MEDICAL PRACTICE

*Mathematics and medicine***Test reduction: I — Introduction and review of published work**

W I CARD, P A EMERSON

Summary and conclusions

The word "test" is used in a more general sense to describe the process of eliciting evidence of any kind from a patient, and this paper explores the thesis that much evidence is unnecessary and that therefore much test reduction is possible. The value of test reduction can be measured by its efficiency—that is, its effect on the misclassification of disease—or, preferably, by some measure of its cost-effectiveness.

We propose to examine the thesis that much medical evidence that doctors elicit is unnecessary and could safely be reduced. A word is needed to describe this process of eliciting, by whatever means, an element of evidence, and we use the word "test," its application being extended from its more usual restriction to a laboratory "test" to include items from the patient's history, responses to treatment, etc. To obtain evidence the doctor applies a test. But the test has a cost in terms both of money and of inconvenience or even danger to the patient. Tests also incur certain sorts of error. If our aim is to reduce the amount of evidence we need and therefore its cost, we can do this only by reducing the number of tests.

In the traditional system of clinical medicine the doctor uses his training and experience to arrive at a judgment and a decision about a patient. But this whole process, which may be very effective, is private; no precise measurements are made of probabilities or of evidence, and thus rational reduction of tests is difficult. This becomes generally possible only if we are

prepared to use a formal system of medicine that includes these types of measurements, so that we can calculate whether the inclusion of a test and the resulting evidence improves the accuracy of diagnosis or prognosis or whether it could safely be discarded.

For the element of evidence provided by the result of a test the word "indicant" has been proposed by Card and Good.¹ Given the set of indicants, the patient can be allocated with a certain probability to a disease class and a prognostic class. Allocation to a class follows a sequence of decisions entailing the choice of certain tests; the mathematics needed are those of probability and decision theory.

Initially the doctor is seeking evidence relevant to a hypothesis that a certain "disease" is present. To elicit evidence he asks questions; performs a physical examination; and may order various laboratory, radiological, and special examinations. In our terminology all these are tests, and each test produces one of a set of results that are more or less characteristic of particular aspects of the disease process—for example, the set of enlargements of the spleen or the set of values of the serum cholesterol. Each such test supplies evidence that alters the probability that the patient belongs to one class rather than another according to some method of inference. This sequential decision procedure continues until ultimately the doctor makes a decision on treatment. The effect of treatment may itself provide further evidence—hence the phrase "a therapeutic test." We could even argue that a necropsy is a test; the evidence, though of no value to the dead patient, might be of great value to the community. We will consider methods of combining the evidence from several tests at greater length in future sections.

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

Review of published work

ALLOCATION TO A DISEASE CLASS

In the past 20 years several papers have been published showing that test reduction is possible. One of the earliest was a careful

study by Zieve and Hill² of nine liver function tests performed on a large group of normal people and also on a group of 41 patients with cirrhosis of the liver of varying severity. Using a linear discriminant function, they showed that four tests discriminated normal subjects from cirrhotic patients as accurately as nine tests. Wilson *et al*³ analysed the radiological signs that were most helpful in distinguishing between benign and malignant gastric ulcer. They were able to reduce an original set of 70 variables to 17. Using this reduced group they correctly diagnosed all of a new series of 14 ulcers. In the diagnosis of patients with upper abdominal pain Scheinok and Rinaldo⁴ chose 11 attributes and then examined all possible subsets from three, the minimum necessary to diagnose six diseases, to 11. In this way they showed that six attributes carried nearly all the information contained in the original 11. In a similar type of study Youker *et al*⁵ investigated the appearances that differentiated benign and malignant polyps of the colon and showed an 83% accuracy in differentiation when they used seven attributes. They then investigated all possible sets of three and found that most polyps fitted into relatively few combinations and that some of these were very powerful.

In a survey of 1238 patients Pipberger *et al*⁶ investigated the amount of information necessary to distinguish causes of chest pain. The symptoms, signs, and results of laboratory tests (which included 429 questions with "yes-no" answers) amounted to 498. To differentiate between four diseases only 55 items were found to be necessary. When the disease groups were reduced to two, coronary artery disease and pneumonia, they could be separated by discriminant function analysis with only six tests. Influenced by the original work of Scheinok and Rinaldo,⁴ Bouckaert⁷ sought for subsets of symptoms for diagnosing five kinds of goitre. Using 12 clinical symptoms, he examined the diagnostic information for each set of three symptoms using measurements of uncertainty. In his search he used a stepwise method, though realising that this did not necessarily identify the best set of symptoms. The triad he obtained proved only slightly less accurate than the conventional Bayesian method using all 12 symptoms.

Spicer *et al*⁸ investigated the number of variables needed to discriminate between two non-specific inflammatory diseases of the large bowel, proctocolitis and Crohn's disease. From 91 possible variables they found, by a stepwise linear discriminant analysis, that seven variables carried most of the discriminating information; further study, which took their interdependence into account, reduced the number to five. The authors suggested that this empirical procedure was a possible method for other diseases. In a study designed to identify patients with hysteria, Woodruff *et al*⁹ reduced the original 60 symptoms used to 14. This reduced number, arranged in a decision "tree," gave results comparable to the more extensive and lengthier interview normally used. They suggested that the method could serve as a screening procedure. Krypsin and Norwich¹⁰ used an application of information theory to reduce the number of symptoms needed, and showed that in the analysis of brain tumours 41 variables could be reduced to eight with minimal loss of information. In a study of the differentiation between "medical" and "surgical" jaundice Teather¹¹ showed that the original set of 64 variables could be reduced to four without any increase in misclassification. Gardner and Barker¹² sought a rule for referring to hospital for biochemical tests patients who had been treated with radioactive iodine and who had been sent a postal questionnaire with nine items relating to hypothyroidism. They tried several refined statistical techniques but found that a simple rule, that of referring any patient with four or more symptoms, proved the method of choice.

ALLOCATION TO A PROGNOSTIC CLASS

All these examples of test reduction refer to the allocation of a patient to a disease class. But allocation to a prognostic class is mathematically similar, and test reduction has been successfully applied to this also (for example, by Jennett *et al*¹³ and Spicer¹⁴).

Carney *et al*¹⁵ were concerned with the diagnosis of the depressive syndrome and selected 35 items that were scored for all patients. The analysis of 129 depressive patients was consistent with two distinct classes of depression—endogenous and exogenous—and a smaller set of 10 items proved as accurate as the full set of 35 in predicting the outcome of electroconvulsive therapy. In predicting the outcome after two years in patients with Hodgkin's disease Brook and Gocka,¹⁶ after a preliminary screening of the predictor variables, showed that five of the 14 chosen for regression analysis contained nearly all the information. Alvord *et al*¹⁷ showed that the probability of survival of patients with subarachnoid haemorrhage due to ruptured intra-

cranial aneurysm could be estimated from the length of survival from the last haemorrhage and the clinical grading of the patient at examination. Bevegaard *et al*¹⁸ in predicting the outcomes of survival or death in acute myocardial infarction observed 24 cardiorespiratory variables in 19 patients. Using discriminant function analysis, they showed that most of the information needed to separate the two groups was contained in three items. Helmers¹⁹ studied both the short-term and the long-term indices of prognosis in acute myocardial infarction. The final indices, which were tested prospectively, contained few factors and pointed to considerable test reduction.

Measurement of success in test reduction

The degree to which tests can be reduced is therefore considerable and we require methods for assessing the value of such reduction. They may be classified under the headings of efficiency and cost effectiveness.

EFFICIENCY

In all the above examples the efficiency of reduction was measured by the reduced number of tests that produced no increased misclassification. The degree of misclassification is measured in the simplest case by a 2×2 table, where the allocation to the cells is made on the basis of "correct" and "calculated" classification (figure). The incorrectly classified are in the

		Correct diagnosis	
		D	$\bar{D}$
Calculated diagnosis	D	a	b
	$\bar{D}$	c	d

A table from which degree of misclassification can be estimated. Diagnosis is between D, disease, and $\bar{D}$, absence of disease. "a" is number of patients calculated as having D and who actually had D; true positives. "b" is number of patients calculated as having D but who in fact had $\bar{D}$; false positives. "c" is number of patients diagnosed as $\bar{D}$, but who actually had D; false negatives. "d" is number of patients calculated correctly to have $\bar{D}$; true negatives.

$$\text{Degree of misclassification is } \frac{b+c}{a+b+c+d}$$

bottom left and top right cells, corresponding to a false-negative and a false-positive class. The total degree of misclassification is the proportion of incorrect allocations. The tabulation of more outcomes requires a larger matrix, the correctly classified patients all lying on the main diagonal running from the top left-hand to the bottom right-hand corner. While this method is useful as a first approximation, it assumes that all misclassification mistakes are of equal importance. A little reflection will show that mistakes in misclassification range from the trivial to the extremely serious. We have therefore to consider the cost, in some sense, of misclassification, which requires the introduction of new concepts. Such considerations were first clearly discussed by Ledley and Lusted²⁰ in a masterly paper.

COST EFFECTIVENESS

A basic concept is that of utility. By this concept we postulate that a number can be attached to any state of health that in some way represents its value or worth, and is called a utility. If an effective treatment is given to a patient there will, on average, be a gain in utility. If, owing to misclassification—that is, a misdiagnosis—the wrong treatment is given, no gain but even a loss of utility may result; this loss can, in principle, be measured. An example is the treatment of an ulcer in the stomach, which is either benign, and should heal with simple medical treatment, or malignant, and requires operation. There is then a loss of utility if the malignant ulcer is called benign and is treated medically so that potentially curative surgery is delayed. But there is also a loss of utility if a benign ulcer, which would have healed completely, is operated on unnecessarily.

But this analysis is insufficient since we have to balance the benefits, the gains in utility, against the costs of treatment. These costs have to be considered as monetary costs and what may be called biological costs, which include pain and danger to life. If the balance is to be assessed the utilities must be measured in the same units as the costs, which means that we have to find a monetary equivalent to each state of health or devise some other method to avoid this difficulty.

In the above example concerning the distinction between a benign and a malignant gastric ulcer, an additional test could be endoscopy of the stomach. This test will, on average, improve accuracy and thus reduce misclassification, and this in its turn will improve the chances of successful treatment. There will thus be a gain in expected utility, where “expected” is the technical term for “on average.” This gain in expected utility has now to be set off against the cost of the test. The gain in expected utility for ordinary treatment divided by its cost can be compared with the gain in expected utility from the additional use of endoscopy divided by the increased cost. If this second fraction is greater than the first the test should be included, otherwise its inclusion cannot be justified. This problem is dealt with further in the third paper in this series.

References

- ¹ Card WI, Good IJ. In: Passmore R, Robson JS, eds. *A logical analysis of medicine in companion to medical studies*. Vol 3. Edinburgh: Blackwell, 1974.

Are there any health hazards associated with using unlined cast-iron cooking pots? I understand the haemosiderosis prevalent in southern Africa is associated with the use of iron cooking pots?

In Britain legislation issued by the Department of Prices and Consumer Protection safeguards the public from specific health risks associated with cooking utensils. Unlined cast-iron cooking pots are not, however, mentioned in this legislation and are not thought to present any specific risks. Although the iron overload described in adult Bantu has been attributed to the preparation of acid-fermented foodstuffs in iron utensils—for instance, porridge (marewa) and kaffir beer—malnutrition and infections may also be important factors.¹ Whether this iron overload is detrimental to wellbeing is debatable.¹

- ¹ Walker ARP, Arvidsson UB. Iron “overload” in the South African Bantu. *Trans R Soc Trop Med Hyg* 1953;47:536-48.

What is the mechanism of the Babinski reflex?

There are wide gaps in our knowledge of the neurophysiological basis for many of the physical signs we elicit during a neurological examination, and, although there are many theories concerning the Babinski reflex, the precise mechanism is uncertain.¹ What may be stated with certainty is that, other than in infants, the presence of a Babinski

- ² Zieve L, Hill E. An evaluation of factors influencing the discriminative effectiveness of a group of liver function tests. The utilisation of multiple measurements in medicine. *Gastroenterology* 1955;28:759-84.
- ³ Wilson WJ, Templeton AW, Turner AH, Lodwich GS. The computer analysis and diagnosis of gastric ulcers. *Radiology* 1965;85:1064-73.
- ⁴ Scheinok PA, Rinaldo JA. Symptom diagnosis: a comparison of mathematical models related to upper abdominal pain. *Comput Biomed Res* 1968;1:475-89.
- ⁵ Youker JF, Welin S, Main G. Computer analysis in the differentiation of benign and malignant polypoid lesions at the colon. *Radiology* 1968;90:794-6.
- ⁶ Pipberger HV, Klingerman JD, Cosma J. Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain. *Methods Inf Med* 1968;7:79-97.
- ⁷ Bouckaert A. Computer diagnosis of goitres. 111. Optimal subsymptomatology. *J Chronic Dis* 1971;24:321-7.
- ⁸ Spicer CC, Jones J, Jones JE. Discriminant and Bayes analysis in the differential diagnosis of Crohn's disease and proctocolitis. *Methods Inf Med* 1973;12:118-22.
- ⁹ Woodruff RA Jr, Robins LN, Taibleson M. A computer assisted derivation of a screening interview for hysteria. *Arch Gen Psychiatry* 1973;29:450-4.
- ¹⁰ Krypsin J, Norwich AM. Application of information calculus to medical data analysis and reduction. *Mathematical Biosciences* 1973;17:165-172.
- ¹¹ Teather D. Statistical techniques for diagnosis. *Journal of the Royal Statistical Society A* 1974;137, 2:231-44.
- ¹² Gardner MJ, Barker DJP. Diagnosis of hypothyroidism: a comparison of statistical techniques. *Br Med J* 1975;ii:260-2.
- ¹³ Jennett B, Teasdale GM, Knill-Jones RP. Predicting outcome after head injury. *J R Coll Physicians Lond* 1975;9:231-7.
- ¹⁴ Spicer CC. The prediction of success or failure of medical treatment for acute colitis. *J R Coll Physicians Lond* 1975;9:252-6.
- ¹⁵ Carney MWP, Roth M, Garside RF. The diagnosis of depressive syndromes and the prediction of ECT response. *Br J Psychiatry* 1965;11:659-74.
- ¹⁶ Brook J, Gocka EE. Prediction and analysis of survival in advanced Hodgkin's disease. *Comput Biomed Res* 1972;5:659-72.
- ¹⁷ Alvord EC, Loeser JD, Bailey WL, Copass MK. Subarachnoid haemorrhage due to ruptured aneurysms. *Arch Neurol* 1972;27:273-84.
- ¹⁸ Bevegaard S, Blomquist P, Hulting J, Matell G, Nygaard ME, Wigertz O. Assessment of prognosis in acute myocardial infarction by computer analysis of cardiorespiratory variables. *Acta Med Scand* 1973;194:75-80.
- ¹⁹ Helmers C. Short and long-term prognostic indices in acute myocardial infarction. *Acta Med Scand* 1974;555,suppl:1-86.
- ²⁰ Ledley RS, Lusted LB. Medical diagnosis and modern decision making. Mathematical problems in the biological sciences. *Proceedings of Symposia in Applied Mathematics* 1962;14:117-58.

This is the first of three papers in this series.

(Accepted 21 February 1980)

reflex (Babinski response or extensor plantar response) is abnormal and indicative of a disturbance of the pyramidal tract. Occasionally, this may be transient, such as after a seizure. The main value of the sign is that it may be present when other signs of a pyramidal tract lesion, such as increased tone, increased reflexes, or weakness, are absent. Its main weakness is that it may be absent despite overt pyramidal signs.

- ¹ Van Gijn J. *The plantar reflex. A historical clinical and electromyographic study*. Krips Repro—Meppel 1977.

Three Indian children all suffer from severe eczema that has been treated in the normal way with an avoidance of allergens, in this case milk, and with emulsifying ointment and 1% hydrocortisone when necessary. Their skin, however, has improved since the mother used an Indian treatment of mustard oil. This had cured one of her other children in the past, and she finds it more effective than the emulsifying ointment. Is this treatment common, and is there any suspicion of carcinogenic action of the mustard oil?

This treatment is used as a “village medicine” in some parts of the Indian subcontinent. The type of mustard oil used is bland, not irritant, and there is no evidence of carcinogenicity. The Indian *Pharmacopoeia* also specifies *Brassica juncea* as the source of expressed mustard oil that can be used as a mild rubefacient.