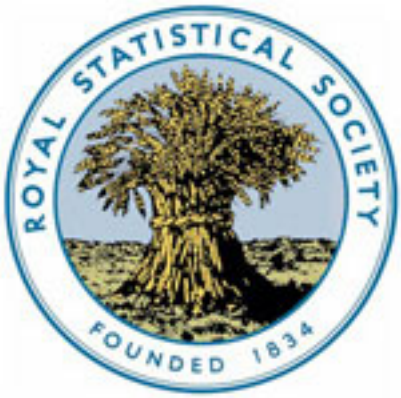

J. R. Statist. Soc. A (1981),
144, *Part* 2, *pp.* 145–175

Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients

By D. M. TITTERINGTON[1], G. D. MURRAY[1], L. S. MURRAY[2], D. J. SPIEGELHALTER[3], A. M. SKENE[3], J. D. F. HABBEMA[4] and G. J. GELPKE[5]

[1] *Department of Statistics, University of Glasgow;* [2] *Department of Neurosurgery, University of Glasgow;* [3] *Department of Mathematics, University of Nottingham;* [4] *Department of Public Health and Social Medicine, Erasmus University, Rotterdam; and* [5] *Department of Neurosurgery, Erasmus University, Rotterdam*

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday December 17th, 1980, the President, Professor D. R. COX, in the Chair]

SUMMARY

Several techniques for discriminant analysis are applied to a set of data from patients with severe head injuries, for the purpose of prognosis. The data are such that multidimensionality, continuous, binary and ordered categorical variables and missing data must be coped with. The various methods are compared using criteria of prognostic success and reliability. In general, performance varies more with choice of the set of predictor variables than with that of the discriminant rule.

Keywords: DISCRIMINANT ANALYSIS; MEDICAL DATA; HEAD INJURIES; MULTIVARIATE; ORDERED CATEGORIES; MISSING DATA; INDEPENDENCE MODEL; LANCASTER MODEL; LINEAR DISCRIMINATION; QUADRATIC DISCRIMINATION; LOGISTIC; LATENT CLASS; KERNEL; DENSITY ESTIMATION; POSTERIOR ODDS; SEPARATION CRITERIA; RELIABILITY CRITERIA; FEATURE SELECTION

1. INTRODUCTION

THE purpose of this paper is to discuss the application of several methods of discriminant analysis to a large complex data set. The complexity was caused by multidimensionality, the inclusion of data which could be regarded as being categorical or continuous, and the occurrence of missing data. The particular data set which we used was from the medical field, made up of patients with severe head injury, and discriminant analysis was used for prognosis (to predict the future outcome), as opposed to the more common application of assisting diagnosis. In statistical terms, however, the problems are technically identical. In both, data are used to estimate the probability that a given patient belongs to a certain class or category.

The medical literature already contains reports of the application of discriminant analysis based on an independence model for this data set (Jennett *et al.*, 1975, 1976, 1979; Teasdale *et al.*, 1979b; see also Becker, 1979, who criticized the use of such a model). This paper compares the performance of several discriminant analysis techniques, some of which have been developed since work began on this data set 10 years ago.

Before describing the data set in detail, it is convenient to introduce the notation and terminology of discriminant analysis (Duda and Hart, 1973; Aitchison and Dunsmore, 1975, Chapter 11; Lachenbruch, 1975).

(i) Individuals in the study are assumed to belong to one or other of k *outcome categories* $\pi_1, \ldots, \pi_k$.

(ii) Associated with these outcome categories there may be a set of *prior probabilities*, *arrival rates* or *relative incidences*, $p(\pi_1), \ldots, p(\pi_k)$, which sum to unity.

(iii) Information is available on each individual in the form of measurements on *feature variables* or *indicants*, making up a *feature vector*.

(iv) A *discriminant rule* is set up for assigning an individual to one of the outcome categories or for specifying appropriate odds for the different outcome categories given the feature vector, y, appropriate to that individual.

(v) A *training data set* D of n individuals whose outcome categories and feature vectors are known, represented as

$$D = \{(c_i, x_i),\ i = 1, \ldots, n\}.$$

The outcome category of individual i is denoted by c_i and the feature vector by x_i. D is used to construct the discriminant rule.

(vi) In comparative studies such as this we require a *test data set*, also of individuals whose outcome categories and feature vectors are known, for the purpose of evaluation of the different discriminant rules used. Often the training set and the test set are the same and less biased evaluation can be achieved provided cross-validatory assessment is used (Lachenbruch and Mickey, 1968) but in the sample considered here there are enough data to permit separate test and training sets.

In this paper, a discriminant rule will be a procedure for obtaining, for an individual having feature vector y, conditional probabilities of the form

$$\{p(\pi_i \mid y, D),\ i = 1, \ldots, k\},$$

which may or may not be used to assign that individual to a single outcome category.

In general this means that we are trying to estimate, using D, some model

$$\{p(\pi_i \mid y),\ i = 1, \ldots, k\} \tag{1}$$

and there are two possible approaches, called by Dawid (1976) the *diagnostic paradigm* and the *sampling paradigm*. In the former, direct models are proposed for (1), whereas the sampling paradigm approach exploits the fact that, by Bayes' Theorem,

$$p(\pi_i \mid y) \propto p(y \mid \pi_i)\, p(\pi_i), \quad i = 1, \ldots, k. \tag{2}$$

Both components of the right-hand side are modelled. Most effort has to go into the first factor which means that *density estimation*, parametric or otherwise, is of prime concern. This latter approach tends to give wider scope for different methods and indeed, of the techniques used in this study, only the logistic (see Section 3) represents the *diagnostic paradigm* principle.

The basic structure is well known, as is the variety of density estimation procedures. We now give empirical illustration of application of a few techniques to a medical problem in which the feature vector from an individual or patient is multivariate and possibly incomplete, in that certain measurements are missing.

2. History and Description of the Data

The data set is a series of 1000 patients with severe head injury collected prospectively by neurosurgeons between 1968 and 1976. This head injury study was initiated in the Institute of Neurological Sciences, Glasgow. After 4 years two Netherlands centres (Rotterdam and Groningen) joined the study, and later data came also from Los Angeles.

The original purpose of the head injury study was to investigate the feasibility of predicting the degree of recovery which individual patients would attain, using data collected shortly after injury. Severely head injured patients require intensive and expensive treatment; even with such care, almost half of them die and some survivors remain seriously disabled for life. Clinicians are concerned to recognize which patients have potential for recovery, so as to concentrate their endeavours on them. Facilities for this kind of treatment are limited even in the most affluent countries, and the rational allocation of resources requires the identification of which patients to treat, or to continue to treat if they do not respond favourably.

The details of data collection have been reported in the clinical literature (Jennett *et al.*, 1979). Patients entered the study on the basis of a minimum degree of brain damage; all were in

coma for at least 6 hours, where coma was rigorously defined. Outcome was categorized according to the Glasgow Outcome Scale (Jennett and Bond, 1975), but the five categories described therein were reduced to three for the purpose of prediction in the present paper. These were

(a) dead or vegetative;
(b) severe disability;
(c) moderate disability or good recovery.

Death usually occurs within a month of injury, but recovery in survivors can be a slow process. However, analysis of the survivors suggests that few improve sufficiently to move from category (b) to category (c) after six months. We have so far limited our work to estimating the probability of attaining one or other of the three outcome categories six months after injury; comparisons of such predictions with actual recovery are based on the patient's actual state at six months, even if his condition changes at a later date.

Apart from the patient's age, the most important predictive factors are the various indicators of the degree of brain damage, as reflected in brain dysfunction. The various factors considered in this paper are listed in Table 1, and these include measures of the depth of coma, state of pupils, motor responses in all four limbs and eye movements. The clinicians concerned with data collection spent much time and effort in evolving these measures, which included conducting observer-error studies to discover what aspects of the patient's state could be reliably recorded by different clinicians in different countries (Teasdale *et al.*, 1978; van den Berge *et al.*, 1979). One particularly powerful indicator is the EMV score (see Table 1) which has become known in the medical literature as the Glasgow Coma Scale (Teasdale and Jennett, 1974; Teasdale *et al.*, 1979a). It can be seen that the variables are all categorical and are either binary or ordered. This means, among other things, that methods based on continuous data might be considered as possible, albeit unsatisfactory, alternatives to categorical data techniques.

TABLE 1
The feature variables used in the comparisons

Variable	*Description*
Age	Age, grouped into decades 0–9, 10–19, ... , 60–69, 70+
E score	Eye opening in response to stimulation, graded 1 (nil) to 4 (normal), but grouped as 1 and 2–4 for these comparisons
M score	Motor response of best limb in response to stimulation, graded 1 (nil) to 6 (normal)
V score	Verbal response to stimulation, graded 1 (nil) to 5 (normal), but grouped as 1 and 2–5 for these comparisons
EMV score	The sum of the raw E, M and V scores, in the range 3 to 15, but grouped as 3, 4, 5, 6, 7, 8, 9–15 for these comparisons
MRP	Motor response pattern, an overall summary of the motor responses in all four limbs, graded 1 (nil) to 7 (normal)
Change	Change in neurological function over the first 24 hours, graded 1 (deteriorating), 2 (static) or 3 (improving)
Pupils	Pupil reaction to light, graded 1 (non-reacting) or 2 (reacting)
SEM	Spontaneous eye movements, graded 1 (nil) to 4 (normal).
OCS	Oculocephalics, graded 1 (nil) to 4 (normal)
OVS	Oculovestibulars, graded 1 (nil) to 4 (normal)
Eye indicant	A summary of SEM, OCS and OVS, graded 1 (bad), 2 (impaired) or 3 (good)

Indicators of brain dysfunction can vary considerably during the few days after injury. Measurements were therefore taken frequently, and for each indicant the best and worst states during each of a number of successive time periods were recorded. The predictions in the present study are based on the best state during the first 24 hours after onset of coma.

The different subsets of variables were chosen to compare how well the various methods were able to exploit the information in subsets of different sizes, and to see how the methods

reacted to the degree of dependence among the variables and to the proportion of missing data. The four subsets used are given in Table 2. Set I consists of four weakly dependent variables with appreciable missing data, whereas set II consists of four highly dependent variables with little missing data. Set III is an extension of set I, and set IV is obtained from set III by expanding the "created indicants" *EMV score* and *Eye indicant*. There is therefore high dependence and appreciable missing data with set IV.

TABLE 2
The subsets of feature variables used in the comparisons

Variable set	*Variables*
I	Age, EMV score, Change, Eye indicant
II	Age, E score, M score, V score
III	Age, EMV score, MRP, Change, Pupils, Eye indicant
IV	Age, E score, M score, V score, MRP, Change, Pupils, SEM, OCS, OVS

Since comparisons will be made on specific sets of predictor variables, which were not chosen according to formal statistical criteria, no discussion will be given on methods of variable selection although this is a very important aspect of applied discriminant analysis.

The dimensionalities of the four data sets are 4, 4, 6 and 10 respectively. The data set of 1000 cases is unusually large for a medical application of discriminant analysis and, as remarked earlier, this allowed us to use separate training and test sets. The set of 1000 cases was split randomly into two groups of 500 for this purpose which gave the distribution of outcomes tabulated here.

	Frequencies	
	Training set	*Test set*
Dead/vegetative	259	250
Severely disabled	52	48
Moderate or good recovery	189	202
Total	500	500

The proportions among the training cases were used as plug-in estimates of the prior probabilities or relative incidences $\{p(\pi_1), p(\pi_2), p(\pi_3)\}$, in (2) for instance.

Table 3 gives an indication of the amount of missing data on the different feature variables by showing the number of times that each variable was observed jointly with each other variable. It can be seen that the bulk of the gaps occurs on the variables *Change*, *SEM*, *OCS*, *OVS* and *Eye indicant*. Indeed the *Eye indicant* was devised to overcome the problem of the high proportion of missing data on *SEM*, *OCS* and *OVS*. It summarizes these three variables and can be obtained whenever at least one of them is recorded.

3. THE REPERTOIRE OF STATISTICAL TECHNIQUES USED

The statistical methods used can be brought together under the following general headings.

(i) Independence-based models for unordered categorical data, allowing for a single overall association parameter.
(ii) Lancaster first-order interaction models for unordered categorical data.
(iii) Latent class models.
(iv) Kernel-based procedures for categorical data.
(v) "Linear and quadratic" discrimination based on normality assumptions.
(vi) Linear logistic discrimination.

TABLE 3

The number of times each two variables are observed jointly among the 500 training cases

	Age	*E score*	*M score*	*V score*	*EMV score*	*MRP*	*Change*	*Pupils*	*SEM*	*OCS*	*OVS*	*Eye indicant*
Age	500											
E score	488	488										
M score	495	488	495									
V score	478	472	478	478								
EMV score	472	472	472	472	472							
MRP	479	471	478	462	456	479						
Change	368	359	366	353	347	368	368					
Pupils	487	479	484	467	463	471	359	487				
SEM	326	321	324	313	310	319	272	322	326			
OCS	328	323	326	316	313	318	292	322	273	328		
OVS	230	229	229	224	224	222	207	228	190	225	230	
Eye indicant	390	384	386	375	372	378	321	384	326	328	230	390

As remarked in Section 1, all but (vi) boil down to density estimation in one form or another. We now give brief notes about the different techniques, with references and specific remarks about the facility for dealing with missing data.

3.1. *Independence Models*

In these, perhaps the simplest models, the density estimates took the form, for y complete,

$$p(y \mid \pi_i, D) \propto \left\{ \prod_{r=1}^{d} \frac{n_i(y_r) + 1/C_r}{N_i(r) + 1} \right\}^B, \tag{3}$$

where

d is the number of variables,

y_r denotes the rth component of y,

$n_i(y_r)$ is the number of patients in the training set in outcome category π_i with score y_r on variable r,

C_r is the number of categories in variable r,

$N_i(r)$ is the number of patients in the training set in outcome category π_i with variable r not missing,

and B denotes an overall association factor representing the "proportion of non-redundant information" in the variables; see Hilden and Bjerregaard (1976). A small amount of smoothing is therefore imposed. In the most familiar case of $B = 1$, $p(y \mid \pi_i, D)$ is given as the product of the estimates of the marginal probabilities. This is essentially the model originally used with these data (Jennett *et al.*, 1976; Teasdale *et al.*, 1979b). When y_r is missing, the appropriate factor on the right-hand side of (3) is replaced by unity.

Three choices of B were used, 1·0, 0·8 and 0·5, and these values characterize the methods which we shall call INDEP1, INDEP2 and INDEP3.

3.2. *Lancaster Models*

The structure of Lancaster models is described for example in Zentgraf (1975) or Goldstein and Dillon (1978). With a factorial-type structure, which for binary variables simplifies to the

Bahadur expansion (Moore, 1973), they permit a full range of models from basic independence models to the full multinomial. A natural version to try is that containing only first-order interactions. Difficulties arise when fitting these models. In particular (Moore, 1973; Trampisch, 1976),

(i) some of the probability estimates may be negative;

(ii) all cells in all two-dimensional cross-tabulations are used, and have consequently to be established from the training samples. Thus further grouping of the variables would be desirable to reduce the number of parameters in the model. However, this was not done in this paper, in order to have better comparability with the other approaches.

Problem (i) was combatted by reverting to the independence model whenever it occurred.

Missing data treatment is straightforward, as in the independence models, and the probability estimates are based on the non-missing cases. The two-dimensional marginal estimates were taken as

$$p(y_r, y_s \mid \pi_i, D_{rs}) = \frac{n_i(y_r, y_s) + \{1/(C_r C_s)\}}{N_i(r, s) + 1} \quad \text{for each } i, r, s,$$

where D_{rs}, $n_i(y_r, y_s)$ and $N_i(r, s)$ are analogous to D_r, $n_i(y_r)$ and $N_i(r)$ in the independence case.

When the independence model was used to avoid problem (i), three choices of B were again made, 1·0, 0·8 and 0·5, and these values characterize the methods which we shall call LANC1, LANC2 and LANC3.

3.3. *Latent Class Models*

In latent class analysis, mixture models are assumed for the density functions being estimated. Thus, for each π_i, it is assumed that

$$p(y \mid \pi_i) = \sum_{j=1}^{L} w_{ij} p_j(y),$$

where L denotes the number of terms in the mixture, i.e. the number of latent classes and $\{w_{ij}\}$ are, for each i, a set of mixing weights; see Fielding (1977). If, for fixed L, maximum likelihood estimation is used, the *EM* algorithm provides a convenient, if potentially slow, iterative technique and, if it is assumed that the mixed densities $\{p_j(\cdot), j = 1, ..., L\}$ represent independence models, the M-step (Dempster *et al.*, 1977) is easy, whether or not there is missing data; see Skene (1978). This conditional independence structure was used here and for each data set two sets of results are quoted, LATCL1 and LATCL2. These correspond to the two best "consecutive" numbers of latent classes.

3.4. *Kernel-based Procedures*

If the density function $p(\cdot \mid \pi_i)$ is estimated by the non-parametric kernel method then

$$p(y \mid \pi_i, D) = \frac{1}{N_i} \sum_{j=1}^{N_i} K(y \mid x_{ij}, \lambda),$$

where

N_i = number of patients in the training set in category π_i,
$(x_{i1}, ..., x_{iN_i})$ denote their feature vectors,
$K(\cdot \mid x, \lambda)$ is a probability density over the sample space of y,

and λ, possibly a vector, describes the degree of smoothing of the relative frequencies achieved by this method of estimation.

As in Aitchison and Aitken (1976), we take the kernel function to be of factorized form

$$K(y \mid x_{ij}, \lambda) = \prod_{r=1}^{d} K_r(y_r \mid x_{ij}^{(r)}, \lambda_r)$$

where d denotes, as before, the dimensionality of y, λ_r is a scalar smoothing parameter and $x_{ij}^{(r)}$ is the rth component of x_{ij}.

The basic kernel used when there is no missing data takes the form

$$\begin{aligned} K_r(y_r \mid x^{(r)}, \lambda_r) &= \lambda_r && \text{if } y_r = x^{(r)} \\ &= c(y_r, x^{(r)})(1-\lambda_r) && \text{if } y_r \neq x^{(r)} \end{aligned} \tag{4}$$

For a nominal (unordered-categories) variable of q categories the natural choice is $c(y_r, x^{(r)}) = (q-1)^{-1}$. if $y_r \neq x^{(r)}$.

Titterington (1980) suggests formulae for $\{c(y_r, x^{(r)})\}$ which satisfy properties demanded by Aitchison and Aitken (1976) for kernels for ordered-category variables. An advantage of kernels of the form (4) is that optimal smoothing parameters $\{\lambda_r\}$ according to a minimum mean squared error criterion can be calculated explicitly on a marginal basis: see Titterington (1980) where a more complicated multivariate pseudo-Bayesian updating technique is also described. Two techniques were used for dealing with missing data. These involved using the kernels (for ordered or unordered variables) described in Murray and Titterington (1978) or regarding missing as an extra category for each variable. The latter means that the ordered nature of the categories is destroyed.

The following kernel methods were used.

KERUN1: Kernel of Murray and Titterington (1978), unordered categories, smoothing parameters chosen marginally.

KERUN2: As KERUN1 but smoothing parameters chosen by multivariate pseudo-Bayes.

KERORD1, KERORD2: As KERUN1, KERUN2 but assuming ordered categories.

KEREX1, KEREX2: Marginal and multivariate choices of smoothing parameters treating "missing" as an extra category.

KEREX3: "Missing" treated as an extra category, but a single smoothing parameter, u, for all dimensions, chosen on the lines of Habbema *et al.* (1978a). Take $0<u<1$ and, for a q category variable, $\lambda = 1-(q-1)/uq$.

For this data set a suitable u was chosen rather subjectively to perform as well as possible for the criteria described in Section 4.

3.5. *Normal-based Methods*

These methods follow the common if dubious practice of assuming multivariate normality and estimating mean vectors and covariance matrices by maximum likelihood. They lead to linear or quadratic discriminant rules depending on whether or not the covariance matrices are assumed equal. They are justified a priori here only on the grounds that their performance would be viewed with interest and that, since the categories in the non-binary variables are ordered, the feature variables can be given meaningful scores. Results from three methods are reported.

NORLIN1: Covariance matrices are assumed equal and sample means from available data are substituted for missing values.

NORLIN2: As NORLIN1, but with proper maximum likelihood treatment for missing data via the *EM* algorithm.

NORQUAD: As for NORLIN2, but without the assumption of equal covariance matrices. With all three methods the incomplete test cases were classified on the basis of the relevant marginal distributions.

3.6. *Linear Logistic Method*

As remarked earlier, this is the only method in which models are set up directly for $\{p(\pi_i \mid y), i = 1,...,k\}$. The models take the parametric form

$$p(\pi_i \mid y)/p(\pi_k \mid y) = \exp(\alpha_i + \beta_i^T y), \quad i = 1,...,k-1,$$

where $\{\alpha_i\}$ and $\{\beta_i\}$ are to be estimated. The technicalities are described by Anderson (1972), where the links between this and some of the density estimation techniques are made. In the absence of more sophisticated ideas, missing values were replaced by group means in the training cases and grand means in the test cases, giving the method to be called LINLOG.

3.7. *Other Methods not used*

The above does not exhaust the range of procedures that are available. As well as quadratic logistic discrimination (Anderson, 1975) there are other density estimation techniques. For categorical data some are given in Goldstein and Dillon (1978) and a few are listed below.

(i) *Loglinear models*

It is possible to construct models of this class to deal with ordered categories (Haberman, 1974) and to cope with missing data (Chen and Fienberg, 1974). The calculations are however likely to be very heavy and this method was not used. Perhaps this can be excused in view of the inferior performance of the Lancaster models relative to the independence models.

(ii) *Non-parametric methods based on orthogonal series* (Goldstein and Dillon, 1978).

These have not as yet been adapted to cope with missing values or high dimensional data, and so they are not included.

(iii) *Continuous kernel methods*

If we can assume normality as in Section 3.5, there seems no reason why we should not try kernel methods based on continuity of the data. This could be done for at least some of the variables, such as *Age*, *M score*, *V score* and *EMV score*, thus regarding the data as of *mixed* type. Kernel methods do exist for such data (Aitchison and Aitken, 1976, Habbema *et al.*, 1978a) and research into useful ways of coping with missing data is under way but still embryonic.

(iv) *Location model*

This is a parametric approach to mixed data (Krzanowski, 1975) and, again, dealing with missing data is, at best, difficult, in conjunction with parameter estimation.

(v) *Others*

A criticism of the independence and latent class models is that the basic models, such as the conditional (on the latent class) densities do not reflect the ordered nature of the variables. This could be rectified either by assuming continuity (*latent structure*, Fielding, 1977) or by setting up discrete distributions appropriate to ordered categories; see McCullagh (1978).

In the analysis of all models the mechanism by which a measurement became "missing" was ignored. In fact, the data were implicitly assumed to be "missing at random" within each prognostic category. Here, "at random" is according to the following definition given by Little (1979) as being equivalent to that given by Rubin (1976). If we have n d-variate observations, denote the $(n \times d)$ data matrix by $X = \{x_{ij}\}$ and define the $(n \times d)$ random matrix $R = \{r_{ij}\}$ by $r_{ij} = 0$ or 1 according to whether x_{ij} is missing or observed. Then any missing values are "missing at random" if the distribution function of the conditional distribution of R given X is functionally independent of the missing values. In particular, the probability that a value x_{ij} is observed must not depend on the value x_{ij} (thus excluding truncation from the definition), although it might depend on the values of an observed variable x_{ik}. Rubin (1976) gives this as the weakest definition of missing at random which allows us to ignore the mechanism generating the missing values. It is fair to say that this is unrealistic in this application. It is, however, difficult to avoid this assumption by convenient realistic modelling and, in the case of the Head Injury Study, the incorporation of the incomplete data using the above techniques does add useful information. This can be demonstrated by carrying out similar analyses on complete

feature vectors only, and for the categorical models, carrying out similar analyses where "missing" is taken to be an additional category (Murray, 1979).

4. Criteria for Evaluation of the Discriminant Rules

Two quite separate aspects of performance must be considered when trying to assess a discriminant rule. The more important aspect is how well the groups corresponding to the outcome categories are separated but we should also like to know whether any probabilities assigned to each group for a given feature vector are realistic. For example, if there are two outcome categories, a rule which invariably assigns a probability of 0·51 to the correct group gives perfect separation but unrealistic probabilities. At the other extreme, a rule which for every case just assigns the arrival rate probabilities gives, in a sense, accurate probabilities but is not useful for separation.

Habbema *et al.* (1978b, 1980), Habbema and Hilden (1980) and Hilden *et al.* (1978a, b) give a wide discussion of these points and present a large number of measures of efficiency for a discrimination procedure. We shall give results for only a few.

4.1. *Measures of Separation*

Error rate

This most commonly used measure of separation is the proportion of cases in the test set allocated to an incorrect class. It is very insensitive because it takes no account of the relative seriousness of different errors, or of "near misses", although it does have respectable decision theoretic foundations.

Average logarithmic score

For a patient whose true category is, say, π_1, his logarithmic score is

$$-\log p(\pi_1 \mid y, D) = -\log p_1, \quad \text{say.}$$

Average quadratic score

For the above patient the quadratic score is

$$(1-p_1)^2 + \sum_{i=2}^{k} p_i^2.$$

Both these scores take account of the actual probabilities assigned and the different discriminant rules are assessed using the average scores over the test set. The logarithmic score is sensitive to small values of the probability, p_1, although this can be dampened by setting a lower limit to p_1 in the formula.

Average loss

A criticism of the logarithmic and quadratic scores is that they still do not reflect the relative seriousness of the different types of error. This is particularly relevant to the Head Injury Study in which it is clearly a far worse error to predict that a patient will die when he actually recovers than to predict recovery when, in fact, he dies. The neurosurgeons involved in this study were asked to construct a loss matrix which would reflect, albeit in a crude and arbitrary way, how they would judge the relative losses associated with the different errors. This matrix, developed in 1975 by Dr R. Knill-Jones, was adopted by us to assign patients to that outcome category which was associated with minimum expected loss; see Table 4.

The average loss measure has the drawback of being very sensitive to minor modifications to the discriminant rule. In fact research is currently under way where the rather unrealistic

TABLE 4
The loss matrix used for head injury predictions

		Predicted outcome		
		D/V	*SEV*	*M/G*
Actual outcome	*D/V*	0	10	75
	SEV	10	0	90
	M/G	750	100	0

requirements of the fixed loss values are relaxed, giving rise to so-called stochastic loss measures which do not suffer from this oversensitivity; see Habbema and Hilden (1980).

For good performance from the point of view of separation we should desire all these measures to be close to zero, and a useful benchmark for comparison is the performance obtained by assigning the prior probabilities to each case. This discriminant rule would score 0·500, 0·939, 0·579 and 45·4, respectively, on the four measures.

4.2. *Measures of Reliability*

Reliability, or trustworthiness, of the probabilities resulting from a discriminant rule is in general assessed by comparing the degree of separation which is actually obtained by the rule with the degree of separation which should be expected if the probabilities given by the rule were perfectly reliable. When the actual separation equals the expected one, apart from random fluctuations, the rule may be called reliable. If not so, the discriminant rule is overconfident when the actual separation is less then the expected one, and underconfident when the actual separation is greater than the expected one.

The most thorough way of checking the reliability of the probability assessments obtained would be to obtain a large number of patients for each possible feature vector and see what proportions do actually attain the different levels of recovery. These proportions should be close to the predicted probabilities afforded by the discriminant rule. This procedure is out of the question because of the very large number of cases involved, but what can be done is to group cases with similar predictions and see what proportions of these groups attain the various outcome levels.

A set of three probabilities can be represented, in terms of areal coordinates, as a point in an equilateral triangle. The following method of constructing a reliability measure is suggested.

Divide the triangle into suitable smaller ones $\{T_1, \ldots, T_m\}$. Let $\{p_{ij}: j = 1, 2, 3\}$ be the probability triple at the centroid of T_i. Suppose $n_{i.}$ patients in the test set have predicted probabilities falling in T_i and suppose that, for each j, n_{ij} of them came from outcome category j. For T_i evaluate

$$D_i = \sum_{j=1}^{3} (n_{ij}/n_{i.}) \log (n_{ij}/n_{i.}\, p_{ij}), \quad i = 1, \ldots, m.$$

These are weighted to obtain an overall measure

$$DT = \sum_{i=1}^{m} n_{i.}\, D_i/n,$$

where n is the total number in the test set, in our case 500.

This was evaluated using 25 small equilateral triangles. Against this rather large choice of 25 can be weighted the criticism that there will be, on average, only two "severe" cases in each small triangle.

For the independence and Lancaster models another approach was used in assessing the reliability. There, the expected quadratic score was computed under the hypothesis of perfect

reliability, and deviations from this expected score gave the measure *DQ*; see Hilden *et al.* (1978a). Both measures are given for the latent class and logistic methods.

For both these reliability measures, high reliability is represented by scores close to zero.

5. Results and Discussion for Head Injury Study

The results are given in full in Tables 5–8. With 19 methods, 4 sets of variables and 6 performance criteria there are obviously many comparisons which can be made and we shall break down the discussion into three sections. Firstly we shall make comparisons *within* groups of similar methods, then *among* groups of similar methods and finally, we shall make an overall comparison among the sets of variables. The reliability measures will be compared only within groups of similar methods.

TABLE 5

The results for variable set I: Age, EMV score, Change, Eye indicant

	Measures of separation				*Measures of reliability*	
Method	*Error rate*	*Average logarithmic score*	*Average quadratic score*	*Average loss*	*DT*	*DQ*
INDEP1	0·278	0·685	0·377	23·3		0·010
INDEP2	0·268	0·681	0·379	22·8		−0·023
INDEP3	0·268	0·708	0·400	23·2		−0·068
LANC1	0·292	0·737	0·397	27·3		0·031
LANC2	0·294	0·735	0·398	26·3		0·021
LANC3	0·296	0·742	0·404	26·6		0·008
LATCL1	0·264	0·719	0·390	25·7	0·037	−0·012
LATCL2	0·290	0·752	0·409	25·2	0·028	−0·009
KERUN1	0·316	0·934	0·467	32·5	0·075	
KERUN2	0·308	0·925	0·449	30·1	0·078	
KERORD1	0·292	0·874	0·443	31·6	0·078	
KERORD2	0·302	0·900	0·430	28·8	0·075	
KEREX1	0·320	0·889	0·453	32·4	0·076	
KEREX2	0·328	1·037	0·477	32·4	0·102	
KEREX3	0·282	0·800	0·420	24·4	0·042	
NORLIN1	0·286	0·707	0·396	25·4	0·054	
NORLIN2	0·284	0·702	0·396	25·9	0·040	
NORQUAD	0·294	0·779	0·404	34·5	0·040	
LINLOG	0·290	0·721	0·400	26·4	0·045	−0·010

5.1. *Comparisons within Groups of Similar Methods*

We shall consider first the discrete parametric models, namely the independence, Lancaster and latent class models. For variable set I the simple independence model (INDEP1) performs well, with INDEP2 giving similar results. Those for INDEP3 are poorer, as were those for the latent class models. There is little to choose among the Lancaster models, whose performance is inferior to the simpler independence models.

The variables in set II are highly dependent, yet the independence model, INDEP1, still does well. INDEP2 gives a slight improvement but, perhaps surprisingly, INDEP3 is not better than INDEP2. Again there is little to choose between the three Lancaster models, but their performance is now superior to the independence models in terms of error rate, average loss and reliability. It is only for the average logarithmic score that the independence models better the

TABLE 6
The results for variable set II: Age, E score, M score, V score

	Measures of separation				*Measures of reliability*	
Method	*Error rate*	*Average logarithmic score*	*Average quadratic score*	*Average loss*	*DT*	*DQ*
INDEP1	0·338	0·775	0·438	29·0		0·055
INDEP2	0·340	0·762	0·436	29·5		0·018
INDEP3	0·338	0·771	0·445	29·0		−0·037
LANC1	0·298	0·808	0·435	26·7		0·039
LANC2	0·298	0·809	0·437	26·9		0·029
LANC3	0·296	0·818	0·445	25·6		0·018
LATCL1	0·328	0·819	0·447	31·3	0·032	−0·007
LATCL2	0·310	0·822	0·446	30·2	0·037	−0·011
KERUN1	0·364	0·924	0·481	33·9	0·072	
KERUN2	0·328	0·872	0·463	24·9	0·071	
KERORD1	0·352	0·905	0·471	32·5	0·055	
KERORD2	0·332	0·856	0·454	26·9	0·068	
KEREX1	0·334	0·953	0·491	35·2	0·074	
KEREX2	0·326	0·903	0·475	29·4	0·063	
KEREX3	0·340	0·852	0·466	25·4	0·058	
NORLIN1	0·316	0·760	0·433	26·6	0·040	
NORLIN2	0·306	0·757	0·431	26·2	0·035	
NORQUAD	0·304	0·884	0·450	34·9	0·052	
LINLOG	0·314	0·764	0·436	25·7	0·031	−0·010

Lancaster ones. The latent class results are generally poor, except that, as in variable set I, the reliability scores are good.

For variable set III there is little to choose among the independence models, and again there is little to choose among the Lancaster models. However the Lancaster results are consistently poorer than the independence ones and the latent class results display a now familiar pattern.

Variable set IV is perhaps the most interesting, because the differences in results are quite marked. INDEP3 clearly betters INDEP2, which in turn betters INDEP1, and we see the same pattern for the three Lancaster models, with LANC3 bettering LANC2, which in turn betters LANC1. We also see that INDEP3 is superior to LANC3, and that for this variable set alone the latent class models perform well.

The comparisons among the discrete kernel models are quite clear cut with KEREX3 standing out as being easily the best. Although the methods based on marginally chosen smoothing parameters are to be favoured on computational grounds they give density estimates which are too sharp from a multivariate point of view. For this data set the pseudo-Bayesian method too has not produced enough smoothing. This has however been achieved much more successfully with KEREX3, although by a subjective method of choice.

The final set of models comprises the continuous parametric models, NORLIN1, NORLIN2 and NORQUAD. We shall also include LINLOG with this group, although strictly speaking it is not restricted in application to continuous data. With these models we see a clear pattern over the four variable sets with the quadratic method always performing poorly. With the linear method it is always preferable to use the *EM* algorithm for parameter estimation. The differences with and without the *EM* algorithm (which admittedly requires considerable computation) are small for the first three variable sets, but marked for set IV. The results for the linear logistic method are comparable with those of NORLIN2 for variable sets I–III, but its performance drops off with

TABLE 7
The results for variable set III: Age, EMV score, MRP, Change, Pupils, Eye indicant

	Measures of separation				*Measures of reliability*	
Method	*Error rate*	*Average logarithmic score*	*Average quadratic score*	*Average loss*	*DT*	*DQ*
INDEP1	0·248	0·686	0·364	21·9		0·073
INDEP2	0·246	0·656	0·358	21·9		0·030
INDEP3	0·232	0·652	0·362	21·3		−0·046
LANC1	0·254	0·738	0·382	22·9		0·057
LANC2	0·256	0·728	0·378	22·9		0·040
LANC3	0·244	0·727	0·376	22·9		0·013
LATCL1	0·298	0·726	0·412	25·2	0·019	−0·002
LATCL2	0·262	0·718	0·372	26·9	0·026	−0·003
KERUN1	0·332	1·103	0·500	33·8	0·150	
KERUN2	0·338	1·267	0·537	35·9	0·190	
KERORD1	0·328	1·030	0·482	31·2	0·158	
KERORD2	0·316	1·270	0·514	30·7	0·177	
KEREX1	0·310	1·013	0·467	36·3	0·116	
KEREX2	0·344	1·412	0·548	36·6	0·197	
KEREX3	0·278	0·769	0·395	22·5	0·046	
NORLIN1	0·256	0·665	0·368	22·8	0·056	
NORLIN2	0·258	0·661	0·367	21·7	0·048	
NORQUAD	0·276	0·907	0·411	30·5	0·080	
LINLOG	0·272	0·676	0·370	24·4	0·046	−0·007

set IV. However, when one considers how crudely the missing data were handled with LINLOG this can be regarded as a good result for the linear logistic method, and further work on the treatment of missing values in logistic models might well be rewarding.

5.2. *Comparisons among Groups of Similar Methods*

Table 9 gives a summary of Tables 5–8. In it we have chosen, for each variable set, the best method from each of the three groups of methods discussed in Section 5.1. In cases where it was not obvious which method was best, they were ordered by their average quadratic scores.

It can be seen that the kernel methods have disappointing logarithmic and quadratic scores and it is only for variable set II that they even approach the other methods. This is perhaps quite simply because we are trying to be too ambitious in using a discrete kernel approach in this problem. For variable set IV, for instance, we are dealing with a contingency table with over 500 000 cells, while imposing very little structure on the cell probabilities. The sparseness of the data can be illustrated by the fact that when, as in KEREX1, KEREX2 and KEREX3, "missing" is regarded as an extra category, only 37 out of the test set of 500 have feature vectors matching one of the training set. This means that, because of the type of smoothing produced by the kernel method, very small probability estimates are obtained for $\{p(y \mid \pi_i, D), 1, \ldots, k\}$, leading to unreliable $\{p(\pi_i \mid y, D), i = 1, \ldots, k\}$. If even a small extra smoothing factor is introduced, for example adding 0·01 to all cell frequencies, results are obtained comparable to simple multinomial smoothing and therefore, for this data set, to assignment simply by the prior probabilities. In some of the methods this insufficient smoothing may be due to the fact that marginally-chosen smoothing parameters are used in a multivariate context.

The results for the linear method are remarkably similar to those achieved with the discrete models. For sets I and III the discrete models have the edge and for sets II and IV the dominance

TABLE 8

The results for variable set IV: Age, E score, M score, V score, MRP, Change, Pupils, SEM, OCS, OVS

	Measures of separation				*Measures of reliability*	
Method	*Error rate*	*Average logarithmic score*	*Average quadratic score*	*Average loss*	*DT*	*DQ*
INDEP1	0·272	0·839	0·399	28·1		0·166
INDEP2	0·264	0·757	0·385	25·4		0·121
INDEP3	0·264	0·673	0·368	23·9		0·029
LANC1	0·286	0·829	0·410	26·7		0·146
LANC2	0·286	0·800	0·403	25·3		0·128
LANC3	0·280	0·768	0·395	22·9		0·097
LATCL1	0·282	0·726	0·396	27·4	0·022	−0·015
LATCL2	0·244	0·709	0·381	25·3	0·032	−0·011
KERUN1	0·350	1·417	0·566	39·2	0·240	
KERUN2	0·390	1·932	0·645	44·3	0·304	
KERORD1	0·340	1·414	0·543	31·5	0·219	
KERORD2	0·374	1·923	0·628	39·3	0·302	
KEREX1	0·388	1·645	0·634	36·0	0·303	
KEREX2	0·398	2·143	0·652	45·3	0·301	
KEREX3	0·298	0·806	0·412	22·7	0·049	
NORLIN1	0·270	0·804	0·404	26·3	0·052	
NORLIN2	0·250	0·663	0·361	23·2	0·053	
NORQUAD	0·274	0·947	0·424	27·5	0·103	
LINLOG	0·286	0·772	0·412	27·8	0·053	−0·035

TABLE 9

An overall summary of the results

Method	*Error rate*	*Average logarithmic score*	*Average quadratic score*	*Average loss*	*Variable set*
INDEP1	0·278	0·685	0·377	23·3	
KEREX3	0·282	0·800	0·420	24·4	I
NORLIN2	0·284	0·702	0·396	25·9	
LANC1	0·298	0·808	0·435	26·7	
KERORD2	0·332	0·856	0·454	26·9	II
NORLIN2	0·306	0·757	0·431	26·2	
INDEP2	0·246	0·656	0·358	21·9	
KEREX3	0·278	0·769	0·395	22·5	III
NORLIN2	0·258	0·661	0·367	21·7	
INDEP3	0·264	0·673	0·368	23·9	
KEREX3	0·298	0·806	0·412	22·7	IV
NORLIN2	0·250	0·663	0·361	23·2	

is reversed, although the differences are so small as to be of little importance in practice. However, it should be said that with the linear method we have a single method which performs very well for each variable set, whereas with the discrete models the choice of the appropriate model can be critical.

5.3. *Comparisons among the Variable Sets*

So far we have seen how the various methods perform relative to each other for each data set, but there still remains the important question of which variable set gives the best overall performance. Indeed we see, from Table 9, that the variation in performance among the methods tends to be smaller than that among the variable sets. The best overall set of results is obtained with method INDEP2 on variable set III, and it is interesting that, although set IV contains strictly more information, the discrete models cannot exploit this. In contrast, we see that, with the linear method, the performance improves as we go from set I to set III to set IV, although the results for sets III and IV are very similar. This again emphasizes the robustness of the linear approach, which appears to make sensible use of the available information, whereas the discrete parametric models have to be matched carefully to the variables being used.

This suggests the general feeling that the linear approach, with the *EM* algorithm, is preferable for a quick, uninformed analysis, but, with more effort and using prior background information, as in our illustration from the neurosurgeons, to combine groups of highly dependent variables into single "created indicants", it should be possible to achieve similar, if not better, performances using a much simpler independence model. This certainly was true for this problem.

This simplicity of the independence model relative to the linear rule could well make the difference between an acceptable and an unacceptable system in the clinical environment and this is particularly so if predictions have to be based on incomplete feature vectors. This is achieved easily with the simple independence model, but to implement it with the linear model it is necessary to invert the relevant part of the covariance matrix to evaluate the coefficients.

At the hospitals where the Head Injury Study is based there are excellent computing facilities, and even the heavy computations involved in predicting a new patient using a kernel technique are quite feasible. However, these facilities are exceptional, and to assess the feasibility of making predictions in a more realistic setting we conducted a trial in Glasgow where the junior doctors responsible for the data collection made predictions themselves based on an independence model using a programmable calculator. This proved to be quite practicable, and was readily accepted, but until now it has been the policy not to give the predictions to the doctors treating the patients. This was so that the predictions would not affect the management of the patients in any way, thus avoiding any accusations of making self-fulfilling prophecies. One other interesting point which emerged from this trial was that once the doctors had had some experience of making predictions their own prognositic accuracy increased markedly.

In Rotterdam, a set of tables is under construction where the probabilistic prognosis can be read off for all kinds of symptom combinations.

Some further comments should be made about the success of discriminant rules in general when applied to the Head Injury data.

One of the more important aims is to identify, at an early stage, the small proportion of persons who will be severely disabled, and thus in need of continuing medical and social care. It must be admitted that no method has been very successful in this context. This is undoubtedly due to the fact that, geometrically, the severe group is overlapped by those in the other two outcome categories whose members dominate the total sample numerically. This *ordering* of the outcome categories should perhaps be acknowledged and developments of the methods of McCullagh (1978) would be useful here. It should also be stated that, by considering data from the patients at later epochs than 24 hours from onset of coma, somewhat better separation of the outcome categories can be achieved; see Braakman *et al.* (1980) and Teasdale *et al.* (1979b).

6. General Conclusions

Some general points have emerged from this particular case study which we consider have wide import to complex discrimination problems in general. They are as follows.

(i) The robustness of independence models, allied to the ease with which incomplete data are dealt with.

(ii) The robustness of the normal based linear discriminant rule, in particular the favour it receives over the quadratic rule (cf. Aitchison *et al.*, 1977), and its facility for recognizing ordered categories.

In view of (i) and (ii) there is the strong suggestion that latent structure models (Fielding, 1977) should be tried. Skene (1978) describes such an application to a restricted version of this data-set.

(iii) The lack of success of the straightforward application of categorical interaction models, which also involve complications with parameter estimation.

(iv) The lack of success of the kernel methods, particularly with high dimensional data. This emphasizes the need for further work on the choice of suitable smoothing parameters.

(v) It should be emphasized that a final choice of a "best" method and variable set depends also on non-statistical criteria specific to the practical problem under investigation, as in the use of prior knowledge to construct sensible "created indicants".

Acknowledgements

We are grateful to the senior neurosurgeons involved in the Head Injury Study, particularly Professor B. Jennett and Mr. G. Teasdale, who were responsible for the collection of this numerically large and well-defined data set, and with whom we have had many consultations over the years. This data collection was part of a larger programme of research into head injury supported in the Department of Neurosurgery in Glasgow by the Medical Research Council; this includes the support of L. S. M. Two of us (D. J. S. and A. M. S.) were supported by the Royal College of Physicians of London and the DHSS, through the Computer Committee of the College. Dr R. Knill-Jones, of that Committee, was involved in the early years in the design, data collection and analysis, and was responsible for the initial application to the Head Injury Study of the independence model. Whilst working on this study G. J. G. received a grant from the fund for juvenile victims of traffic accidents of the Royal Dutch Automobile Club (KNAC).

References

AITCHISON, J. and AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413–420.

AITCHISON, J. and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

AITCHISON, J., HABBEMA, J. D. F. and KAY, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Appl. Statist.*, **26**, 15–25.

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**, 19–36.

—— (1975). Quadratic logistic discrimination. *Biometrika*, **62**, 149–154.

BECKER, D. P. (1979). Comments on Jennett *et al.* (1979).

VAN DER BERGE, J. H., SCHOUTEN, H. J. A., BOOMSTRA, S., VAN DRUNEN-LITTEL, S. and BRAAKMAN, R. (1979). Inter-observer agreement in assessment of ocular signs in coma. *J. Neurol. Neurosurg. Phychiat.*, **42**, 1163–1168.

BRAAKMAN, R., GELPKE, G. J., HABBEMA, J. D. F., MAAS, A. I. R. and MINDERHOUD, J. M. (1980). Systematic identification of prognostic features in patients with severe head injury. *Neurosurgery*, **6**, 362–370.

CHEN, T. and FIENBERG, S. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, **30**, 629–642.

DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics*, **32**, 647–658.

DEMPSTER, A. P., LAIRD, N. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

DUDA, R. O. and HART, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

FIELDING, A. (1977). Latent structure analysis. In *Exploring Data Structures* (C. A. O'Muircheartaigh and C. Payne, eds), pp. 125–157. New York: Wiley

GOLDSTEIN, M. and DILLON, W. R. (1978). *Discrete Discriminant Analysis*. New York: Wiley.

HABBEMA, J. D. F., HERMANS, J. and REMME, J. (1978a). Variable kernel density estimation in discriminant analysis. In *Compstat 1978* (L. C. A. Corsten and J. Hermans, eds), pp. 178–185. Vienna: Physica Verlag.

HABBEMA, J. D. F. and HILDEN, J. (1980). The measurement of performance in probabilistic diagnosis IV. Utility considerations in therapeutics and prognostics. (Submitted for publication.)

HABBEMA, J. D. F., HILDEN, J. and BJERREGAARD, B. (1978b). The measurement of performance in probabilistic diagnosis I. The problem, descriptive tools and measures based on classification matrices. *Meth. Inform. Med.*, **17**, 217–226.
—— (1980). The measurement of performance in probabilistic diagnosis V. General recommendations. (Submitted for publication.)
HABERMAN, S. J. (1974). Loglinear models for frequency tables with ordered classifications. *Biometrics*, **30**, 589–600.
HILDEN, J. and BJERREGAARD, B. (1976). Computer-aided diagnosis and the atypical case. In *Decision Making and Medical Care: Can Information Science Help?* (F. T. De Dombal and F. Gremy, eds), pp. 365–378. Amsterdam, North-Holland.
HILDEN, J., HABBEMA, J. D. F. and BJERREGAARD, B. (1978a). The measurement of performance in probabilistic diagnosis II. Trustworthiness of the exact values of the diagnostic probabilities. *Meth. Inform. Med.*, **17**, 227–237.
—— (1978b). The measurement of performance in probabilistic diagnosis III. Measures based on continuous functions of the diagnostic probabilities. *Meth. Inform. Med.*, **17**, 238–246.
JENNETT, B. and BOND, M. (1975). Assessment of outcome after severe brain damage. *Lancet i*, 480.
JENNETT, B., TEASDALE, G., BRAAKMAN, R., MINDERHOUD, J. and KNILL-JONES, R. (1976). Predicting outcome in individual patients after severe head injury. *Lancet i*, 1031–1034.
JENNETT, B., TEASDALE, G. M. and KNILL-JONES, R. P. (1975). Predicting outcome after head injury. *J. Roy. Coll. Physns. Lond.*, **9**, 231–237.
JENNETT, B., TEASDALE, G., BRAAKMAN, R., MINDERHOUD, J., HEIDEN, J. and KURZE, T. (1979). Prognosis of patients with severe head injury. *Neurosurgery*, **4**, 283–288.
KRZANOWSKI, W. J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Ass.*, **70**, 782–790.
LACHENBRUCH, P. A. (1975). *Discriminant Analysis.* New York: Macmillan.
LACHENBRUCH, P. A. and MICKEY, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–10.
LITTLE, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *J. R. Statist. Soc.* B, **41**, 76–87.
MCCULLAGH, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika*, **65**, 413–418.
MOORE, II. D. H. (1973). Evaluation of five discrimination procedures for binary variables. *J. Amer. Statist. Ass.* **68**, 399–404.
MURRAY, G. D. (1979). Missing data problems in discriminant analysis. Ph.D. Thesis, University of Glasgow.
MURRAY, G. D. and TITTERINGTON, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist.*, **27**, 325–334.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
SKENE, A. M. (1978). Discrimination using latent structure models. In *Compstat 1978* (L. C. A. Corsten and J. Hermans, eds), pp. 199–204. Vienna: Physica Verlag.
TEASDALE, G. and JENNETT, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet ii*, 81–84.
TEASDALE, G., KNILL-JONES, R. and VAN DER SANDE, J. (1978). Observer variability in assessing impaired consciousness and coma. *J. Neurol. Neurosurg. Psychiat.*, **41**, 603–610.
TEASDALE, G., MURRAY, G., PARKER, L. and JENNETT, B. (1979a). Adding up the Glasgow Coma Score. *Acta Neurochirurgica Suppl.*, **28**, 13–16.
TEASDALE, G., PARKER, L., MURRAY, G., KNILL-JONES, R. and JENNETT, B. (1979b). Predicting the outcome of individual patients in the first week after severe head injury. *Acta Neurochirurgica Suppl.*, **28**, 161–164.
TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22**, 259–268.
TRAMPISCH, H. J. (1976). A discriminant analysis for qualitative data with interactions. *Computer programs in biomedicine* **6**, 50–60.
ZENTGRAF, R. (1975). A note on Lancaster's definition of higher-order interactions. *Biometrika* **62**, 375–378.

DISCUSSION OF THE PAPER BY DR TITTERINGTON *ET AL.*

Dr J. A. ANDERSON (University of Newcastle upon Tyne): It is a great pleasure to be able to propose the vote of thanks tonight, particularly as it is a "Glasgow" paper. I realize that a minority of the authors are from Glasgow but the work described is essentially "Glasgow". I have been interested in statistical diagnosis for many years and have found colleagues in Glasgow, medical and statistical, to be a valuable source of stimulus.

The paper before us bears witness to an immense amount of hard work by an immense number of people and is destined to become a classic in the field of discrimination. Hence, in the best traditions of the Society, I propose to interpret the results of the paper from a different viewpoint. A partiality for the linear logistic method may be detected.

There is a dilemma which faces anyone proposing to compare statistical techniques of practical interest. Should one attempt a theoretical approach and derive results which apply to a family or families of distributions? At its worst, this leads to conclusions such as normal distribution methods are best for normally distributed data. Alternatively, should one compare the methods using a real data set and run the risk that this is atypical in some way? The authors have taken this braver, second course and invested it with some authority by choosing four different sub-sets of the data with different dependence properties and varying incidences of missing values. This is valuable, but I doubt whether their results are applicable in all contexts. For example, there must be problems where quadratic discrimination is better than linear discrimination.

The paper is concerned with comparisons and choice between discriminant methods but it omits two important considerations. Firstly, there is no mention of the necessity to look at the data before starting to use any method. Standard practice is to consider the data for possible transformations, outliers, collinearity and anything else relevant. Secondly, there is no mention of the effect of sample size on the complexity of model that we are prepared to analyse. Thus, it is sensible to use simple models for small samples. These considerations enable a better choice of model to be made.

Moving on through the paper, I would like to comment on the methods selected for discrimination. One unifying feature of the independence, normal-based and linear logistic methods is that they all give rise to linear discriminators, as consideration of the log-likelihood ratios indicates. The difference between them is in the method of estimating the coefficients.

The independence model has been used here in its most basic "naive Bayes" form which insists on prediction variables that are nominal categorical variables. As we are assuming independence, we can use any convenient model for the marginal distributions. In particular, continuous variables can be used as such if we assume normal margins (after transforming if necessary). This would still give a linear discriminant; ordered variables could be "scored" and handled this way or more complicated models (McCullagh, 1979) could be considered.

There appears to be imbalance in the amount of polish allowed for the techniques. This is perhaps inevitable. There is a world of difference between "a poor thing—but mine own" and "a poor thing but *his*". For example, the independence model has a choice of three values of the discount factor B allowed. This corresponds exactly to shrinking the coefficients in the linear discriminant by the factor B. Shrinkage factors are also available for the normal-based and linear logistic methods but have not been used.

The choice of missing value techniques shows some unevenness. For the normal-based approach, two methods are given, a computationally cheap but crude approach (substituting the overall mean for any missing value), and a computationally expensive and complex approach, using the *EM*-algorithm. However, a straightforward alternative is available which may be preferable. This is to estimate the means and covariances from all observations and pairs of observations that are not missing. A linear discriminant function can then be calculated and amended for any missing values at the cost of one matrix inversion. This provides a useful compromise between the above alternatives and information about its operating characteristics would have been useful.

The missing value technique for the linear logistic method is again to substitute overall means and this is acknowledged to be crude. Better methods are being developed in Newcastle but have not yet been fully tested. These methods have not been needed so far as the medical collaborators of adherents of the linear logistic method have been trained to produce complete data sets.

Guilt is expressed because the ordered nature of some of the variables has not been utilized. I would like to give some reassurance on this point as I believe the potential gain here is likely to be small. John Pemberton, also at Newcastle, and I have investigated the performance of three methods of dealing with ordered categorical predictor variables where the response variable is binary. We considered "binarizing" the ordered variable, scoring it or allowing the ordering to be expressed. There was very little difference between the methods with our data set.

The yardsticks for assessing performance of the discriminant rates do not distinguish between inferential and decision contexts. Statistical diagnosis tends to be concerned with inference but decision problems do occur. For example, I have been concerned in a project designed to select patients for preoperative anticoagulant therapy to reduce the risk of postoperative deep vein thrombosis (Crandon *et al.*, 1980). In contexts like this, a near miss is as good as a mile and the error rate is a good yardstick.

I question the conclusion that the independence method is much simpler to use than the normal-based for missing values provided the alternative approach is taken instead of the approach based on the *EM*-algorithm. In any case, missing values are rarely as common when a diagnostic method has been established and predictors are known to be required. I agree that for clinical practice, simple methods are

still desirable for many reasons. However, the disadvantage of unrealistic models, like the independence model, is that it is not clear how to proceed with them for any problems other than discrimination. For example, if we assume independence, it is not clear how we can select good diagnostic predictors from a large pool of possibles, and yet this is generally regarded as being very important. To take the other extreme, in some diagnostic or prognostic problems, there are very few possible predictors, perhaps one or two. If these are subject to significant measurement or biological variability, a diagnostic system can be improved by replication. How many replicate measurements should be taken if any? It appears that the independence model cannot provide answers to questions like these, whereas the normal-based and linear logistic models can.

I now consider the Head Injuries study itself. Sufficient detail has been given to indicate what an important and exciting field this is but we have not been told about the impact of statistical diagnosis on clinical practice. Are any of the diagnostic methods discussed here thought to be worthwhile by the doctors concerned? Are any methods in routine use? I note that the authors feel that they have failed in one of their primary objectives, to identify the small number of cases who will be severely disabled. For those who are not familiar with this field, it is worth noting that statistical methods have made useful contributions in several diagnostic areas. For example, the linear logistic approach to the selection of patients for preoperative anticoagulant therapy (Crandon *et al.*, 1980) was generally considered to be very successful from the clinical viewpoint.

Finally, I return to the choice between methods. I do not believe there is an overall best method. Each diagnostic problem should be considered on its own merits, taking into account characteristics of the data, the sample size, the context and the objectives. This will eliminate some possible methods and the paper before us will help in the choice between the remaining contenders.

I have pleasure then in proposing the vote of thanks.

Dr C. C. SPICER (University of Exeter): I warmly welcome this paper as a contribution to the problems of discriminant analysis in medicine. It is perhaps worth pointing out that methods of the kind discussed are equally applicable both to diagnosis and prognosis though the latter is that discussed tonight. The difficulties that arise are mainly due to two peculiarities of the medical approach: first the use of categorical or binary variates and second the tendency to use these in large numbers compared with other branches of biology. It is not unusual in these studies for the doctor to propose 100 or more indicants as possibly relevant to prognosis or diagnosis.

No statistician could view with satisfaction the state of his subject in this field.

My own approach was initially based on the idea that the assumption of independence between the indicants was so unreal that a realistic model would certainly pay dividends in reducing the number of tests used and associating them in groups whose clinical meaning was plausible. It also appeared to me that these were some basic statistical tools to hand which would be particularly helpful since the problem can be stated in terms of large sparse dimensional contingency tables. The sample paradigm consists in estimating the frequencies separately in the cells of the table for each of the diseases or conditions to be differentiated and combining these with the priors. In prognosis it seems more natural to use what Dawid has called the "Diagnostic" paradigm in which each cell of the table contains the fraction of cases in which, say, death or recovery has occurred. The ordering of the categories is usually intended by the clinician to reflect the severity of the conditions and the usual methods of linear modelling can then be used to estimate the predictive capacities of the indicants (and their interactions).

In the purely diagnostic problem the categories frequently have no such meaning.

I started by using the model of contingency tables put forward by Lancaster but abandoned it when Nelder and Wedderburn first produced the GLIM system of programs. In any case I do not think Lancaster's model is appropriate to the current problem.

To sum up my own experience in trying these methods I would say that although they are frequently useful and illuminating, and above all economical, they are never quite as good in prediction as the cruder and more empirical methods discussed in this evening's paper. They are also much more liable to be invalidated by missing data. Until some new type of mathematical model is devised in which the number of parameters can be kept within reasonable bounds, the methods discussed by the authors tonight are likely to be the mainstay of statistics as an aid to medical decision making.

I was interested, though not too surprised, that the linear discriminant came through so well. I have used this a great deal in stepwise mode to try and reduce the total number of indicants. What seems to happen is that it either works reasonably well or it is hopeless. In the latter case this is usually quite obvious as it will not even classify the training set effectively. In this connection I should like to emphasize that the

linear discriminant is *not* dependent on the assumption of multivariate normality. It was introduced by Fisher as a function which maximized the ratio of two quadratic forms representing variation between and within groups. It will do this regardless of the type of numerical variate involved, if there is a reasonable separation of groups in some multidimensional space. I do not think that the quadratic discriminant is necessarily a natural generalization of the linear and can be quite unsuitable.

The general question of linear and quadratic discriminants in which the scaling matrix is not necessarily a covariance matrix has been admirably discussed by Day and Kerridge (1967) and Day (1969).

I would advise anyone who is involved in this kind of work to try out a linear discriminant as it is often illuminating and failures can be discreetly suppressed.

I was glad to see that the clinicians were well integrated into this investigation. There is little possibility of achieving accurate and complete recording of the data unless one or more (preferably senior) clinicians are involved. Doctors do not like filling in long pro-formas and are apt to feel that if a technique merely confirms their own opinions in the first dozen or so patients that it has nothing to teach them.

It would be nice to end this contribution with some trenchant criticisms of the paper and suggestions on the way forward. I have no major objection to the methodology though some criticisms are implied, perhaps, in what I have already said. In this field the Bayesian can usually lie down with the frequentist. But I think that if some Bayesian could come up with a more suitable prior distribution than the Dirichlet, which seems to be the usual choice for contingency tables, some real progress might be made. Unless something quite new in the way of statistical methodology appears then I feel that the next advances lie in machine intelligence and/or the use of micro-computers and software of the "perceptron" variety. (In one experiment of mine the perceptron did about as well as the independent Bayes.)

In view of the recent rumpus about the diagnosis of brain death, which is closely related to prognosis of head injury, it is not necessary to stress the practical side of the present paper. It gives me great pleasure to second the vote of thanks to the authors for their valuable contribution and for bringing this important subject to the notice of statisticians.

The vote of thanks was passed by acclamation.

Dr R. P. KNILL-JONES (Department of Community Medicine, University of Glasgow). I also congratulate the authors on an interesting and timely paper and hope that similar comparisons of discrimination techniques will be undertaken on different data sets in the future. It would be unwise to overgeneralize from the results of this study until other discrimination problems have been examined in this way.

There have been relatively few attempts to compare techniques in the past, and certainly none as extensive as in this paper. However, Croft and Machol (1974) did use ten different discrimination techniques on a large body of data from patients with liver disease—1991 training cases and 437 test cases. There were 20 disease classes and 50 indicants of which 38 were discrete. Assessment was by error rate only. The independence model gave the best results compared both to a number of linear discriminant functions, and to classification procedures using a nearest neighbour rule. There were some difficulties in using the latter methods because of singular matrices.

The construction of loss matrices for medical problems has improved in recent years. Technically a loss matrix is equivalent to a matrix of regrets (a regret being the difference between the utility of a state of health following optimal treatment, and that following suboptimal treatment consequent upon a wrong allocation to a diagnostic class). Formal construction of a regret matrix may be carried out by a combination of ranking the individual regrets, followed by a wagering procedure using indifference points. A good agreement between clinicians has been found (Le Minor *et al.*, 1981) and a "consensus" matrix may be useful for evaluating discrimination rules.

Construction of a regret matrix may be feasible when the preferred classical approach of obtaining utilities of the health states of each consequence is infeasible because of the complexity of the decision structure of the problem.

A more precise overall association parameter for each set of variables used with the independence model could easily be obtained from the training data set by simulation and numerical approximation. It might also be possible to obtain a maximum likelihood estimate for this parameter. Either method would improve the reliability (or calibration) of posterior probabilities produced by the independence model.

Dr P. J. BROWN, (Imperial College, London): Problems associated with medical diagnosis and sparse contingency tables have tried and tested many statisticians over recent years. The problems are

considerable, and tonight's authors are to be congratulated on producing such definitive answers for us to examine. Perhaps they could have been braver and allowed us a few data tabulations, but maybe we have sufficient to go on.

I wonder about the adopted separation of model formulation and method of estimation. If one insists on maximum likelihood or some computationally cheap surrogate, indeed problems will arise if many variables and many parameters are included. However, if estimation methods are allowed which smooth in a data-based fashion, problems need not arise. Why jump from all main effects to all main effects plus interactions, for example? If such implicit exchangeabilities are really *a priori* meaningful, why not introduce a prior assumption to the effect that interactions are likely to be smaller than main effects? One formulation of this would lead to procedures analogous to ridge regression arising in normal theory models. Or, more traditionally one can test whether certain interactions or main effects should be set equal or equal to zero. Indeed, model criticism, re-emphasized recently by Box (1980) is rather absent from tonight's paper. The weakness of assumptions in non-parametric kernel methods, for me, is a serious handicap because consequently they do not allow such modelling.

In this general context, perhaps the authors could explain their notion of comparability given in Section 3.2(ii).

Another aspect that might be modelled is the multicentre nature of the data. Are the data consistent from centre to centre? Certainly, protocols of safety, for example in wearing of crash helmets on motorcycles, differ from country to country. Since the authors sensibly choose to split their data at random, the various centres will be represented similarly in the validation and training set, but of course this need not be the case in any future application. If between-centre differences are real, this could be important at least for those methods which do not model outcome conditional on symptoms. Such modelling may still require some different parameters for different centres for example if effective patient care varies, but avoids the intercentre weighting difficulty and focuses attention on the important discriminatory variables. The advantage is not specific to logistic models, and it is shared, for example, by rather more specialized key models which I have given in an earlier paper, in 1973. Separate population modelling of outcomes contributed to the difficulties exposed in Brown (1976).

Finally, a not unimportant technical detail is that, by definition, maximum likelihood estimates of probabilities for Lancaster models cannot be negative since the likelihood is zero for negative probabilities. If many estimates turn out to be zero, or rather unstable, this may be an indication that the model needs modification. The method of moments used by the authors has the virtue of simplicity, but considering the computational slack available when this method is compared to kernel methods perhaps the more efficient maximum likelihood method is warranted, in the absence, that is, of more genuine parametric smoothing.

Professor J. B. COPAS (University of Salford): However much mathematical theory we have about statistical methods, a thorough study of how well such methods work on real data is always to be welcomed, and I am sure that tonight's paper is an important contribution to the practice of discriminant analysis. If I were allowed just one criticism I think it would be that relatively little attention is given to a preliminary analysis of the data. For instance, when the incidence rates are tabulated against the explanatory variables taken 2 or 3 at a time, how large are the interactions? Is it really sensible to fit a very complicated model unless there is reason to believe that these various interactions are important? The fact that some of the measures of separation given in the tables exceed the null values given at the end of Section 4.1, meaning that the relevant method is actually worse than ignoring the data altogether, suggests that some of the models are grossly overfitted. I doubt if I am the only person here tonight whose reaction to the paper is "Interesting, but what a pity they haven't tried *my* favourite method". My method would depend on a preliminary analysis of the data, and would probably end up with a linear discriminant or logistic regression based on whatever variables and interactions seemed to be important as judged from both empirical and clinical points of view.

Many of the methods in the paper are essentially linear; even the independence model can be viewed as a sum of estimated marginal log likelihood ratios. If we have a linear score $\boldsymbol{\beta}'\mathbf{y}$, then over the test data its variance will be $\boldsymbol{\beta}'V\boldsymbol{\beta}$ where V is the variance–covariance matrix of $\mathbf{y}$. Similarly, the score as estimated, $\hat{\boldsymbol{\beta}}'\mathbf{y}$, will have variance $\hat{\boldsymbol{\beta}}'V\hat{\boldsymbol{\beta}}$ which, on average, will exceed $\boldsymbol{\beta}'V\boldsymbol{\beta}$ by a positive amount $\mathrm{tr}(V\ \mathrm{var}(\hat{\boldsymbol{\beta}}))$. Thus the predicted scores will tend to be too extreme, particularly when the dimension of the model is large. The fact that DT for LINLOG increases with dimension, and is always better than that for NORQUAD, suggests that the variance inflation dominates the reliability assessment. Rather than use omnibus measures of accuracy as proposed in the paper, it is perhaps worthwhile to monitor this effect directly. This can be done graphically

by displaying, for any pair of outcomes (i,j),

$$\sum_{\pi_i} \psi\left(\log\frac{p_i}{p_j} - x\right) \Big/ \sum_{\pi_j} \psi\left(\log\frac{p_i}{p_j} - x\right)$$

against x. Here, ψ is a suitable kernel window function, and the summations are over the test cases in π_i and π_j. This estimates the actual log probability ratio of $\pi_i : \pi_j$ as a function of the estimated value of this ratio. A method which is "reliable" in the sense of the paper will give a slope equal to 1; variance inflation will result in a slope less than 1. It is likely that this slope will be particularly small for the more complicated models, indicating that estimated probabilities are considerably biased.

Professor A. P. DAWID (The City University): In the early days of the century, before anybody knew about sufficiency or the Cramér–Rao inequality, several papers were written on the possibility of finding a better unbiased estimator than the sample average for the mean of a normal distribution. The subject provided ample scope for ingenious suggestions, for their theoretical and empirical study, and for the comparison of competing ideas; but of course it was ultimately fruitless. Whatever empirical support a contending estimator may once have possessed, or however useful it might be as a robust estimator when normality is in doubt, no one would now prefer it if the model of normality were truly believed. Where modelling rules undisputed, empiricism must give way.

Tonight's paper is a masterful collaborative empirical comparison of a number of ingenious suggestions in the field of discrimination. However, the authors have completely ignored any theoretical considerations of the suitability of different models for the data considered. In this problem of prognosis, the feature vector y is measured soon after injury; the outcome π is assessed 6 months later. Is it not obvious that the probability $p(\pi \mid y)$ is directly meaningful, and describes the natural evolution of the recovery process over time? On the contrary, the "backward" probability $p(y \mid \pi)$ is a contorted and meaningless product of an irrelevant analogy with the problem of diagnosis, in which the disease exists prior to the symptoms it produces. I therefore strongly dispute the statement in the first paragraph: "In statistical terms, however, the problems [of prognosis and diagnosis] are technically identical."

However well methods based on modelling $p(y \mid \pi)$ may turn out empirically, it seems to me misguided to use such a logically unsatisfactory approach. More effort should be applied to extending direct analyses of the LINLOG type. Not only did this perform moderately well, but the estimated $p(\pi \mid y)$ is a meaningful quantity. Moreover, what I have called the diagnostic paradigm (but would here be better called the *prognostic* paradigm) offers the only hope of taking into account changes in the outcome category over time. (An exception might perhaps be made for the latent class models, in which the latent class is imagined determined at injury, and itself governs the development of both y and π, independently; then we should need to model the separate factors of $p(y, \pi, l) = p(y \mid l)\, p(\pi \mid l)\, p(l)$.)

My final point is that the empirical comparisons of this paper constitute the outcome of a single trial. If the series had been divided into two or more subseries, and the methods of construction and assessment applied to each such subseries (itself divided into training and test sets), we would have had some valuable empirical information on the reliability of the comparative results presented.

Mr G. J. GOODHARDT (Thames Polytechnic, London): I very much welcome empirical studies comparing different statistical techniques. The authors were fortunate in having as many as 1000 cases to deal with. This enabled them to divide the data into a training set and a totally independent test set. The more usual situation is that, due to scarcity of data, we are forced to use all the available cases as the training set. We may still try a number of different procedures but the only way we can compare their expected efficiency for future classifications is to assess them on the way they classify the cases we already have. That is, we use the training set as a test set. This clearly gives us a biassed estimate of the power of each procedure, but we console ourselves with the thought that the bias is possibly much the same for all procedures and so the comparison may not be affected. But is this in fact so?

Could the authors of tonight's paper cast some light on this problem? What would happen if they were to apply their 19 different methods to the cases in the training set and were to assess their relative performance there? If the rank orderings of the methods on the various criteria were the same as obtained on the independent test set, this would be very reassuring to those who are used to making the best they can of much smaller data sets.

Finally, despite the last sentence in the paper, would the authors be prepared to say which of the 19 methods they thought was the best—or would there be seven different answers to that question?

Drs GRAHAM TEASDALE and BRYAN JENNETT (Institute of Neurological Sciences, Southern General Hospital, Glasgow): Doctors are gradually accepting the value of applying the techniques of statistics and the tools of computing to ordinary clinical practice, as well as in research. Unfortunately, what too often happened in the past was that data were handed over to statisticians with inadequate information either about the limitations of the process of data collection or the uses likely to be made of the analyses. A consequence has been that doctors can be provoked into controversy, each championing what "his statistician" has told him about the relative validity of their own work and that of others. This has already happened in the application of discriminant techniques to the problem of predicting outcome after severe head injury (Jennett, 1980). By contrast, the study that formed the basis for this paper was characterized by extremely close collaboration between clinicians and statisticians, over a period of many years, as was recommended by Professor Healy in his recent inaugural address (Healy, 1979). One component of this collaboration was an extensive analysis of the data available in the bank, from which those prognostic factors were identified that were most appropriate, both from clinical and statistical standpoints.

The data collected in the International Collaborative Study have already been used to compare the efficacy of different methods of treatment, by analysing the effect of alternative methods on actual outcome, as compared with that predicted (Jennett, 1980; Jennett *et al.*, 1980). However, predictions have not yet been used in making management decisions about individual patients. One reason for delaying this application was the need first to explore the "best" statistical methods. An important finding of the present study is the demonstration that arguments about the relative value of different methods of discriminant analysis are of little profit, particularly when taken out of context of the nature of the data analysed. What is now to be discovered is if calculated outcome probabilities will, in practice, aid clinicians in taking decisions such as whether or not a severely injured patient needs, or can benefit from, a certain type of treatment; or alternatively, whether it is appropriate that he be entered into a trial of some new form of treatment (Teasdale, 1980). Doctors should be reminded that this is really not a new or threatening step. Over 200 years ago, while Bayes was still alive, but before his work was published, Sir Percival Potts, a London surgeon, in his book *Observations on the Nature and Consequences of Wounds and Contusions of the Brain* wrote "In matters of this sort, positive proof and conviction are not in our power; all that we can do is, by making a comparison of the conduct and event of a number of similar cases, to come as near to the truth as we can, and to get probability on our side" (Potts, 1768).

Professor M. AITKIN (University of Lancaster): The linear logistic model has had its hands tied behind its back in being restricted to main effects. This is particularly noticeable in set IV: in sets I and II it performs as well as the latent class models, and in set III substantially better. It is not clear why the training set was not modelled by an appropriate logistic model incorporating necessary interactions, and then this model evaluated on the test data set.

Missing data can be handled for the log-linear or logistic model by the *EM* algorithm just as for the normal regression or discriminant analysis model, but a model is then required for the joint distribution of all the explanatory variables. When these explanatory variables are categorical, they can be modelled by a suitable log-linear model.

Professor D. R. COX (Imperial College, London): This very interesting paper represents an impressive collaborative effort. It is similar in spirit to investigations on the forecasting of time series, in which different procedures are applied in a fairly mechanical fashion to empirical series and success measured say be mean squared error. While such studies are interesting, they inevitably raise the question of explanation: what aspects of the series are critical in influencing choice of predictor or discriminator? It would be most helpful if the authors in their reply would indicate the features of the data they think "explain" their conclusions.

Stability of relations in time and between centres is important. What checks of this have been made?

The following contributions were received in writing after the meeting.

Mrs BARBARA GREGSON (University of Newcastle upon Tyne): As an applied statistician using discriminant analysis, I welcome this paper devoted to a case-study and its practical implications. Unfortunately, the authors have excluded the question of variable selection and taken little account of the information that the outcome variable is ordered. I am concerned that these omissions bias their comparison of alternative methods against linear logistic discrimination. This method is well equipped in

both respects and any practical application of it to this dataset would exploit both these advantages to the full.

We in the Health Care Research Unit have developed a procedure for applying LINLOG to problems arising from our research in the N.H.S. More recently we have extended this procedure to the case where the outcome variable is not binary but ordered, by substituting ordinal logistic discrimination (Anderson and Philips, 1981) for the original method of discrimination between unordered populations (Anderson, 1972). We would therefore approach the analysis of the authors' dataset along the following lines.

We would apply ordinal logistic discrimination in stepwise fashion. At each stage the criterion for the inclusion of a variable in the discriminant function would require that it should increase the log likelihood by more than any other variable and by at least $\frac{1}{2}\chi^2_1(\alpha \text{ per cent})$. α is chosen so that account is taken of the number of simultaneous comparisons, as suggested by Russell (1977).

When there are few missing values they can be dealt with quite conveniently: in binary variables they can be replaced by the midrange and in ordinal variables by the median group value. In this dataset, however, five of the variables are missing in more than 20 per cent of the cases. Whether one of these variables is recorded could be as important as its value when recorded. Therefore we would create five additional binary variables to indicate whether each of these five variables is missing. These additional variables would compete for inclusion in the discriminant function. Our experience has shown that such additional variables are rarely needed where there are few data missing, since they usually do not increase the maximum log likelihood significantly.

A similar approach would be used to choose whether the EMV score or its constituent scores should appear in the discriminant function.

Dr P. A. LACHENBRUCH (University of Iowa): Lack of symmetry seems to affect normal theory models seriously, particularly quadratic discriminants. This does not seem to be the case here possibly because most of the variates are dichotomies; those with more categories have a clear upper limit. How bad was the skewness? Could a transformation have helped? The results regarding the independence model(s) are found repeatedly in simulation studies. A recent dissertation (Chang, 1980) found that for small sample sizes in categorical models, the independence model performed well. Log-linear models including interactions often could not be fit, and when they were they did not do as well as independence models. However, if the samples were large enough (twice the number of cells works) categorical models are satisfactory. For ordered categories they do better than scoring methods with discriminant functions because of the many ties.

In missing values problems, there are usually patterns among the missing observations. Does the *EM* algorithm account for these patterns? Also, in this study it might be more likely that a poor outcome was associated with missing values. Thus, the number of missing values might be a good predictor, although it would have little use in practice.

Finally all the centres were pooled. Were there differences among centres? If so, this might be useful for the clinicians to know.

Dr PETER MCCULLAGH (Imperial College, London): Tonight's paper raises many points, some general and some specific to the particular data set. I restrict my comments to a general point which is partly, although not entirely, a matter of terminology and concerns the distinction between response variables and explanatory or feature variables. Discriminant analysis is concerned with allocating individuals to groups or populations on the basis of the observed response; prediction analysis, on the other hand, is concerned with predicting the response given the explanatory or feature variables. In the context of tonight's paper the response variable from either a clinical or from a statistical viewpoint is the outcome or degree of recovery so that the topic of the paper is prediction, not discrimination.

If this were simply a matter of terminology no particular difficulty would arise but a closer examination of equation (2) reveals that in all methods tried except for the linear logistic model it is assumed that for each possible value of the response, π_i, the distribution of the feature vector is assumed constant over centres. This is a very strong and, indeed, counterintuitive assumption because it assumes that, for example, the age by sex distribution of head-injured patients who die is the same from Glasgow to Groningen and from Rotterdam to Los Angeles. It is easy to think of social and legal reasons why this might not be so. The direct method in Section 3.6 avoids such assumptions though I would have preferred a model taking account of the order in the response categories (McCullagh, 1980). Furthermore, the direct approach permits an examination of the consistency of any observed relationship over the different centres

and if necessary an appropriate allowance for between centre differences can be made although this might be undesirable from a purely scientific viewpoint. The treatment of missing values in this context deserves further work.

Dr B. J. T. MORGAN (CSIRO, Melbourne) and Mr D. DAVENPORT-ELLERBY (University of Kent): Tonight's interesting paper has been of particular relevance to us, as we have also been analysing data on individuals with head injuries. In our case only 172 individuals were involved, data having been obtained, mostly retrospectively from records, by Mr John Bartlett and Mr Glen Neil-Dwyer, neurosurgeons at the Brook Hospital, Woolwich. There were no missing data! The reason for the small sample size was simply because a particular form of head injury—the extra-dural haematoma—was being considered. We would like to ask the authors of tonight's paper how the 1000 individuals of the paper were distributed between different types of head injury, and indeed how many had an extra-dural haematoma. Such individuals may, but need not, become severely injured. If the injury is correctly diagnosed, then the haematoma may be simply evacuated by drilling a burr-hole in the skull. Unfortunately, individuals with this injury, resulting from a blow to the head, may not even sustain a fracture, and so early diagnosis is difficult. If distinct types of head injury are present in the sample of 1000 individuals, then prediction of recovery could conceivably be improved if it was done separately for the different injury types.

In our study, we found that individuals with a fast development of the haematoma frequently had a poor outcome after operation, while the converse was sometimes true for individuals with a slow development. This leads us to wonder whether the entry requirement of individuals to the sample of tonight's paper—a minimum of 6 hours in coma—might result in predictions of degree of recovery which may not in general apply to individuals with fewer than 6 hours in a coma.

We found that the best measures for discriminating between individuals with poor and good outcomes after operation were age, two measures of coma state on hospital admission (one of verbal ability and one of motor response), and the same two measures prior to operation at the Brook Hospital. In our case also the linear discriminant function performed well, correctly classifying 79·9 per cent of individuals, as compared with 84·9 per cent resulting from a non-parametric procedure using a kernel suggested by Aitchison and Aitken (1976). However we used equal values for the $p(\pi_i)$, the same test and training set, and no cross-validation.

There are always, of course, unusual features to data sets which confound any discrimination technique. In our study a number of individuals with a good outcome which was not predicted were traced to a surgeon who had amassed more than a lifetime's experience in locating haematomas when working with the allied forces in North Africa during the war!

The AUTHORS replied later, in writing, as follows.

We should like to thank all discussants for taking the trouble to make their contributions. We are warmly grateful for the many positively *complimentary* remarks and we are, perhaps, even more appreciative of what we feel are *complementary* comments to the material of our paper, namely the wide range of critical contributions and queries. We have reported on a small aspect of the practice of discriminant analysis and of the statistical analysis of this data-set in particular. To the discussants is due the credit for any lasting usefulness the publication may have as a more rounded reference on the subject. In reply we should like to discuss some themes which several people have raised.

(i) *A comparison or a case-study?*

We should like to emphasize that our work should not be regarded as a fully-reported case study. For that it would have been appropriate to discuss in more detail the methods of data-collection, the validity of combining the data-sets from different centres (Dr Brown, Professor Cox, Professor Lachenbruch, Dr McCullagh), variable selection (Mrs Gregson), exploratory analysis with a view to selection of method or transformation (Dr Anderson, Professor Copas, Professor Lachenbruch) and the impact on the clinical protocol of implementation of the chosen procedure or procedures. Our aim was the more restricted one of paving the way for case studies by trying to assess the suitability and sensitivity (or otherwise) of a repertoire of techniques. Our results, we hope, may help practitioners to choose what is likely to be a good method or, at worst, one that is liable to be reasonably reliable and comparatively easy to apply. It is, however, fair that we should have to comment on the activities alluded to above.

(ii) *What do the data look like?*

Several contributors have asked about the pooling of the centres, both from the point of view of estimation and of prediction (Dr Brown, Professor Cox, Professor Lachenbruch and Dr McCullagh). Although the present study has not included predictions within the different centres, and although there are certainly differences among the centres, such as that Glasgow patients are received after primary admission to other hospitals, the protocols in the centres are carefully standardized and the distributional patterns for informative indicants and the outcomes are fairly similar; see, for instance, Jennett *et al.* (1977) and Jennett *et al.* (1980) where some information about the stability in time of some features is described. Predictions for Glasgow and Dutch patients using a Glasgow training set are reported by Teasdale *et al.* (1979c) and, in an unpublished study, it was shown that predictions for Glasgow patients using a Glasgow training set were very similar to those using a Dutch training set.

As far as exploratory work is concerned (Dr Anderson, Dr Brown, Professor Copas and Professor Lachenbruch), there is much on which we could, and perhaps should, have reported. The problems are exemplified by the following cross-tabulation of Change with Outcome.

		Outcome		
		D/V	*SEV*	*M/G*
Change	1	202	23	54
	2	99	34	99
	3	74	17	132

Clearly no transformation will symmetrize all column patterns, or row patterns, for that matter. Assumption of Normality is laughable in principle for the conditional distributions of change scores given the outcome and a categorical model appropriate for an ordered response variable with an ordered predictor (Ms Gregson, Dr McCullagh) seems much more suitable. Hopefully, refinements of such models to deal with multivariate and patchy predictors will lead to the robustness of independence or "Normal" (Fisher-linear-discriminant, as Dr Spicer reminds us) based approaches being outweighed. The collinearity problem referred to by Dr Anderson is eased by the use of created indicants.

Recently the data-base has been extended to 1250 patients and some information about trends and interactions is given in Teasdale *et al.* (1981), for "complete" portions of the data. In particular, the log-odds of "death" as against "survival" is linear in age, and the slope seems the same independently of the initial coma-score (*EMV*). Analysis by loglinear models showed that coma-score and "Pupils" score provided significant further discriminatory information.

Although there are strong first-order interactions within many pairs of indicants, these interactions are fairly similar across the outcome categories. Three-way interactions involving outcome are, generally, small and this appears to explain why NORLIN2 is not inferior to NORQUAD, why the linear logistic approach should do well and why the independence model also seems robust. This last can be illustrated with the following extract from the extended data-base, with two outcome categories and using two highly dependent indicants, E score and V score.

		Deaths *V-score*			*Survivors* *V-score*	
		1	2–5		1	2–5
E-score	1	488	59	1	290	109
	2–4	24	28	2–4	56	122

Given these frequencies we obtain the following estimated conditional probabilities of death, with and without the assumption of conditional independence of E-score and V-score.

Indicant scores		*Conditional probabilities of death*	
E	*V*	*Independence*	*No independence*
1	1	0·66	0·62
1	2–5	0·32	0·34
2–4	1	0·29	0·29
2–4	2–5	0·09	0·18

The similarity between the two sets of probabilities is clear.

(iii) *Generalizability of the results*

Dr Anderson and Dr Knill-Jones have issued the caveat that our findings will not universally be true. We hope that this warning will stimulate further studies on real data. Such studies are, we feel, of considerable importance, because real data simply do not come from a parametric model. In theoretical work, as Professor Dawid points out, empiricism must give way. With real data, empirical work is seldom valueless. We do feel, however, that our general conclusions about the robustness of independence-based and linear discriminant function methods will often hold. Even in simulation studies the latter can be disconcertingly efficient compared to, say, the method based on unequal covariance matrices (Aitchison *et al.*, 1977; Remme *et al.*, 1980) and the robustness of independence models has seemed to perplex Professor Healy (1979) as well. We hope that our comments in (ii) above are helpful in this context.

As well as comparison across different real-life problems, replication within a large data-set is both feasible and desirable, for the reasons mentioned by Professor Dawid. The table shows the quadratic scores for four replications, using all four combinations of training (*TR*) and test (*TE*) sets.

Variable set	$TR \rightarrow TE$	$TE \rightarrow TE$	$TE \rightarrow TR$	$TR \rightarrow TR$
I	0·377	0·354	0·412	0·381
II	0·438	0·422	0·435	0·422
III	0·364	0·334	0·396	0·364
IV	0·399	0·352	0·411	0·392

In general, each column picks out set II as the "worst" set, set III as the "best". There is, however, a noticeable effect across the columns. The column 2 scores are uniformly better than those in column 1 and those in column 4 better than those in column 3. The cause of this is the optimistic bias caused by using the same set both for developing a discriminant rule and for testing it, as Dr Goodhardt mentions and as Dr Morgan and Mr Davenport-Ellerby practise. Had it not been for wanting to illustrate this point we should have used the cross-validatory technique of Lachenbruch and Mickey (1968) and, thereby, have reduced this bias.

(iv) *Diagnostic paradigm versus sampling paradigm*

There are two major questions in this context that arise from our paper. Firstly there is the general problem of which approach is more correct in principle. Secondly, there are the practicalities which set the linear logistic approach apart from other methods which lead to log-likelihood ratios that are linear in the observables.

First, it should be said that, since this is a comparison study, we have in principle no axe to grind and we have presented results from both approaches. It is fair comment that sampling paradigm techniques have been looked at in greater number and we are grateful to Dr Anderson, Professor Aitkin and Ms Gregson for reporting on both current and potential increases in sophistication.

In many respects the diagnostic paradigm is intuitively the more appealing one, particularly when Professor Dawid argues as in his 1976 paper. Also, it is often more plausible, say, that the distribution of an indicant for disease sufferers is a truncated Normal than a complete Normal. The linear-logistic formulation is then still nicely appropriate. We are not so sure about the time-based argument that Professor Dawid, with support from Dr Spicer, offers here. If the appropriateness of the factorization of $p(y, \pi)$ should be determined by the time-ordering in which y and π are observed or develop, then it seems

that, in Dawid (1976), the diagnostic paradigm should be called the prognostic paradigm and the sampling paradigm should be called the diagnostic paradigm. What seems to be more important is whether the *particular* model adopted using one factorization is more suitable than the particular model from the other. Sometimes, as in direct applied discriminant analysis, the choice of paradigm has little effect; see Krzanowski (1975). When, however, there are unconfirmed cases in the data bank, the choice of paradigm may lead to quite different results, if we are interested in discovering how much extra information the unconfirmed cases provide about the discriminant rule. With the factorization

$$p(x, \pi) = p(x)\, p(\pi \mid x),$$

there is no extra information unless the above two factors have parameters in common (Anderson, 1979) whereas, with the alternative, it can be shown that unconfirmed cases do improve the performance of the discriminant rule; see, for instance, O'Neill (1978), Ganesalingam and McLachlan (1979) and the dialogue between Dawid (1980) and Makov (1980).

(v) *Treatment of missing values*

Missing values constitute one of the major features of this data-set. Many statisticians will envy the complete response that Dr Anderson stimulates from his collaborators. Undoubtedly, we have used wrong approaches in that the missing data process will not be ignorable. It seems unlikely that a single convincing model will be developed for this process, so that we are led to something like the pragmatic but computationally demanding approach of Rubin (1978). For different suggestions for the missing data process, the missing values are simulated and the resultant "complete" data set is analysed accordingly. Many replications from each process are made and the results are compared. If the choice of missing data process has little effect, it may be safe to assume that any of them can be accepted and the results are fairly reliable. Otherwise, nothing should be done but to obtain more information about the missing data process.

The "unevenness" referred to by Dr Anderson is largely due to the unevenness in the literature. His suggestion of substitution of regression-based estimates in the normal-based approach is worth noting (see also Little, 1978) and, for illustration, we indicate some results from this method, which we call NORLIN3, on Variable set IV. The five performance scores are, respectively, 0·260, 0·670, 0·365, 22·4 and 0·048. By comparison with Table 8 we see that they are comparable with the *EM* results. The relative computational burden may not be quite as at first sight, because the *EM*-algorithm (or some alternative) is only used *once*, with the training-set. As far as application to the test set, or to new patients is concerned, it is simply a matter of evaluating a linear discriminant function, or density estimates. In principle, for the linear logistic approach, prediction for an incomplete case would require re-estimating the parameters for the relevant missingness pattern. Important difficulties may arise if the discriminant rules are to be updated as new prognoses are confirmed. Then, at least, current parameter estimates provide good starting values for a new maximization, or an approximate, sequential version, on the lines of recent work by one of us (D. M. T.) and J-M. Jiang, is possible. Also, as Dr Anderson suggests, maybe later observations will not be incomplete. Unfortunately, good management may not be enough in this study, in which facial damage is often the cause of missing data on eye responses.

We look forward hopefully to the development of missing-data techniques for the logistic methods and for the models of Dr McCullagh. The suggestions of Ms Gregson are interesting, particularly the creation of indicator variables for missingness. The median substitution method is in the same spirit as our mean substitution and may be misleading if, say, the marginal distribution of that variable is bimodal. A similar comment seems appropriate to Professor Aitkin's approach to missing data substitution. A major benefit of the logistic approach is the lack of dependence on a specific model for the explanatory variables (Anderson, 1972). Formally, the missing data problem may be resolved using non-parametric density estimation. Suppose a complete observation is bivariate but that in a particular case from π_1, say, only x_1 is available. We require, in the likelihood, a factor $p(\pi_1, x_1)$. This can be written

$$p(x_1) \int p(\pi_1 \mid x_1, x_2)\, p(x_2 \mid x_1)\, dx_2.$$

With a non-parametric (consistent) estimator for $p(x_2 \mid x_1)$ and if the integration can be done, this gives a possible approach. If a kernel-based density estimate is used with a Gaussian kernel function and if the logistic distribution function is approximated by a Normal one, the integration can be done (Aitchison and

Begg, 1976). The resulting expression is quite complicated, and some work into practical simplification must be done. The same difficulty is inherent in the missing-data modification of continuous kernel methods alluded to in Section 3.7.

The introduction of missing-value indicator variables mentioned by Ms Gregson is in similar spirit to kernel functions like that in Table 1 of Murray and Titterington (1978). The method has been tried on the head injury data, but without improving the performance noticeably. Both approaches react to some extent to the missing data pattern, which is in Professor Lachenbruch's mind. This contrasts with the *EM* algorithm's assumption that the missing data are "missing-at-random".

(vi) *Miscellaneous comments*

We are grateful for Dr Anderson's reference to Crandon *et al.* (1980). The error rate is a good yardstick if the loss-structure is well specified. Otherwise, the arbitrariness of the decision boundary is a worry. The severes in our study are very difficult to identify. Several two-dimensional plots of the data have shown how they lie sparsely scattered in the overlap zone of the other two groups.

Discriminant analysis and multivariate analysis in general are areas where a technique does not, to paraphrase Dr Spicer's words, "merely confirm their own opinions". Often the picture is too complicated for the clinicians, but becomes clear via the computer. Perhaps the recent promotion of the logistic Normal distribution by Aitchison and Shen (1980) may prove to be Dr Spicer's sought-for alternative to the Dirichlet; see also Leonard (1973). Is Dr Spicer using the Fisher separation statistic for stepwise variable selection? Habbema and Hermans (1977) suggest that this should be discouraged.

Maximum likelihood estimation of the overall interaction factors and of the Lancaster model, proposed by Dr Knill-Jones and Dr Brown, may be worth including. Dr Brown mentions data-based smoothing and, of course, this may well be yet another good alternative. Parallels between ridge regression and kernel-based smoothing of relative frequencies can be drawn by comparing the papers of La Motte (1978) and Titterington (1980, Section 7).

Dr Brown asks us about our use of "comparability" in Section 3.2(ii). The point was that, in order to use the *same* data sets for all methods, the "further grouping" mentioned there was avoided.

We are sorry that time has not permitted us to report on application of Professor Copas' promising reliability test.

We hope that other clinicians and scientists will be encouraged by Mr Teasdale and Professor Jennett to seek and implement statistical help. The plethora of useful reports and analyses from this International Study and the increasingly routine implementation of discriminant analysis by the neurosurgeons point to the fruitfulness of practical Medical Statistics.

Variable selection in this study is discussed in Braakman *et al.* (1980). We do not agree with Ms Gregson that the linear logistic method can be singled out as especially suited to dealing with variable selection or ordered outcomes. It would be interesting to compare Ms Gregson's inclusion criterion in practice with other techniques.

In reply to Dr Morgan and Mr Davenport-Ellerby we can report that only 84 of the 1000 patients had an extra-dural haematoma only. Although about half the patients in our study had a haematoma of some sort (Jennett *et al.*, 1979), many more haematomas are treated in the various centres but are excluded because the patients were not in coma for more than 6 hours. The discriminant rules we have developed certainly may not be appropriate for such patients. Prognosis patterns may be very different preoperatively (Teasdale *et al.*, 1976) from those postoperatively, as a current study is confirming. Six-months prognosis depends much more on the age of the patient and the *degree* of brain damage than on its cause.

(vii) *Which method is "best"?*

Dr Goodhardt would like us to answer this. Professor Cox would like the answer explained. "Best" could mean "guaranteed consistent" (kernel, apart from the missing-at-random assumption), "most robust to model variation" (independence or NORLIN2 or NORLIN3, on the basis of simulation studies), "ease of practical application without a large computer" (independence, methods based on functions of a linear discriminant function). For this data-set and for reasons which, we hope, have been indicated already in the reply, the independence and NORLIN2, or NORLIN3, approaches seem as good as any. With further methodological development and more data the kernel methods must come into contention and appropriate missing data modifications to the linear logistic method *ought* to make it a winner. If we add to this the possibility of acknowledging the ordering in the outcome itself, in any of the above methods, then one thing seems clear. The "best" is yet to come!

References in the Discussion

Aitchison, J. and Begg, C. B. (1976). Statistical diagnosis when cases are not classified with certainty. *Biometrika*, **63**, 1–12.

Aitchison, J. and Shen, S. M. (1980). Logistic normal distributions: some properties and uses. *Biometrika*, **67**, 261–272.

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, **66**, 17–26.

Anderson, J. A. and Philips, P. R. (1981). Regression, discrimination and measurement models for ordered variables. *Appl. Statist.*, **30**, 22–31.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with Discussion). *J. R. Statist. Soc.* A, **143**, 383–430.

Brown, P. J. (1973). Aspects of design for binary key models. *Biometrika*, **60**, 309–317.

——— (1976) Remarks on some statistical methods for medical diagnosis. *J. R. Statist. Soc.* A, **139**, 104–107.

Chang, Y. (1980). Discriminant analysis with categorical data. Ph.D. Dissertation, University of Iowa.

Crandon, A. J., Piel, K. R., Anderson, J. A., Thompson, V. and McNicol, P. P. (1980). Prophylaxis of pre-operative deep vein thrombosis: selective use of low-dose heparin in high-risk patients. *Brit. Med. J.*, **281**, 345–347.

Croft, D. J. and Machol, R. E. (1974). Mathematical models in medical diagnosis. *Ann. Biomed. Engng*, **2**, 69–89.

Dawid, A. P. (1980). Contribution to the discussion of Makov (1980).

Day, N. E. (1969). IEEE *Trans. Information Theory*, May 1969. (Correspondence: Linear and quadratic discriminations in pattern recognition.)

Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, **23**, 313–323.

Ganesalingam, S. and McLachlan, G. J. (1979). Small sample results for a linear discriminant function estimated from a mixture of normal populations. *J. Statist. Comput. Simul.*, **9**, 151–158.

Habbema, J. D. F. and Hermans, J. (1977). Selection of variables in discriminant analysis by *F*-statistic and error rate. *Technometrics*, **19**, 487–493.

Healy, M. J. R. (1979). Does medical statistics exist? *Bias*, **2**, 137–182.

Jennett, B. (1980). Comments on paper by Stablein et al. *Neurosurgery*, **6**, 246–248.

Jennett, B., Pitts, L. H. and Murray, L. (1980). Management of severe head injury (letter). *Lancet ii*, 370.

Jennett, B., Teasdale, G., Fry, J., Braakman, R., Minderhoud, J., Heiden, J. and Kurze, T. (1980). Treatment for severe head injury. *J. Neurol. Neurosurg. Psychiat.*, **43**, 289–295.

Jennett, B., Teasdale, G., Galbraith, S., Pickard, J., Grant, H., Braakman, R., Avezaat, C., Maas, A., Minderhout, J., Vecht, C. J., Heiden, J., Small, R., Caton, W. and Kurze, T. (1977). Severe head injuries in three countries. *J. Neurol. Neurosurg. Psychiat.*, **40**, 291–298.

Lamotte, L. R. (1978). Bayes linear estimators. *Technometrics*, **20**, 281–290.

LeMinor, M., Alperovitch, A. and Knill-Jones, R. P. (1981). Applying decision theory to medical decision-making: the concept of regret and error of diagnosis. (In preparation.)

Leonard, T. (1973). A Bayesian method for histograms. *Biometrika*, **60**, 297–308.

Little, R. J. A. (1978). Consistent regression methods for discriminant analysis with incomplete data. *J. Amer. Statist. Ass.*, **73**, 319–322.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

Makov, U. E. (1980). Approximations to unsupervised Bayes learning procedures. *Proc. Int. Mtg. Bayesian Statist.*, Valencia.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Ass.*, **73**, 821–826.

Potts, P. (1768). Observations on the Nature and Consequences of those Injuries to which the Head is Liable from External Violence. Hitch, C. and Hawes, L. pp. 141–142.

Remme, J., Habbema, J. D. F. and Hermans, J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. *J. Statist. Comput. Simul.*, **11**, 87–106.

Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. In *Imputation and Editing of Faulty or Missing Survey Data* (F. Aziz and F. Scheuren, eds). Washington: U. S. Department of Commerce.

Russell, I. T. (1977). British patient's choice of care for minor injury: a case study of separate sample logistic discrimination. *Proc. Amer. Statist. Ass., Social Statistics section*, pp. 548–553.

Teasdale, G. (1981). Prognosis of coma after head injury. Paper presented at the Symposium of Advanced Medicine. Newcastle, 1980. Pitman.

Teasdale, G., Galbraith, S., Parker, L. and Knill-Jones, R. (1976). Prediction of outcome after surgery for traumatic intracranial haematoma. *Brit. J. Surgery*, **63**, 150.

Teasdale, G., Parker, L., Murray, G. and Jennett, B. (1979c). On comparing series of head injured patients. *Acta Neurochirurgia Suppl.*, **28**, 205–208.

Teasdale, G., Skene, A., Spiegelhalter, D. and Murray, L. (1981). Age, severity and outcome of head injury. *Proc. 4th Chicago Conf. Neurol Trauma. In Seminars in Neurol. Surgery.* New York: Raven.

As a result of the ballot held during the meeting, the following were elected Fellows of the Society.

ADEDOYIN, Solomon A.
AL-KASSAB, Mowafaq M.
BAXTER, Michael J.
BECKETT, Valerie R.
BENNETT, Stephen
BROMWICH, Laureen
CRICHTON, Nicola J.
DONNELLY, Peter J.
GARTHWAITE, Paul H.
GENTLE, James E.
HAN, Juat J.
HICKIE, James S.
HUNTINGTON, Eric
ISLES, John E.
JENSEN, Jens L.
KANE, Michael C.
KIM, Geung-Ho
KURIMBOKUS, Nazma
LIM, Suk L.
MANDALLAZ-BISHOP, Daniel M. E.
MATTHEWS, John N. S.
NI BHROLCHAIN, Maire
PALMSTRØM, Stephen H.
PASCHENTIS, Ioanna
RUNGER, George C.
SØRENSEN, Michael
TOYOOKA, Yasuyuki
WATSON, Peter T.
WEBBER, Christopher D.
WHITNEY, Andrew
YUEN, Pui-Ham A.