Mathematics and Medicine

Test reduction: III—Practical applications

W I CARD, P A EMERSON

Summary and conclusions

While there is no generally applicable method of test reduction, ways in which the problem can be tackled can usually be suggested. These include the examination of error rates of tests and the use of current statistical methods. When the cost-effectiveness of some more expensive test has to be measured, a model of the decision problem needs to be developed, with an attempted estimation of the values or utilities of the states of health that result from treatment.

Introduction

This third and last paper in the series is intended to show the practising doctor how these statistical concepts could be applied to his own clinical decision making. Unfortunately there is not sufficient experience to give advice applicable to every case. Ways in which the problem can be tackled can, however, be suggested and useful empirical techniques described.

Why has the application of these techniques to everyday clinical practice been so slow? Though Bayes's theorem is some 200 years old, decision theory, on which much of the thinking of these papers is based, is very recent and dates only from the work of von Neumann and Morgenstern.[1] Further, the existence of a formal theory may produce an unjustified conflict with the art of medicine in the mind of the clinician and may be considered as a reflection on his professional skill. Considerations of cost have encountered even greater resistance in the past—but not now. There can be few who have not been forced to accept that the finances of the NHS are limited, and there must be more rational use of the available resources.

Error rates of tests

Every test, whether it elicits a symptom, sign, or the result of some laboratory investigation, has an error rate such that it gives true-positive results ($p(S/D)$), false-positive results ($p(S/\bar{D})$), true-negative results ($p(\bar{S}/\bar{D})$), and false-negative results ($p(\bar{S}/D)$), where S is a symptom or equivalent, D is a disease, $\bar{S}$ the absence of the symptom, and $\bar{D}$ the absence of the disease. Ideally the doctor should know these probabilities for every test he uses. If experienced he may have some idea of these probabilities for clinical symptoms and signs but be less certain about the probabilities for some laboratory tests. While laboratories ought to provide these probabilities they often do not have the necessary data to provide them with much confidence.

The use of a test can be excluded on its error rates alone. It has been recognised for many years[2] that if the false-positive rate of a test is sufficiently high it may, when applied to a population where the incidence rate—that is, the initial probability of a disease—is sufficiently low, be more often false than true. The test in such an application is valueless and its use a waste of time and money.

Efficiency of tests

An application of test reduction is to the problem of the differential diagnosis of some limited group of disorders—for example, the acute abdomen or dyspepsia—or the problem of deciding which patient with chest pain needs admission or which patient with jaundice needs a laparotomy. Other examples may be concerned with prognosis—for example, forecasting which of several different outcomes may follow severe head injury—or the recognition of which patient with chronic air flow obstruction may be expected to respond to corticosteroids. All these problems are medically different but they are similar mathematically, and consist in assigning a patient to a certain class, either diagnostic, prognostic, or therapeutic, on the basis of evidence supplied from a series of tests.

In all these problems large numbers of tests may be used. This is particularly so when the study has been prospective and a group of clinicians have included their own tests, what might be called their private hunches, so that their value can be experimentally measured. Test reduction here aims to reduce the original set of tests so that the degree of misclassification, or misdiagnosis, is not increased. We call this problem one of efficiency and it is not directly concerned with questions of cost.

The analysis of simple test selection was described in the second of these papers, but the problem becomes difficult or unmanageable when there are large numbers of tests, and dependency between tests exists. The concept of dependency is understood by all doctors, though they may have difficulty in expressing it formally. The symptoms of the passage of dark urine and pale stools are obviously dependent. By this we mean that the probability of the symptom of pale stools being present is greatly increased when, in a jaundiced patient, we have already elicited the history of the passage of dark urine. If these two symptoms were statistically independent the probability of pale stools would be unaltered by this previous knowledge. Though dependence is obvious in this example it is not of great importance since its existence does not help us to distinguish between different kinds of obstructive jaundice. The important examples of dependence are those that occur within one disease but not in another and thus give rise to a pattern of symptoms or signs that are recognised as being highly diagnostic. An example would occur in a man aged over 55 who has never

Diagnostic Methodology Research Unit, Southern General Hospital, Glasgow G51 4TF

W I CARD, MD, FRCP, professor emeritus University of Glasgow

Westminster Hospital, London SW1 2AP

PETER A EMERSON, MD, FRCP, consultant physician

previously suffered from dyspepsia, who gave a recent history of abdominal discomfort or pain, which was not episodic and which was getting worse. We recognise that such a pattern is highly suggestive of carcinoma of the stomach, though individually each of the indicants might occur in a patient with a chronic duodenal ulcer.

If we have many test results (indicants) the number of possible combinations can be very great. If there are 50 possible indicants, the number of possible combinations or patterns taken four at a time is nearly a quarter of a million. If we want to look at combinations of five at a time, there are over two million. These numbers are unmanageable. Naively we might suppose that if we found the single most useful test for our discriminative procedure, retained this, and sought for the next most powerful test to be added to it and continued in this way, we could accumulate the most powerful combination of any size for our purpose. Unfortunately this method does not work when dependency is present. Teather, who investigated the statistical techniques useful in distinguishing "medical" from "surgical" jaundice,[3] identified the single most powerful discriminative test, the most powerful pair, and then the most powerful triad (personal communication). The most powerful triad did not contain the single most powerful test or either of the tests used in the most powerful pair. This paradoxical result rests on dependency. Teather in the same paper[3] showed in a theoretical study that two indicants, each of which alone gave no discrimination whatever, can be combined to give perfect discrimination between two classes. The position may be summarised by saying that, while we have at present no mathematically perfect method for selecting tests, current statistical methods can give valuable guidance.

Test reduction need not be complete since the retention of some redundancy helps to compensate for the error that is inherent in all tests. A familiar example of this mechanism occurs with the English language, which has been calculated to be 50% redundant.[4] If English were rewritten to exclude redundancy, there would be a great saving of time in reading or listening to it but if one word were missed the sentence containing it would become meaningless. Redundancy can be thought of as a simple error-correcting mechanism.

Cost-effectiveness of tests

The third main problem is measuring the value of a single test when the test has a significant cost. This may be monetary in terms of equipment and staff or cost to the patient in terms of discomfort and danger. Both these sorts of costs must be set against the expected improvement in the patient's health. We may call this the problem of cost effectiveness. Sometimes a simple test would lead to completely unjustifiable costs. The American Cancer Society advised that, to detect cancer of the bowel, six tests should be made for occult blood in the stool and that a positive result should be followed by a barium enema examination. Neuhauser and Levicke[5] analysed this advice and by making certain plausible assumptions of the costs of tests and of barium examinations, showed that, if it were followed, detecting a case of cancer of the colon after the sixth occult blood test would cost \$47 000 000. While such advice stands self-condemned, the analysis of cost effectiveness usually requires the concept of utilities discussed in the first paper and the technique of decision theory.[6] The following example illustrates the method of analysis.

We suppose that a man aged 55 has had a history of indigestion for some months and a gastric ulcer is seen on radiological examination. If we have no access to endoscopic examination we have to decide, on the basis of the radiological examination alone, on either medical treatment, assuming that the ulcer is benign, or surgical resection, assuming that the ulcer is malignant. The components of the decision problem can be set out in a table.

At the top of each column are what we may call the events, the designations being shortened here to "benign" and "malignant," and the rows follow from the decisions, "medical treatment" and "surgical treatment," on the left. By "medical treatment" we mean a course of cimetidine and a further *x*-ray examination in six weeks. In the body of the table are the "utilities," which are numbers corresponding to the states of health which result from taking particular decisions. A convenient scale for utilities is from zero, representing death, to one, representing perfect health. For example, if the ulcer is really malignant and the patient has medical treatment, the resulting state of health is represented by the utility 0·281. We must emphasise, however, that these numbers are no more than plausible estimates; the method of estimating them will be discussed later. At the bottom of the columns are the probabilities of the events named at the top. In this example we could

Treatment of gastric ulcer: decision table

	Benign	Malignant
Medical treatment	1·00	0·281
Surgical treatment	0·970	0·331
Probabilities	0·96	0·04

assume that the probability of the ulcer's being malignant was 0·04—that is, a 4% chance—and therefore that the probability of its being benign was 0·96—that is, a 96% chance. We can now estimate the expected value of a treatment—that is, its value on the average. The expected value of medical treatment will be $(0{\cdot}96\times1{\cdot}00)+(0{\cdot}04\times0{\cdot}281)=0{\cdot}971$, while the expected value of surgical treatment will be $(0{\cdot}96\times0{\cdot}970)+(0{\cdot}04\times0{\cdot}331)=0{\cdot}944$. For the moment, ignoring the costs of treatment and with no other information, we can only choose the treatment with the greatest expected utility, in this case medical treatment. Informally, we may say that this is the best bet.

This is the course of action that we pursue on the evidence so far available but we can now ask the question: would gastroscopy be worthwhile? Gastroscopy in the hands of an experienced endoscopist with biopsy examination is a highly accurate procedure and we may assume, for simplicity, that endoscopy gives perfect information. By perfect information we mean that it will tell us with certainty whether the ulcer is benign or malignant, so that the expected utility of treatment with perfect information is obtained by multiplying the maximal utility in each column by the appropriate probability and summing the answers—that is, $(0{\cdot}96\times1{\cdot}00)+(0{\cdot}04\times0{\cdot}331)=0{\cdot}973$. The difference between this and the greatest expected utility derived from deciding on the best treatment as discussed above, $0{\cdot}973-0{\cdot}971=0{\cdot}002$, may be called the expected value of perfect information. This gain in utility has now to be set against the cost of the procedure to decide whether the additional cost is "worthwhile." A possible method of comparison is to say: Let C_1 be the total cost of treating the patient without endoscopy and let C_2 be the cost of an endoscopy, then endoscopy should be performed if

$$\frac{\text{Expected utility with endoscopy}}{C_1+C_2} > \frac{\text{expected utility without endoscopy}}{C_1}$$

In the practical problem of estimating costs of medical and surgical procedures we have found the cost accountants of the area health board most helpful. Using the costs they estimated, and given all the assumptions, we found that gastroscopy could not be justified. But we have also to consider the further situation in which the ulcer fails to heal after six weeks' medical treatment. The probability that it is now malignant will rise steeply to perhaps 0·5, and we must then recalculate the decision after adjusting the matrix of utilities and including the change in probabilities. Calculation now shows that, given the assumptions, gastroscopy is justified in preference to surgical treatment, which is otherwise the treatment of choice. We have made all calculations on the assumption that gastroscopy is error free. This is untrue but the error is slight, can be measured, and can easily be included in the calculations.[6]

This model of the decision problem is an accepted one and the development of a satisfactory model is usually half the battle. The other half is the estimation of values for the variables, in particular the estimation of the utilities for states of health. We can only touch on this difficult problem.

Utilities have an "as if" existence. If doctors act consistently then their actions can be thought of *as if* there were a value or utility attached to a state of health coupled with a probability of attaining it. We can elicit this implicit utility only by testing a doctor in either a real or a simulated situation and finding out with what probabilities of success or failure he would advise a particular course of action such as a surgical operation. From his replies to a group of problems, we can estimate his utility function—that is, the values or utilities he attaches to a set of states of health. The method is promising but is in its infancy.[7-10] We give this example of an "ulcer-cancer" to show that we can now envisage the possibility of measuring the cost effectiveness of a test with some degree of objectivity.

Professor W I Card acknowledges with gratitude a personal grant from the Scottish Home and Health Department.

These papers have been written by members of the Computer Workshop of the Royal College of Physicians, which seeks to study the extent of the contribution that mathematical methods of decision making can make to clinical medicine. The workshop is open to research workers, and advice is available to those who are interested in attacking the sort of problems discussed in this paper. Letters should be addressed to Dr Peter Emerson, chairman of the Computer Committee, Royal College of Physicians, 11 St Andrews Place, London NW1 4LE.

This is the third of three papers in this series.

References

1 von Neumann J, Morgenstern O. *Theory of games and economic behaviour.* Princeton; Princeton University Press, 1944.
2 Meehl PE, Rosen A. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol Bull* 1955;**52**:194-216.
3 Teather D. Statistical techniques for diagnosis. *J R Statist Soc A* 1974; **137**:231-44.
4 Shannon CE, Weaver W. *The mathematical theory of communication.* Urbana: University of Illinois Press, 1964;56.
5 Neuhauser D, Levicke AM. What do we gain from the sixth stool guaiac? *N Engl J Med* 1975;**293**:226-8.
6 Lindley D. *Making decisions.* London: Wiley-Interscience, 1971.
7 Card WI, Rusinkiewicz A, Phillips CI. Utility estimation of a set of states of health. *Methods Inf Med* 1977;**16**:168-75.
8 Bunker JP, Barnes BA, Mosteller F. *Costs, risks, and benefits of surgery.* New York: Oxford University Press, 1977.
9 Card, WI. "Is it worth it?" *J R Coll Physicians Lond* 1979;**13**:213-6.
10 Card WI. Rational justification for therapeutic decisions. *Metamedicine* 1980;**1**:11-28.

(*Accepted 21 February 1980*)

For Debate . . .

Migraine prodromes separated from the aura: complete migraine

J N BLAU

Summary and conclusions

Detailed questioning of 50 patients with uncomplicated migraine has shown that 17 had symptoms that preceded the headache phase by several hours. These prodromes consisted of changes in mood, behaviour, wakefulness, appetite, bowel activity, or fluid balance. The term "complete migraine" is proposed for attacks that include prodromal symptoms, whose occurrence implies an initial diffuse cerebral or hypothalamic disturbance.

Introduction

A migraine attack may start suddenly with an aura or develop slowly with a mounting headache. There is a phase, however, even before the headache or aura that is recognised by most clinicians as well as some patients. Thus George Eliot felt "dangerously well before an attack"[1] but Sir John Forbes had an "irresistible and horrid drowsiness"[2]; Lady Conway ate her supper with "a greedy appetite" and her "pain would almost certainly follow the next morning,"[3] whereas Du Bois Reymond's migraines were "in general preceded by constipation."[2]

In a recent review Friedman[4] wrote, "Some investigators believe that these changes (prodromal symptoms) are not related to the migraine syndrome . . . [my] "experience, however, indicates that these symptoms are part of the migraine attack and are therefore truly prodromal." Surprisingly, this facet of the clinical picture of migraine has received little attention. Hence I undertook this prospective study of 50 patients with migraine uncomplicated by other conditions.

National Hospitals for Nervous Diseases, London WC1, and City of London Migraine Clinic, London EC1

J N BLAU, MD, FRCP, consultant physician

Patients

All patients were seen by me in outpatient clinics in neurological or general hospitals or in a migraine clinic, having been referred by a doctor. The history was taken to establish the diagnosis of migraine, as defined,[5] and whether it was the common or classical variety. Patients were then asked if they had any indication beforehand that an attack might follow. Some gave spontaneous descriptions of preheadache symptoms. Others had to be asked directly if they noticed variations