The Analysis of Diagnostic Data

D. TEATHER, MSc, Lecturer in Statistics, Polytechnic of Central London, and W. HILDER, BSc, Division of Medical Computing, M.R.C. Clinical Research Centre, Northwick Park, Harrow

In considering a mathematical approach to diagnosis, attention is confined to situations in which the patients under study may be assumed to have one of a fixed list of diseases. To decide which disease a particular patient has, the doctor or clinician obtains information from the patient by means of questions, examinations, and tests. Any characteristic observed, such as age, body temperature, presence or absence of pain, is called an ‘attribute’. Thus, the clinician is required to make two decisions: firstly, which attributes should be observed, and secondly, which disease should be predicted on the basis of these observations.

The main difficulty facing both clinician and mathematician is the large number of attributes available; typically well over a hundred. One approach to the problem is based on Bayes’s theorem and the use of computer assistance (de Dombal et al., 1972).

THE BAYES THEOREM AND INDEPENDENCE

A simple example will illustrate the application of Bayes’s theorem to diagnosis. Consider the problem of differentiating between two diseases, D1 and D2, using two laboratory tests, LAB1 and LAB2, whose results are simply classified as normal or abnormal.

Given a patient with a normal result for LAB1, Bayes’s theorem gives

\[ p(D1|LAB1 \text{ normal}) = \frac{k \cdot p(LAB1 \text{ normal}|D1) \cdot p(D1)}{p(D2|LAB1 \text{ normal})} \]

or equivalently

\[ \frac{p(D1|LAB1 \text{ normal})}{p(D2|LAB1 \text{ normal})} = \frac{p(LAB1 \text{ normal}|D1) \cdot p(D1)}{p(LAB1 \text{ normal}|D2) \cdot p(D2)} \]

Notice that LAB1 is of no value for diagnosis if the incidence of normal test results is the same in both diseases.

In the case of both LAB1 and LAB2 giving normal results we have,

\[ p(D1|LAB1 \text{ normal, LAB2 normal}) \propto p(LAB1 \text{ normal, LAB2 normal}|D1) \cdot p(D1) \]
Further, assuming the test results as independent we have,
\[ p(\text{LAB1 normal, LAB2 normal}|D1) = p(\text{LAB1 normal}|D1) \cdot p(\text{LAB2 normal}|D1) \]

Thus, in studying the usefulness of both tests we need consider only the marginal incidence of normal results for LAB1 alone and LAB2 alone in the two diseases, if the independence assumption is made. This, however, can be quite misleading, as a hypothetical example illustrates.

**Table 1. Two laboratory tests, each individually giving no discrimination**

|       | LAB1 |   | LAB2 |   |
|-------|------|---|------|---|
|       | D1   | D2|      |   |
| Normal| 25   | 25|      |   |
| Abnormal| 25  | 25|      |   |

|       | LAB1 |   | LAB2 |   |
|-------|------|---|------|---|
|       | D1   | D2|      |   |
| Normal| 25   | 0 |      |   |
| Normal| 0    | 25|      |   |
| Abnormal| 0   | 25|      |   |
| Abnormal| 25  | 0 |      |   |

Table 1 shows the incidences, expressed as percentages, of normal test results in both diseases. Note that for both tests the incidence of normal results is the same in both diseases. Assuming independence, we conclude that neither test is useful for diagnosis. However, the entire joint distribution (Table 2) for the two tests and the two diseases shows that the test results give perfect discrimination between the diseases. Practical examples of such difficulties have been discussed in detail by Teather (1974a).

**DATA ANALYSIS**

In practice, the incidences of the various symptoms and diseases are unknown. Typically, these must be estimated from a data base of past patient records
for which both attributes and disease are known. Clearly, if the independence assumption is not made, there are too many probabilities to estimate if all the attributes are considered. However, some progress can be made if only a few of the attributes are considered. The data base may then be used to construct a contingency table showing the observed incidences of a subset of the attributes and the diseases. This contingency table may then be used to construct a simple diagnostic aid. The statistical aspects of this approach are described by Teather (1974b). The method can be illustrated by considering the differential diagnosis of jaundice, using the simplified problem of differentiating between a non-surgical (NS) and a surgical (Surg) form of jaundice. The analysis is based on a series of 207 consecutive admissions to the Liver Unit, King's College Hospital. This is an earlier version of the data base discussed by Knill-Jones et al. (1973). For simplicity, we restrict attention to attributes from the history and physical examination. Continuous attributes such as age were simply categorised so that all attributes were dichotomous. There were 64 available attributes to discriminate between the two diseases.

Table 3. Diagnosis of jaundice, typical subset of three attributes

| Age | Trans | Drugs | NS | Surg | Predict |
|-----|-------|-------|----|------|---------|
| <50 | No    | No    | 36 | 8    | NS      |
| <50 | No    | Yes   | 8  | 2    | NS      |
| <50 | Yes   | No    | 21 | 6    | NS      |
| <50 | Yes   | Yes   | 7  | 0    | NS      |
| >50 | No    | No    | 20 | 50   | Surg    |
| >50 | No    | Yes   | 11 | 2    | NS      |
| >50 | Yes   | No    | 20 | 7    | NS      |
| >50 | Yes   | Yes   | 8  | 1    | NS      |

Table 3 shows the observed frequencies of the two diseases and the three attributes—age, whether the patient was transferred to King's from another hospital (Trans), and whether the patient had been exposed to any hepatotoxic drugs (Drugs). For each of the eight possible outcomes of the three attributes we select the disease that occurred most frequently among the patients in the data base. This defines the prediction rule based on this particular subset of three attributes. The accuracy of the resulting prediction rule is estimated from the observed percentage of data base patients whose disease is correctly predicted. For the rule obtained from Table 3 the accuracy is estimated as 79 per cent (161/207).

This analysis is repeated for all subsets of three attributes from the 64 avail-
able, and the subset with the greatest estimated accuracy is selected. For the jaundice data this turns out to be the three attributes shown in Table 3.

Examining the table we note that for all patients below 50 years of age we predict non-surgical jaundice regardless of the two attributes transfer and drug exposure. We may indicate this simplification by presenting the prediction rule either in the form of a ‘collapsed’ table (Table 4) or diagrammatically in the form of a tree (Fig. 1). The tree evolves from left to right and the path taken for a particular new patient will depend on the observed values of the
attributes for the patient concerned. Thus, in Fig. 1, for a patient under 50 we predict a non-surgical cause which we estimate to have probability 0.82.

**Table 4.** Prediction rule for diagnosis of jaundice based on the best subset of three attributes

|       | Age | Trans | Drugs | NS | Surg | Predict | % Correct |
|-------|-----|-------|-------|----|------|---------|-----------|
| <50   | Yes | 72    | 16    | NS | 82   |         |           |
| >50   | No  | 28    | 8     | NS | 77   |         |           |
| >50   | Yes | 11    | 2     | NS | 84   |         |           |
| >50   | No  | 20    | 50    | Surg| 72   |         |           |

This analysis of subsets, consisting of only a few of the attributes, forms the basis of the method for obtaining simple pencil and paper rules such as that in Fig. 1. Teather (1974b) illustrated this approach and obtained simple diagnostic aids for jaundice that are of comparable accuracy to the computer-aided approach using Bayes's theorem.

**The Computer Program**

The diagnostic procedure (Fig. 1) considered above requires no computer aid when applied to the diagnosis of new patients. However, a large amount of computation has been involved in the production of the tabulations necessary in the search for the best combination of signs.

To aid the scientist in performing the above analysis, a comprehensive program has been written for the production of such tabulations. A detailed account of the program will be given elsewhere by the authors.

The program suite consists of two main sections, the edit program and the analysis program. The analysis program is written in ANSI FORTRAN and, although not completely machine independent, the edit program will produce a version suitable for a particular computer installation.

The following data is utilised by the analysis program in the production of the diagnostic aids—

(i) A data base of past patient records.
(ii) A set of identifiers for the attributes and their associated states (e.g. attribute SEX, 2 states MALE, FEMALE).
(iii) A list of instructions specifying the tabulations to be produced.

The instructions in Table 5 identify the attributes and direct the program to
Table 5. Typical list of instructions for the analysis program

ATTRIBUTES
AGE
SEX FEM MALE
TRAN NO YES

LIVP NO YES
LIVT NO YES
DIAG NS SURG
END
PREDICT DIAG
ADD CATEGORIZED ATTRIBUTE AGE [0-49] < 50 [50-99] > 50
NUMBER OF ATTRIBUTES 3
PRODUCE RULES
KEEP BEST RULE
DISPLAY CONTINGENCY TABLE, PERFORMANCE

Table 6. Typical tabulations obtained from the analysis program

INCIDENCE MATRIX

| AGE | TRAN | DRUG | DIAG | DIAG | TOT | RULE DIAG | CORRECT |
|-----|------|------|------|------|-----|-----------|---------|
| <50 | NO   | NO   | 36   | 8    | 44  | NS        | 81.8    |
| <50 | NO   | YES  | 8    | 2    | 10  | NS        | 80.0    |
| <50 | YES  | NO   | 21   | 6    | 27  | NS        | 77.8    |
| <50 | YES  | YES  | 7    | 0    | 7   | NS        | 100.0   |
| >50 | NO   | NO   | 20   | 50   | 70  | SURG      | 71.5    |
| >50 | NO   | YES  | 11   | 2    | 13  | NS        | 84.4    |
| >50 | YES  | NO   | 20   | 7    | 27  | NS        | 74.1    |
| >50 | YES  | YES  | 8    | 1    | 9   | NS        | 88.9    |

TOTAL       207
MISSING VALUES 0
TOTAL ALL RECORDS 207

PERFORMANCE TABLE

| ACTUAL | DIAG NS SURG |
|--------|--------------|
| PREDICTED NS 111 26 |
| SURG 20 50 |
| PERCENTAGE CORRECT 78.6 |
| .95 CONFIDENCE LIMITS 73.3-82.9 |

produce the rule for each combination of three attributes. The best rule is selected (that which allocates the most cases correctly) and the incidence table, together with the rule, is printed, as shown in Table 6.
Acknowledgements

This work arose out of a recommendation of the Computer Committee of the Royal College of Physicians to make this method of analysis more generally available. The work is being undertaken with the guidance of D. V. Lindley and C. C. Spicer. The data on jaundice was collected at the Liver Unit, King's College Hospital, London, in co-operation with the Department of Medicine in Relation to Mathematics and Computing, Glasgow.

References

de Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P. and Horrocks, J. (1972) British Medical Journal, 2, 9.

Knill-Jones, R. P., Stern, R. B., Girmes, D. H., Maxwell, J. D., Thompson, R. P. H. and Williams, R. (1973) British Medical Journal, 1, 530.

Teather, D. (1974a) Bulletin of the Institute of Mathematics and its Applications, 10, 37.

Teather, D. (1974b) Journal of the Royal Statistical Society, Series A, 137, 231.

The Role of Utility in Decision-making

D. V. LINDLEY, M.A, Professor of Statistics, University College, London

In an earlier article (page 197) I discussed the contribution that a statistician might make toward the problem of diagnosis. It was argued that diagnosis is properly conducted by using the Bayes rule, and that effective diagnosis might be accomplished using relatively few carefully selected tests or symptoms. This article extends the discussion from diagnosis to include the action that is taken as a result of the diagnosis, one such action being the possibility of further tests, but I want to go beyond this and consider the choice of treatment, management and surgery that might be given to a patient, for diagnosis is only a stepping-stone on the path to action.

Let us begin with an example, similar to one already considered in that it concerns jaundice, involving surgery. A simplification of a situation that can arise is to suppose that all the tests have been performed and it is now a question of whether to resort to surgery or not. In further simplification let it be supposed that there are only two types of jaundice: hepatitis, which is treated without surgery, and stones in the biliary tree, which require surgery. The