; guarding present) is estimated from them as:

$$0{\cdot}661\times0{\cdot}395\times0{\cdot}789=0{\cdot}206$$

for cases needing admission, and

$$0{\cdot}903\times0{\cdot}173\times0{\cdot}125=0{\cdot}0195$$

for those not needing admission. These values are quite close to the observed frequencies, 0·183 and 0·029.

The independence assumption has been found in practice to give very useful results even when it is known to be quite untrue, mainly because the estimated values of the pattern frequencies are regarded as unimportant, provided that they arrange the possible diagnoses in their correct order of relative likelihood. If the probabilities were to be used in conjunction with costs to make a quantitative estimate of the value of a test or the cost of a wrong decision an accurate estimate of their value would be needed.

Computers

An alternative approach which takes account of the interdependence of the tests, is to use brute computer force to examine all combinations of indicants and select those patterns which give the best discrimination between the groups. It is not usually practicable to grind out combinations of more than about five indicants because so many possible patterns exist that either the frequencies in any one pattern are too small to be reliable or the task is simply too big for present-day computers—for example, there are about 75m patterns of five in 100 two-state attributes. The relative usefulness of this approach and the independence assumption is still unsettled. In general the combinatorial method produces much simpler rules that need no computer for their application, but is less accurate. Use of the independence assumption, if it is to improve on this, requires in practice a small computer and gives little or no understanding of what basic patterns are concerned. In any individual problem it is always worth trying both. Some other possible methods will be discussed in the last part of this series.

In spite of these complications it is hoped that this exposition has shown how clearly the use of Bayes's theorem sets out the essential elements in the quantitative evaluation of diagnostic tests.

Bibliography

[1] Card WI, Good IJ. Logical foundations of medicine. *Br Med J* 1971;i:718-20.

[2] de Dombal FT, Leaper DJ, Horrocks Jane C, *et al*. Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians. *Br Med J* 1974;i:376-80.

One of the best and least technical accounts of the ideas discussed in this paper is *Making Decisions* by Professor D V Lindley (New York, John Wiley Interscience, 1971). Some practical examples and further references are to be found in the *Journal of the Royal College of Physicians*, 1975, vol 9.

Clinical Topics

Low-birth-weight infants in Bradford 1972-9

P J CONGDON, G T LEALMAN

Many reports emphasise the improved outlook for low-birth-weight babies (those under 2500 g) as a result of using modern methods of intensive care,[1] [2] although the outlook for those weighing less than 1500 g may be no better than it was 15 years ago.[3] Because these babies need such a high degree of care, regional perinatal centres have been developed to accept either the high-risk pregnancy before delivery[4] or the ill newborn baby.[5] It is, however, difficult to judge from figures published from referral centres the effect that modern intensive care has on reducing neonatal mortality rates in the regions they serve.

We report the outcome for all babies under 2500 g born and treated in two Bradford hospitals from January 1972 to 31 August 1979. Analysing our figures over this period should show the effect that the provision of improved perinatal care has had on altering neonatal mortality in a given community. Although only a relatively few infants needed respiratory support, the overall raising of standards has resulted in a consistent improvement in survival rates at all birth weights. For those under 1000 g, however, the mortality still remains very high.

Department of Paediatrics, Bradford Royal Infirmary, Bradford BD9 6RJ

P J CONGDON, MRCP, DCH, senior paediatric registrar
G T LEALMAN, MRCP, BSC, consultant paediatrician

Patients and methods

Bradford district with a total population of 343 200 (1978 estimate) is served by two maternity hospitals with a total of 5000-6000 deliveries a year. (This includes a net inflow to the Bradford Health District of about 500 maternity patients from surrounding districts, and these patients are included in our figures.) There is a large and increasing immigrant population with an estimated 44 000 New Commonwealth and Pakistani citizens (1978 figures). This is reflected in an increasing number of births to Asian parents, which in 1972 represented 19·7% of all deliveries in Bradford hospitals, but which had increased to 27·9% in 1978 (table I). Almost 99% of deliveries in the district now take place within these two hospitals with consultant and GP units sharing common delivery suites. Only 13·5% of births are to parents of social classes I and II while 34% are to those from social classes IV-V (Registrar-General's classification). There are two special-care baby units, one with 16 cots and one with 20 cots including three neonatal intensive-care cots, which provide ventilatory

TABLE I—*Total numbers of births in Bradford January 1972-August 1979 with percentage of births born to Asian mothers*

Year	Total No of births	% Asians
1972	6025	19·7
1973	5501	20·3
1974	5154	21·8
1975	4993	21·7
1976	5174	23·5
1977	5143	27·4
1978	5549	27·9
1979 (Jan-Aug)	4484	Not available