Logistic Regression in Medical Decision Making and Epidemiology

Financial support by the Netherlands Heart Foundation for the publication of this thesis is gratefully acknowledged.

LOGISTIC REGRESSION IN MEDICAL DECISION MAKING AND EPIDEMIOLOGY

Logistische Regressie in Medische Besliskunde en Epidemiologie

Proefschrift

ter verkrijging van de graad van doctor in de geneeskunde
aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof. Dr. M. W. van Hof
en volgens besluit van het College van Dekanen.
De openbare verdediging zal plaatsvinden op
woensdag 23 april 1986 te 15.45 uur

door

PAULUS IGNATIUS MARIA SCHMITZ
geboren te 's Gravenhage

Promotiecommissie

Promotor : Prof. R. van Strik
Overige leden : Prof. Dr. A. P. J. Abrahamse
Prof. Dr. R. Doornbos
Prof. Dr. P. J. van der Maas

Contents

VOORWOORD

In het voorjaar van 1974, kort nadat ik als consulterend biostatisticus binnen het Instituut voor Biostatistica was begonnen, deed professor R. van Strik verslag van een lezing van Donald Rubin over matchen in observationeel onderzoek. Het onderwerp intrigeerde me: was dit matchen wel verantwoord? Was dit niet het zoveelste onoordeelkundig gebruik van de statistiek? Bij het bestuderen van relevante literatuur bleek mij dat observationeel onderzoek zeer interessante probleemgebieden omvat. Wie de effectiviteit van autogordels wil onderzoeken kan geen gerandomiseerde studie doen, maar zal retrospectief gegevens verzamelen van ongevallen en achteraf vergelijkingen maken. Matchen bleek één van de methoden om voor de mogelijke onzuiverheid van effectschattingen, veroorzaakt door de afwezigheid van randomisatie, te corrigeren. Soortgelijke situaties doen zich voor bij epidemiologisch onderzoek, bijvoorbeeld naar de invloed van roken op de kans op het krijgen van longkanker.

Dit bracht me op het terrein van de etiologische epidemiologie, en daarmee op het gebied waarin meer geavanceerde statistische technieken zeer bruikbaar zijn. Het logistische model is er één van. Dit model is in de loop van de jaren zeventig in steeds breder kader toegepast, vooral nadat bleek dat het ook uitstekend voor de analyse van case-control onderzoek toepasbaar is. Toen dan ook collega ir. Wim Hop het computerprogramma van Elisa Lee voor logistische regressie-analyse binnen ons instituut bracht verschafte dit nieuwe perspectieven. Beide brachten we wijzigingen aan en het programma werd vervolgens op ruime schaal in Nederland verspreid. Een en ander resulteerde in een studieweek voor epidemiologen, in april 1980 op Texel, onder voorzitterschap van prof. van Strik, waarbij ik één van de docenten mocht zijn. In Seattle kreeg ik vervolgens gelegenheid bij prof. N. Breslow en prof. N. Day mijn kennis van dit model te completeren. Een praktische toepassing leverde het Tilburgse materiaal van dr. Bert Herman, een case-control studie van risicofactoren voor een beroerte.

Parallel met deze ontwikkelingen verliep mijn interesse voor de toepassing van het logistische model voor discriminantanalyse. Onderzoekmateriaal van dr. Joop van de Merwe en prof. F. Wensinck (agglutinatietesten voor de diagnose van de ziekte van Crohn) bleek zeer geschikt voor deze toepassing. Dr. Dik Habbema stelde voor een vergelijkende studie op te zetten van discriminantanalysemethoden: het "Leidse" ALLOC-programma en "mijn" logistische programma, aangevuld met enkele algemene methoden. Dit leidde tot een zeer vruchtbare samenwerking: Dik Habbema is van vijf hoofdstukken in dit proefschrift co-auteur. Ook dr. Jo Hermans had een belangrijke inbreng in de vergelijkende studies over discriminantanalyse.

Het was Dik Habbema die mijn belangstelling voor de medische besliskunde heeft gewekt. Medische besliskunde, diagnostische testen en discriminantanalyse zijn begrippen die in elkaars verlengde liggen. Voor zover het dit proefschrift betreft is deze constatering alleen in hoofdstuk 2 terug te vinden. De ontwikkeling gaat verder, waarbij ik veel stimulans ondervind door deelname aan de Leidse studiegroep Medische Besliskunde waarin o.a. Jo Hermans en Berti Zwetsloot participeren.

In het voorgaande zijn enige achtergronden geschetst, maar lang niet alle. Zonder de voortdurende steun en stimulans van mijn promotor prof. R. van Strik bij mijn proefschrift en andere aktiviteiten zou dit boekje er niet zijn gekomen: ik ben hem hiervoor zeer erkentelijk. In de laatste fase van dit proefschrift werd behalve door prof. van Strik, ook door de leden van de promotiecommissie nog veel aandacht aan het manuscript geschonken. Prof. dr. A. P. J. Abrahamse, prof. dr. R. Doornbos en prof. dr. P. J. van der Maas bedank ik voor de tijd die zij in deze fase aan mijn proefschrift hebben besteed. Cobie Jansen gaf blijk van haar inzet bij het typen van een aantal hoofdstukken. Van mijn beide paranimfen, drs. Fred Munning en drs. Wim van Putten, ondervond ik vooral morele steun. En tenslotte: Kiki, na 23 april heb ik in de weekends weer meer tijd voor jou.

INTRODUCTION

INTRODUCTION

In his recent textbook "Primer of Biostatistics", S. A. Glantz refers to the nowadays growing pressure on clinicians for more effective use of medical resources. He asserts that clinicians should be able to make better informed judgements about claims of medical efficacy. They can participate then more intelligently in the debate on how to allocate medical resources. These better informed judgements are the objective of "medical decision making", where the choices to be made in diagnostic and therapeutic strategies are studied.

Medical decision making is to be based for a great part on statistical reasoning. A statistical approach which has proven to be valuable in this field is a technique known as logistic regression analysis. The assessment of the performance of this logistic model is the subject of this thesis. Applications in medical decision making are considered, mainly with respect to medical diagnosis. Logistic regression analysis may be viewed as a sophisticated diagnostic aid. Multiple test outcomes and patient characteristics are incorporated into a logistic model for the probability that a patient belongs to a certain disease class.

Performance of the logistic model in etiologic-epidemiological studies is another area of study. In this context logistic regression is used for the detection of risk factors, adjusted for confounding in order to obtain unbiased assessments.

This thesis includes ten chapters, consisting of papers which have either been published or which were recently submitted for publication by the author, many of them in cooperation with various colleagues. The chapters have been grouped into three parts. Part 1 (chapter 1) presents a review of developments in logistic regression from 1970 up to 1986. It outlines statistical aspects of the model such as estimation, hypothesis testing and model selection. Part 2 (chapters 2 - 7) deals with logistic discriminant analysis in medical diagnosis. An extensive evaluation of a logistic model for the diagnosis of Crohn's disease by agglutination reactions is presented in chapter 2. Actually, this chapter results from our evolving insights since the first application of logistic discriminant analysis for the diagnosis of Crohn's disease (chapter 3). Comparison of logistic discrimination with some other discriminant analysis methods is studied in chapter 4 through application to real data from clinical practice, and in chapters 5 and 6 through the use of simulated data. This comparative evaluation was performed on datasets consisting of mixtures of continuous and discrete data. The underlying distribution in the first simulation study (chapter 5) is a fourdimensional normal from which discrete variables were obtained by discretizing the continuous variables. The simulation study in chapter 6 is based on a location model. Within each outcome combination of the discrete variables a multivariate normal distribution is assumed. The next chapter concerns the evaluation of logistic discriminant analysis for modelling QSAR's, quantitative structure-

activity relationships (chapter 7). It is used there as a technique for detecting which chemical compounds will be useful for the development of new drugs. The third and final part of this thesis (chapters 8 - 10) is concerned with the application and evaluation of the logistic model in etiologic-epidemiological studies, particularly in case-control studies. In chapter 8 the first epidemiologic application of the logistic model in The Netherlands, which uses estimates of the model parameters based on conditional likelihood, is presented. Risk factors for stroke are investigated in a case-control study conducted some years ago in Tilburg. In chapter 9 the (multiplicative) logistic model is compared with an (additive) linear model for case-control studies with one continuous exposure factor and without consideration of confounding variables. Some aspects concerning variables selection in epidemiologic studies, also relevant for logistic regression, are discussed in chapter 10. The thesis is concluded with a summary.

CHAPTER 1

DEVELOPMENTS IN LOGISTIC REGRESSION METHODOLOGY: 1970 - 1986

Devolopments in Logistic Regression Methodology: 1970 - 1986. P. I. M. Schmitz. Submitted for publication, February 1986.

DEVELOPMENTS IN LOGISTIC REGRESSION METHODOLOGY: 1970 - 1986

Paul I. M. Schmitz
Institute of Biostatistics
Erasmus University
P. O. Box 1738
3000 DR Rotterdam
The Netherlands

Summary

Some theoretical developments of the multiple logistic regression model during the period 1970 - 1986 are reviewed. Major topics covered are: 1. Inference: point estimation, hypothesis testing and interval estimation. 2. Model selection: goodness of fit, regression diagnostics and selection procedures. 3. Generalizations: general linear models and polychotomous logistic regression. Emphasis is given to applications in medical decision making and epidemiology.

1. Introduction

In the biomedical field, data in binary form such as disease/no disease or survival/death are very common. Suppose that such binary observations are available on n individuals. The observations can be represented by n random variables $Y_1, Y_2, \ldots, Y_n$, each with possible outcomes 0 or 1. Outcome $Y_i = 1$ may be referred to as a "success", $Y_i = 0$ as a "failure". The main task is to study the dependency of the probability of "success" $P(Y = 1) = \theta = E(Y)$ on a set of p explanatory variables $x_1, x_2, \ldots, x_p$. In the multiple linear logistic regression model, this dependence is assumed to be described by the logistic transform of a linear combination of the explanatory variables:

$$P(Y = 1 \mid x_1, x_2, \ldots, x_p) = 1 / (1 + \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))) \qquad (1.1)$$

(Throughout this paper, stochastic variables are denoted by capital letters, vectors and matrices by underlining).

The earliest applications of this model were in prospective studies of coronary heart disease in which a non-diseased individual's probability of developing the disease was related with risk factors x such as age, serum cholesterol and cigarette consumption (Cornfield, 1962; Truett, Cornfield and Kannel, 1967). In these applications a multivariate normal distribution for the x-variables in both the disease group and the non-disease group was assumed. Since the introduction of weighted least squares and maximum likelihood (Walker and Duncan, 1967; Day and Kerridge, 1967; Cox, 1970) the applicability of the logistic model is considerably enhanced.

This paper reviews some major developments in logistic regression methodology from 1970 onwards. It does not attempt to provide an extensive bibliography of papers on logistic regression or to present a detailed historical and methodological review. Some important recent extensions of the theory are indicated. These are roughly covered by the following three topics: inference (estimation and testing; section 2), model selection (goodness-of-fit, selection of variables; section 3) and generalizations (general linear models, polychotomous logistic regression; section 4). Some applications in medical decision making and in epidemiology are considered in the discussion (section 5).

2. Inference

2.1. Point estimation

Estimation of the parameters in the general logistic model (1.1) is usually based on the unconditional likelihood

$$\prod_{l=1}^{n_1} P(Y = 1 \mid \underline{x}_l) \prod_{l=n_1+1}^{n} P(Y = 0 \mid \underline{x}_l) \qquad (2.1)$$

where n_1 = number of observations with $Y = 1$ and $\underline{x}_l = (x_{1l}, x_{2l}, \ldots, x_{pl})$ is the vector of independent variable values of observation l ($l = 1, 2, \ldots, n$).

A technique for maximization of likelihood (2.1) which is widely applied is the Newton-Raphson method. This technique was implemented in one of the earlier computer programs for linear logistic regression analysis, the program of Lee (1974), which was distributed in the Netherlands in an adapted version (Schmitz and Hop, 1979). The same approach has been implemented in BMDP (Procedure LR, Dixon 1985). Among alternative maximization techniques leading to the same solutions, we mention the method of weighted least squares (GLIM, Baker and Nelder, 1978), and for categorical data the interative proportional fitting using the Deming-Stephan algorithm (BMDP procedure 3F, Dixon, 1985).

Several other estimators of the coefficients of the logistic regression model are described in the literature. From these we mention two methods much used in practical applications: Fisher's linear discriminant analysis and conditional maximum likelihood. Fisher's classical linear discriminant function coefficient estimates are based on the assumption that within the two groups identified by $Y = 0$ and $Y = 1$ the conditional distribution of $\underline{x}$ is multivariate normal with equal covariance matrices. Then, the maximum likelihood estimates of $\underline{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ are obtained as an analytical expression: an iterative Newton-Raphson method is not necessary. This approach became well known through applications by amongst others Cornfield (1962) and Truett, Cornfield and Kannel (1967). Many studies comparing the unconditional maximum likelihood approach and Fisher's discrimination method were focussed on the classification error rate and other performance measures for discriminatory power (Schmitz, Habbema and Hermans, 1983; Schmitz et al. 1983; Schmitz, Habbema and Hermans, 1985). However, in many applications, especially in etiologic-epidemiological studies such as case-control and cohort studies, the magnitude of the estimated coefficients themselves are of interest. Studies considering the error rate show that Fisher's linear discriminant function method is very robust, whereas studies comparing the coefficient estimates show that Fisher's method may give heavily biased estimates in the case of discrete variables (Halperin, Blackwelder and Verter, 1971). From a sampling experiment using data from a large clinical trial Hosmer, Hosmer and Fisher (1983) conclude that Fisher's discriminant function estimates of the coefficients for dichotomous variables will be seriously biased for those binary variables whose univariate odds ratios exceed two.

An alternative to unconditional maximum likelihood estimation of coefficients $(\beta_1, \beta_2, \ldots, \beta_p)$ in (1.1) is the procedure based on the conditional maximum likelihood that $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{n_1}$ are the explanatory vectors of the n_1 observations $Y = 1$, given the n vectors $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$:

$$\frac{\prod_{j=1}^{n_1} P(\underline{x}_j \mid Y=1) \prod_{j=n_1+1}^{n} P(\underline{x}_j \mid Y=0)}{\sum \left[\prod_{j=1}^{n_1} P(\underline{x}_{l_j} \mid Y=1) \prod_{j=n_1+1}^{n} P(\underline{x}_{l_j} \mid Y=0)\right]} \qquad (2.2)$$

(summation Σ is on all n over n_1 manners to divide $\{1, 2, \ldots, n\}$ in a group $\{l_1, l_2, \ldots, l_{n_1}\}$ and a group $\{l_{n_1+1}, \ldots, l_n\}$). The advantage of this method is that nuisance parameters are eliminated. Therefore, this procedure is preferred to the unconditional approach in

situations with many nuisance parameters, such as in matched case-control studies. If many nuisance parameters are involved, this may lead to biased estimators with unconditional maximum likelihood (Cox and Hinkley, 1974; Pike, Hill and Smith, 1980). However, when modelling (1.1) is performed only over a few strata and the number of parameters is relatively small, computation of (2.2) becomes intractable. Since the unconditional likelihood approach will give about the same results in such circumstances, it is the preferred method in those situations.

The estimation methods described above can be applied under several sampling schemes: sampling from the conditional distribution of Y given $\underline{x}$, sampling from the joint distribution of Y and $\underline{x}$ and sampling from the conditional distribution of $\underline{x}$ given Y (separate sampling). In the latter situation the estimate of the constant term β_0 needs to be corrected in an appropriate way (Anderson, 1972). Prentice and Pyke (1979) showed that the unconditional maximum likelihood estimates for $\beta_1, \beta_2, \ldots, \beta_p$ in separate sampling schemes are obtained as in the situation of sampling from the conditional distribution of Y given $\underline{x}$. Anderson and Blair (1982), however, do not believe that this is maximum likelihood estimation in the ordinary sense if $\underline{x}$ contains continuous variables, and suggest an alternative method, penalized maximum likelihood estimation. So far, this approach (which subtracts from the log likelihood a positive term in order to obtain a smooth estimate of the continuous density function $g(\underline{x})$) appears not to be mathematically tractable for $p \geqslant 2$.

It is well known that maximum likelihood estimates are asymptotically unbiased (Cox, 1970). For small samples (say $n < 100$) the ML estimates may be substantially biased. Several authors recommend therefore a bias correction for small sample sizes in logistic regression (Anderson and Richardson, 1979; McLachlan, 1980; Schaefer, 1983).

The existence and uniqueness of the maximum likelihood estimates in the logistic model has received relatively little attention in practical applications. Nevertheless, it is our experience that situations do occur fairly often in which the likelihood does not have a unique maximum attained for finite $\underline{\beta}$. Anderson (1972) describes an example with non-unique maxima at infinity, so-called "complete separation": all observations $\underline{x}$ from group $Y = 1$ lying on one side of a hyperplane and the observations $\underline{x}$ from group $Y = 0$ lying on the other side. In a later paper, Anderson (1974) gives an example of "zero marginal proportions": a binary variable $x_1 = 0$ for all x_1 in group $Y = 1$ and $x_1 \neq 0$ for at least one x_1 from group $Y = 0$) in which the estimate of coefficient β_1 is $-\infty$. In discrimination studies the method still yields good discriminators when "complete separation" occurs, while for "zero marginal proportions" a heuristic solution based on a mixed logistic-independence model is quite acceptable (Anderson, 1974). In other situations (epidemiological applications) the solution of the problem is less clear. Albert and Anderson (1984) describe general rules for identifying infinite parameter estimates in the logistic model.

2.2 Hypothesis testing

Once the maximum likelihood estimates of the model parameters have been obtained, testing of hypotheses concerning these parameters is to be considered. Of the several procedures that exist, the following three asymptotically equivalent methods are frequently applied: likelihood ratio (LR) tests, Wald-tests and efficient scores tests (see e. g. Kleinbaum et al., 1982).

Suppose we have $\underline{\beta}' = (\underline{\beta}_1', \underline{\beta}_2')$, where the dimension of $\underline{\beta}_2$ is $r < p$, and we wish to test the null hypothesis $H_0 : \underline{\beta}_2 = \underline{0}$. Let $L(\underline{\beta})$ be the likelihood function for a model with parameter vector $\underline{\beta}$,

$$\underline{U}(\underline{\beta}) = \partial / \partial \underline{\beta} [\ln L(\underline{\beta})]$$

the efficient scores vector,

$$\underline{I}(\hat{\underline{\beta}}) = -\partial / \partial \underline{\beta} [\underline{U}(\underline{\beta})]_{\underline{\beta} = \hat{\underline{\beta}}} = \hat{\underline{V}}^{-1}(\hat{\underline{\beta}})$$

the observed information matrix, and $\hat{\underline{V}}_{22}(\hat{\underline{\beta}})$ the lower right rxr submatrix of $\hat{\underline{V}}(\hat{\underline{\beta}})$. Then, it can be shown that each of the following test statistics is asymptotically χ^2_r under $H_0 : \underline{\beta}_2 = \underline{0}$:

(1) The likelihood ratio statistic

$$LR = -2 \ln [L(\hat{\underline{\beta}}_1^{(0)}, \underline{0}) / L(\hat{\underline{\beta}}_1, \hat{\underline{\beta}}_2)] \tag{2.3}$$

where $\hat{\underline{\beta}}_1^{(0)}$ is the ML estimator of $\underline{\beta}_1$ under $H_0 : \underline{\beta}_2 = \underline{0}$, and $(\hat{\underline{\beta}}_1', \hat{\underline{\beta}}_2')$ is the ML estimator of $\underline{\beta}' = (\underline{\beta}_1', \underline{\beta}_2')$.

(2) The Wald statistic

$$W = \hat{\underline{\beta}}_2' (\hat{\underline{V}}_{22}^{-1}(\hat{\underline{\beta}}) \hat{\underline{\beta}}_2 \tag{2.4}$$

(3) The efficient scores statistic

$$S = \underline{U}'(\hat{\underline{\beta}}_1^{(0)}, \underline{0}) \underline{I}^{-1}(\hat{\underline{\beta}}_1^{(0)}, \underline{0}) \underline{U}(\hat{\underline{\beta}}_1^{(0)}, \underline{0}) \tag{2.5}$$

In the special case $r = 1$ (thus $\underline{\beta}_2 = \beta_p$) the Wald statistic simplifies to:

$$W = Z^2 = \hat{\beta}_p^2 / \hat{Var}(\hat{\beta}_p).$$

Further, for categorical data the efficient score test is identical to the Mantel-Haenszel test (Day and Byar, 1979).

The test statistics (1) - (3) are applicable to both unconditional and conditional likelihood functions. In variable selection algorithms usually Wald's test is used to exclude variables, because once parameters have been estimated, Wald's test can be used to test the significance of a coefficient of an included variable without additional estimation. On the other hand, the score test is used to include variables since it does not require parameter estimation for testing whether a variable should be included. However, in section 3 warnings are given against cumputerized selection of variables.

As long as the logistic model (1.1) holds and the sample size is not too small, the three tests will produce valid and approximately equal results. A discrepancy between Wald's test and either the likelihood ratio or the efficient scores test is often a simple indication of a need to reparameterize the model (Thomas, 1981; Lustbader et al., 1984). This happens because the

asymptotic chi-square distribution of W depends on the multivariate normality of $\hat{\underline{\beta}}$ whereas that of LR and S does not (Thomas, 1981). In such situations Wald's test should be used with circumspection (Lustbader et al., 1984).

For small sample sizes an exact UMPU test may be used (Cox, 1970, p. 46; Bayer and Cox, 1979).

2.3. Interval estimation

An approximate 95 % confidence interval for parameter β_j in the logistic model (1.1) is:

$$\hat{\beta}_j \pm 1.96 \sqrt{\hat{V}_j(\hat{\underline{\beta}})} \tag{2.6}$$

where $\hat{V}_j(\hat{\underline{\beta}})$ is the appropriate diagonal element of $\hat{\underline{V}}(\hat{\underline{\beta}})$, the inverse of the observed information matrix.

Often, confidence limits of (a function of) a linear function of the estimated β_j's are required, for example for odds ratios or for the probability (1.1) itself. These limits are easily derived from the limits of the linear function $\Sigma\hat{\beta}_j x_j = \underline{\beta}'\underline{x}$:

$$\hat{\underline{\beta}}'\,\underline{x} \pm 1.96\,\{\underline{x}'\hat{\underline{V}}(\hat{\underline{\beta}})\underline{x}\}^{1/2} \tag{2.7}$$

Alternative ways of finding approximate confidence intervals are described by Cox (1970, p. 88) and Thomas (1981, p. 676). Large sample confidence bands for the logistic response curve (1.1) have been derived by Hauck (1983) and Brand, Pinnock and Jackson (1973).

3. Model selection

Inferences based on the logistic model may, just as inferences using other probabilistic models, only be trusted if the model fits the data adequately. Therefore, methods for examining goodness of fit are important. These methods should be used with each model selection procedure or analysis strategy. Before discussing model selection procedures we will describe some goodness of fit statistics for logistic regression models, and we will discuss some methods for detection of outliers, so-called influential cases and other model departures (residual analysis and regression diagnostics).

3.1. Goodness of fit statistics

Two goodness of fit statistics are useful for grouped data, provided the number of $\underline{x}$ values is sufficient small compared with the total sample size. Then, we may use Pearson's X^2-statistic

$$X^2 = \Sigma \frac{(O - E)^2}{E} \tag{3.1}$$

or the likelihood ratio statistic

$$G = 2 \Sigma O \ln(O/E) \qquad (3.2)$$

(see e.g. Breslow and Day, 1980, p.208). Summation is over all cells formed by the combination $(Y, \underline{x})$, O and E are the observed and expected frequencies in the respective cells ($E = N(\hat{P})$ if $Y = 1$ and $E = N(1 - \hat{P})$ if $Y = 0$). Both these statistics have asymptotic chi-square distributions under the null hypothesis with degrees of freedom equal to the number of cells less the number of parameters in the logistic model.

For the general situation of mixed discrete and continuous variables $\underline{x}$ the three asymtotic test statistics for $H_0 : \underline{\beta}_2 = \underline{0}$ (section 2.2) are often seen to be erroneously used as a goodness of fit statistic. These statistics have to be viewed as tests for in- or exclusion of variables in the logistic model rather than as goodness of fit tests per se.

Several alternative statistics have recently been described for assessment of goodness of fit of the general logistic model (1.1) (Hosmer and Lemeshow, 1980). The following one, proposed by Lemeshow and Hosmer (1982) is recommended by these authors. Let $P(\underline{x}_i)$ be the estimated probability (1.1) for subject i ($i = 1, 2, \ldots, n$). Given the values of $P(\underline{x}_i)$ these n values are classified to ten fixed deciles for the values of $P(\underline{x}_i)$: [0 - 0.1), [0.1 - 0.2), . . ., [0.9 - 1.0]. Within each decile l ($l = 1, 2, \ldots, 10$) the number o_{kl} of observed cases with $Y = k$ ($k = 0,1$) is counted and the estimated expected frequency e_{kl} is calculated using

$$e_{1l} = \Sigma P(\underline{x}_i), \; e_{0l} = \Sigma(1 - P(\underline{x}_i))$$

(summation over subjects within decile l). Then, an appropriate goodness of fit statistic which compares observed and estimated expected frequencies is

$$H = \sum_{k=0}^{1} \sum_{l=1}^{10} \frac{(o_{kl} - e_{kl})^2}{e_{kl}} \qquad (3.3)$$

distributed approximately chi-square with 8 degrees of freedom (H corresponds with H_g^*, expression (11), in Lemeshow and Hosmer (1982)).

A useful goodness of fit statistic presented by Brown (1982), is based on the adequacy of fit for the logistic model as a special case of a more general family of models. This statistic has poor power against certain patterns (such as excessive deviations) outside this general class of models (Brown, 1982).

3.2. Residual analysis and regression diagnostics

Reduction of a sample of observations to a single test statistic is not an entirely adequate method of assessing goodness of fit. The plotting and examination of residuals is a useful additional approach to detect certain departures from the fitted model. Recently, several functions of the data, called "regression diagnostics", have been proposed for detecting cases that could have an unwarranted influence on the coefficient estimates. A key paper is that of Landwehr, Pregibon and Shoemaker (1984). This paper proposes and discusses three

graphical methods to detect several kinds of lack of fit of the logistic regression model. First, the local mean deviance plot, especially useful for detecting a missing interaction term. Second, the empirical probability plot, a sophisticated graphical way of comparing observed and expected proportions in order to detect outliers and some general model departures. And third, the partial residual plot, especially useful for misspecifications involving a specific variable (e. g. $\ln(x_1)$ instead of x_1). Most of these approaches are modifications and extensions of similar methods for the ordinary linear regression model, and are computationally intensive. In the discussions following the paper of Landwehr, Pregibon and Shoemaker (1984) several discussants showed that some plots would not work correctly in all situations (e. g. the partial residual plot for probabilities near zero or one), and the general impression is that much work still needs to be done in this field.

3.3. Model selection procedures

Several methods of variable selection in logistic regression have been proposed. Most of them are based on similar principles as in ordinary linear regression. Forward selection, backward elimination and all subset regressions are much used approaches. For each set of selected variables a time-consuming iterative procedure must be carried out in order to obtain the ML estimates of the coefficients in the model. A procedure that automatically selects variables in logistic regression will therefore take much more computer time than the corresponding stepwise ordinary linear regression procedure where parameter estimates are obtained by solving a single system of linear equations. The following algorithm suggested by Schoenfeld (1982) is modeled after the algorithm given by Peduzzi, Hardy and Holford (1980). It requires a minimal amount of computer time. Define the following two sets: IN_k is the set of variables in the model at step k, OUT_k is the set of variables not in the model at step k. The algorithm is initialized by letting IN_1 contain all variables that are known to be important, and OUT_1 be all other variables. Different tests are used for including and excluding variables from the model. To exclude variables Wald's test is used because once parameters have been estimated, this test can be used to test whether an included variable has significant coefficients without additional estimation. To include variables the efficient scores test is used since it does not require additional parameter estimation to test whether a variable should be included. Thus, at each iteration of the algorithm only one maximization of the likelihood is performed.

Stepwise methods have many limitations. When applying logistic regression in discriminant analysis problems the criterion of primary interest is not the significance of a coefficient β but a measure of discriminatory performance like the error rate, a utility function or similar measures (see e. g. Habbema, Hilden and Bjerregaard, 1981). It would therefore be a natural choice then to take such performance measure as criterion for variable selection.

In observational studies, e. g. case-control studies, the use of variable selection algorithms as proposed by Peduzzi, Hardy and Holford (1980) may result in highly misleading interpretations and conclusions. Also in these fields the algorithms are generally focussed on the wrong criterion like significance testing or precision considerations. Instead, unbiasedness of estimates of the coefficients in the logistic model are extremely important in case-control investigations. A fairly general analysis strategy in observational studies is outlined by Schmitz (1982). The essence of the latter approach is based on considerations by Cox and Snell (1974) that lead to selection in three phases. First, the variables concerned are divided

into subgroups of homogeneous batteries of variables. Second, within each subgroup the best fitting model is searched. Finally the variables selected within each battery are combined. The "best fitting model", may then be obtained using a limited number of variables (phases 2 and 3 in the approach above) with a strategy described by Kleinbaum, Kupper and Chambless (1982). Its first step, after specifying the "important" variables, consists of the assessment of significant interaction terms (interaction with a prespecified exposure factor). Next, the possible confounding effects of variables which are not effect modifiers (i.e. which show no interaction with the prespecified exposure factor) are investigated. There are many other strategies that can be followed. The main criterion in epidemiologic applications, however, is to obtain valid estimates of the coefficients of the exposure factors.

4. Generalizations

So far the logistic model (1.1) with a binary response variable Y has been considered. An obvious generalization is the polychotomous regression model with $T + 1$ ($T > 1$) outcome categories $Y = 0, 1, 2, \ldots, T$ where $Y = 0$ represents the baseline category. We use the subscript i ($i = 1, 2, \ldots, n$) to denote each of the n cases, and the subscript d ($d = 0, 1, 2, \ldots, T$) to denote the $T + 1$ outcome categories. Then, $P_{di} = P(Y = d \mid x_{1i}, x_{2i}, \ldots, x_{pi})$ represents the probability that case i belongs to the dth outcome category, and the polychotomous logistic model is defined by:

$$\log P_{di}/P_{0i}) = \beta_{d0} = \beta_{d1}x_{1i} + \beta_{d2}x_{2i} + \ldots + \beta_{dp}x_{pi}$$

$$\stackrel{def}{=} y_{di},\ d = 1, 2, \ldots, T$$

$$\text{with} \sum_{d=0}^{T} P_{di} = 1,\ i = 1, 2, \ldots, n. \tag{4.1}$$

Using $Y = 0$ as baseline category it is easily shown that:

$$P_{0i} = 1/(1 + \sum_{d=1}^{T} \exp(y_{di}))$$

$$P_{di} = \exp(y_{di}) / (1 + \sum_{r=1}^{T} \exp(y_{ri})),\ d = 1, 2, \ldots, T \tag{4.2}$$

Recently, the practical implementation of the polychotomous logistic model has been extensively discussed by Wijesinha et al. (1983) en Begg and Gray (1984). They discuss some problems that occur with this model: large numbers of parameters to be estimated simultaneously, scarcely available statistical software and frequent occurrence of "zero marginal proportions" (see also section 2.1). In order to overcome these difficulties they propose a so-called individualized binary logistic regression approach, based on a one-to-one correspondence between the parameters of the polychotomous model and the parameters of simple logistic regression (Wijesinha et al., 1983, appendix B). A limitation of this approach

is that it does not take into account that the separate analyses are not independent because the baseline category cases are common to all of them. Consequently, contruction of tests concerning parameters over three or more outcome categories simultaneously is not feasible. Applications of the individualized approach in discriminant analysis are described by Wijesinha et al. (1983) and Schmitz, van de Merwe and Habbema (1986). Marshall and Chisholm (1985) discuss a "reduced" logistic model for hypothesis testing in polychotomous logistic regression. Polychotomous logistic models for an ordered outcome variable have been considered especially by Anderson and Philips (1981) and Anderson (1984). Greenland (1985) describes an application of the logistic model for an ordinal response. He cautions for the rather strong assumptions these models embody, and points out that unconditional maximum likelihood methods may have poor small-sample properties here.

Several families of models are described in the statistical literature of which the logistic model may be viewed as a special case. Without doubt the general linear model (GLM, see e. g. Nelder and Wedderburn, 1972, and McCullagh and Nelder, 1983) is the most extensively documented and applied model. Aranda-Ordaz (1981) proposes a family of symmetric transformations for the probability $P(Y = 1 \mid \underline{x})$ to the linear function $\underline{\beta}'\underline{x}$ which contains a parameter λ that denotes the transformation parameter. This class reduces to the logistic model if $\lambda = 0$ and to the linear model when $\lambda = 1$. Actually, both the transformation parameter and the model parameters $\underline{\beta}$ may be estimated by maximum likelihood. A comparable approach is proposed by Guerrero and Johnson (1982). The proposed models may be fitted using the GLIM package (Baker and Nelder, 1978) for each fixed value of the transformation parameter.

Thomas (1981) describes a generalization of the logistic model for application in survival analysis and matched case-control studies. This generalization is obtained by writing the conditional likelihood as a function of $\exp(\underline{\beta}'\underline{x})$. With this approach a flexible comparison of additive and multiplicative models is possible. For the situation of only one risk factor, additive (linear) multiplicative (logistic) models for application in case-control studies are compared by Schmitz (1986) by means of simulation and by applying the methods to an empirical data set.

5. Discussion

The binary logistic regression model is now widely used in medical research, and innumerable many applications have been described in the literature. In the field of medical decision making logistic regression is extensively employed as discriminant analysis method. Anderson (1972, 1974, 1975, 1982) has contributed a great deal to developments in logistic discrimination. We applied logistic discriminant analysis to several medical problems (Van de Merwe, Schmitz, Wensinck, 1981; Schmitz, Habbema and Hermans, 1983; Schmitz, van de Merwe and Habbema, 1985) and compared the performance of logistic discrimination with some alternative discrimination procedures for mixtures of continuous and discrete data (Schmitz et al., 1983; Schmitz, Habbema and Hermans, 1986). It appears that a good performance is obtained when the so called "augmented logistic apprach" is used: logistic discrimination with a linear function of explanatory variables augmented with appropriate interaction terms.

In epidemiologic investigations, especially case-control studies, the logistic model has proven to be a powerful method for risk estimation in the presence of confounding factors. Especially if many confounding factors must be controlled simultaneously, classical stratification methods will break down. The monograph of Breslow and Day (1980) greatly stimulated the use of logistic regression in case-control studies. We applied their computer program MATCH, based on conditional maximum likelihood (Breslow and Day, 1980, appendix IV) for analysis of matched case-control data relating stroke risk to several risk factors in Tilburg, The Netherlands (Herman et al., 1983).

The multivariate analysis of survival data is usually based on proportional hazards models (Cox, 1972) which incorporate the time element of event occurrences. Consequently logistic regression has played a less important role in survival analysis. Some recent papers, however, show that the two models will lead to similar conclusions (a) in studies with short lengths of follow-up and slow rates of disease occurrence (Green and Symons, 1983) and (b) in the setting of grouped event times (Abbott, 1985).

Of course, there are numerous other applications of the logistic model in medically oriented research, besides discriminant analysis, epidemiology and survival analysis. To mention a few examples: bioassay (McLeish and Tosh, 1983), cross-over designs in clinical trials (Le, 1984) and quantitative structure-activity relationships for biological activity (Schmitz and Habbema, 1985).

The logistic model has proven to be widely applicable, but it has its hazards and limitations too. For example, just as in ordinary linear regression, with highly correlated explanatory variables (multicollinearity) problems may occur (Gordon, 1974), and results have to be interpreted cautiously then. Gordon (1974) also asserts that departures from the linear logistic model (such as the presence of interaction or the conditional probability not being monotonic in the independent variable) should present problems. However, regression diagnostic procedures can be used to help reveal deviations from the model and possibly to help improve the fit. In addition it should be stressed that particularly in epidemiologic studies simple techniques which incorporate stratification, such as the Mantel-Haenszel approach, should always be used in combination with the multivariate modelling procedures, in order to reach a practical interpretation of the results. In discriminant analysis applications the results should always be validated with new data.

The techniques of logistic regression diagnostics are still developing, and as with other statistical techniques, the applied statistician has to wait for the useful techniques to filter through, resulting in implementation in general available computer packages. In the forthcoming years statistical analyses using the logistic model will certainly improve in quality by further developments of its theory. Meanwhile the model will undoubtedly maintain its prominent place among the nowadays available statistical tools.

REFERENCES

Abbott, R. D.: Logistic regression in survival analysis. Amer. J. Epid. 121 (1985) 465 - 471.

Albert, A. and Anderson, J. A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrika **71** (1984) 1 - 10.

Anderson, J. A.: Separate sample logistic discrimination. Biometrika **59** (1972) 19 - 35.

Anderson, J. A.: Diagnosis by logistic discriminant function: Further practical problems and results. Appl. Statist. **23** (1974) 397 - 404.

Anderson, J. A.: Quadratic logistic discrimination. Biometrika **62** (1975) 149 - 154.

Anderson, J. A.: Logistic discrimination. Chapt. 7 in: Krishnaiah, P. R. and Kanal, L. N., Handbook of Statistics **2** (1982). North-Holland Publ. Comp., Amsterdam.

Anderson, J. A.: Regression and ordered categorical variables. J. R. Statist. Soc. B. **46** (1984) 1 - 30.

Anderson, J. A. and Blair, V.: Penalized maximum likelihood estimation in logistic regression and discrimination. Biometrika **69** (1982) 123 - 136.

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

Anderson, J. A. and Richardson, S. C.: Logistic discrimination and bias correction in maximum likelihood estimation. Technometrics **21** (1979) 71 - 78.

Aranda - Ordaz, F. J.: On two families of transformations to additivity for binary response data. Biometrika **68** (1981) 357 - 363.

Baker, R. J. and Nelder, J. A.: The GLIM system release 3 generalized linear interactive modelling (1978). Numerical Algorithms Group, Oxford, England.

Bayer, L. and Cox, C.: Exact tests of significance in binary regression models. Algorithm AS 142, Appl. Stat. **28** (1979) 319 - 324.

Begg, C. B. and Gray, R.: Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika **71** (1984) 11 - 18.

Brand, R. J., Pinnock, D. E. and Jackson, K. L.: Large sample confidence bands for the logistic response curve and its inverse. The Amer. Stat **27** (1973) 157 - 160.

Breslow, N. E. and Day, N. E.: Statistical methods in cancer research. Vol. I. The analysis of case-control studies (1980). IARC Scientific Publications No. 32, IARC, Lyon.

Brown, C. C.: On a goodness of fit test for the logistic model based on score statistics. Commun. Statist.-Theor. Meth. **11** (1982) 1087 - 1105.

Cox, D. R.: The analysis of binary data (1970). Methuen, London.

Cox, D. R.: Regression models and life tables (with discussion). J. R. Stat. Soc. B. **34** (1972) 187 - 220.

Cox, D. R. and Hinkley, D. V.: Theoretical Statistics (1974). Chapman and Hall, London.

Cox, D. R. and Snell, E. J.: The choice of variables in observational studies. Appl. Stat. **23** (1974) 51 - 59.

Cornfield, J.: Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. Fed. Proc. **21** (1962) 58 - 61.

Day, N. E. and Byar, D. P.: Testing hypotheses in case-control studies – equivalence of Mantel - Haenszel statistics and logit score tests. Biometrics **35** (1979) 623 - 630.

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

Dixon, W. J.: BMDP statistical software (1985). University of California Press, Berkley CA.

Gordon, T.: Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies. J. Chron. Dis. **27** (1974) 97 - 102.

Green, M. S. and Symons, M. J.: A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. J. Chron. Dis. **36** (1983) 715 - 724.

Greenland, S.: An application of logistic models to the analysis of ordinal responses. Biom. J. **27** (1985) 189 - 197.

Guerrero, V. M. and Johnson, R. A.: Use of the Box - Cox transformation with binary response models. Biometrika **69** (1982) 309 - 314.

Habbema, J. D. F., Hilden, J. and Bjerregaard, B.: The measurement of performance in probabilistic diagnosis. V. General recommendations. Meth. of Inf. in Med. **20** (1981) 97 - 100.

Halperin, M., Blackwelder, W. C. and Verter, J. I.: Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood. J. Chron. Dis. **24** (1971) 125 - 158.

Hauck, W. W.: A note on confidence bands for the logistic response curve. The Amer. Stat. **37** (1983) 158 - 160.

Herman, B., Schmitz, P. I. M., Leyten, A. C. M., Luijk, J. H. van, Frenken, C. W. G. M., Op de Coul, A. A. W. and Schulte, B. P. M.: Multivariate logistic analysis of risk factors for stroke in Tilburg, The Netherlands. Amer. J. Epid. **118** (1983) 514 - 525.

Hosmer, D. W. and Lemeshow, S.: Goodness of fit tests for the multiple logistic regression model. Commun. Statist. - Theor. Meth. **9** (1980) 1043 - 1069.

Hosmer, T. A., Hosmer, D. W. and Fisher, L.: A comparison of three methods of estimating the logistic regression coefficients. Commun. Statist. - Simula. Computa. **12** (1983) 577 - 593.

Kleinbaum, D. G., Kupper, L. L. and Chambless, L. E.: Logistic regression analysis of epidemiologic data: Theory and practice. Commun. Statist. - Theor. Meth. **11** (1982) 485 - 547.

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C.: Graphical methods for assessing logistic regression models. JASA **79** (1984) 61 - 71.

Le, C. T.: Logistic models for cross-over designs. Biometrika **71** (1984) 216 - 217.

Lee, E. T.: A computer program for linear logistic regression analysis. Comp. progr. in biomed. **4** (1974) 80 - 92.

Lemeshow, S. and Hosmer, D. W.: A review of goodness of fit statistics for use in the development of logistic regression models. Amer. J. Epid. **115** (1982) 92 - 106.

Lustbader, E. D., Moolgavkar, S. H. and Venzon, D. J.: Tests of the null hypothesis in case-control studies. Biometrics **40** (1984) 1017 - 1024.

Marshall, R. J. and Chisholm, E. M.: Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer. Statistics in Medicine **4** (1985) 337 - 344.

McCullagh, P. and Nelder, J. A.: Generalized linear models (1983). Chapman and Hall, London.

McLachlan, G. J.: A note on bias correction in maximum likelihood estimation with logistic discrimination. Technometrics **22** (1980) 621 - 627.

McLeish, D. and Tosh, D.: The estimation of extreme quantiles in logit bioassay. Biometrika **70** (1983) 625 - 632.

Merwe, J. P. van de, Schmitz, P. I. M. and Wensinck, F.: Antibodies to Eubacterium and Peptostreptococcus species and the estimated probability of Crohn's disease. J. Hyg., Camb. **87** (1981) 25 - 33.

Nelder, J. A. and Wedderburn, R. W. M.: Generalized linear models. J. R. Statist. Soc. A. **135** (1972) 370 - 384.

Peduzzi, P. N., Hardy, R. J. and Holford, T. R.: A stepwise variable selection procedure for nonlinear regression models. Biometrics **36** (1980) 511 - 516.

Pike, M. C., Hill, A. P. and Smith, P. G.: Bias and efficiency in logistic analyses of stratified case-control studies. Int. J. Epid. **9** (1980) 89 - 95.

Prentice, R. L. and Pyke, R.: Logistic disease incidence models and case-control studies. Biometrika **66** (1979) 403 - 411.

Schaefer, R. L.: Bias correction in maximum likelihood logistic regression. Statistics in Medicine **2** (1983) 71 - 78.

Schmitz, P. I. M.: Univariate dose-response models in case-control studies. Biom. J. **28** (1986). In press.

Schmitz, P. I. M.: Selection of variables in etiologic-epidemiological studies. T. soc. Geneesk. **60** (1982) 851 -854 (In Dutch).

Schmitz, P. I. M. and Habbema, J. D. F.: Logistic discriminant analysis for modelling Quantitative Structure-Activity Relationships (QSAR). (1986) Submitted.

Schmitz, P. I. M., Habbema, J. D. F. and Hermans, J.: The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. Statistics in Medicine **2** (1983) 199 - 205.

Schmitz, P. I. M., Habbema, J. D. F. and Hermans, J.: A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables. J. Statist. Comput. Simul. **23** (1985) 69 - 95.

Schmitz, P. I. M., Habbema, J. D. F., Hermans, J. and Raatgever, J. W.: Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables. Commun. Statist. - Simula. Computa. **12** (1983) 727 - 751.

Schmitz, P. I. M. and Hop W. C. J.: Addendum to "Lee, E. T. A computer program for linear logistic regression analysis. Comp. progr. biom. **4** (1974) 80 - 92" (1979). Institute of Biostatistics. Erasmus University Rotterdam (In Dutch).

Schmitz, P. I. M., van de Merwe, J. P. and Habbema, J. D. F.: Construction and validation of a diagnostic support model for the diagnosis of Crohn's disease by agglutination tests. (1986) Submitted.

Schoenfeld, D. A.: Analysis of categorical data: Logistic models. Chapt. 14 in: Mike, V. and Stanley, K. E., eds. Statistics in medical research (1982). Wiley, New York.

Thomas, D. C.: General relative-risk models for survival time and matched case-control analysis. Biometrics **37** (1981) 673 - 686.

Truett, J., Cornfield, J. and Kannel, W.: A multivariate analysis of the risk of coronary heart disease in Framingham. J. Chron. Dis. **20** (1967) 511 - 524.

Walker, S. H. and Duncan, D. B.: Estimation of the probability of an event as a function of several independent variables. Biometrika **54** (1967) 167 - 179.

Wijesinha, A., Begg, C. B., Funkenstein, H. H. and McNeil, B. J.: Methodology for the differential diagnosis of a complex data set. A case study using data from routine CT scan examinations. Med. Decis. Making **3** (1983) 133 - 154.

CHAPTER 2

CONSTRUCTION AND VALIDATION OF A DIAGNOSTIC SUPPORT MODEL FOR THE DIAGNOSIS OF CROHN'S DISEASE BY AGGLUTINATION TESTS

Construction and Validation of a Diagnostic Support Model for the Diagnosis of Crohn's Disease by Agglutination Tests. Paul I. M. Schmitz, Joop P. van de Merwe and J. Dik F. Habbema. Submitted for publication, October 1985.

CONSTRUCTION AND VALIDATION OF A DIAGNOSTIC SUPPORT MODEL FOR THE DIAGNOSIS OF CROHN'S DISEASE BY AGGLUTINATION TESTS

by

Paul I.M. Schmitz *, Joop P. van de Merwe **
and J. Dik F. Habbema ***

SUMMARY

The use of four agglutination tests in the differential diagnosis of Crohn's disease is studied. The test outcomes of 164 patients with Crohn's disease, 46 patients with ulcerative colitis and 188 healthy subjects are analyzed by logistic discrimination. Patients with ulcerative colitis could not be distinguished from healthy subjects, but differentiation was possible between Crohn's disease and non-Crohn's disease. With these training cases a positive predictive value was obtained of 95% and a negative predictive value of 70% (for priors proportional to sample sizes: probability of Crohn's disease is 0.41).

The logistic approach showed equal discriminatory power but better reliability (goodness-of-fit) than Fisher's linear discrimination and an independence model. Validation with sera from new patients, both from the Netherlands and from an international data set, resulted in a decreased performance. This decrease can probably be partly ascribed to the poor definition of diagnostic categories for the validation cases.

Keywords: Discriminant analysis, Mixtures of variables; Logistic discrimination; Diagnostic support model; Crohn's Disease.

* Institute of Biostatistics, Erasmus University, P.O.Box 1738, 3000 DR Rotterdam, The Netherlands

** Department of Medical Microbiology, Erasmus University, Rotterdam, The Netherlands

*** Department of Public Health and Social Medicine, Erasmus University, Rotterdam, The Netherlands

1. INTRODUCTION

The incidence of inflammatory bowel diseases (IBD), a generic term given to two major diseases (ulcerative colitis and Crohn's disease), has increased worldwide and has emerged as an important biomedical problem. (Kirsner and Shorter (1)). The diagnosis of inflammatory bowel diseases can only be made after exclusion of all other causes of colonic and/or ileal disease. A correct diagnosis is important as the treatment of the former group is markedly different from that of the latter (Johnson et al.(2)). The mean delay between onset of symptoms and diagnosis of IBD is about four years (Dyer and Dawson (3), Brandes and Eulenburg (4) and Mekhjian et al. (5)). Therefore, a diagnostic test for IBD, which may serve as at least a catalyst in the diagnostic process, would be of value.

Wensinck and van de Merwe (6) investigated sera from patients with Crohn's disease (CD), ulcerative colitis (UC), other diseases (diseased controls, DC) and healthy subjects (HS) on the presence of antibodies to four strains of gram-positive coccoid anaerobes using agglutination tests. Two strains (Me44 and Me47) were identified as Eubacterium contortum, one (Me46) as Coprococcus comes (previously identified as Eubacterium rectale) and one (C18) as Peptostreptococcus productus. Two drops of serum and one drop of bacterial suspension were thoroughly mixed. Results were scored after 5 minutes as negative (0) or positive (1,2 or 3 according to strength of agglutination).The results showed that the four strains were agglutinated more frequently by sera from patients with Crohn's disease than by sera from healthy subjects. Patients with ulcerative colitis only had a higher percentage of positive agglutination for strains Me46 and C18. Among the 'control' diseases of the intestinal tract and liver, carcinoma of large bowel showed an arised percentage of positive agglutinations for Me47 and Me46, while cirrhosis of liver had a higher percentage positive for strain C18.

In order to study the possibilities of using the agglutination tests for differentiation between patients with either Crohn's disease or ulcerative colitis and healthy subjects, the agglutination tests with sera from Dutch patients from three different populations: CD, UC and HS, will be analyzed by logistic discriminant analysis (section 2). Some alternative statistical approaches to clinical decision-support systems and a number of different measures for ascertainment of validity are described and applied to the Dutch patients in section 3. The model derived from these so called training cases

will be validated with Dutch validation cases and with data from international cases (section 4). Implementation in clinical practice is considered in section 5. Finally, the results are discussed in section 6.

2. LOGISTIC DISCRIMINATION FOR THE DUTCH TRAINING CASES

2.1 Crohn's Disease, Ulcerative Colitis and Control Subjects.

A summary of the composition of the training and test cases used in this paper is shown in Table 1. From October 1975 onwards sera obtained from patients with CD or UC in the Departments of Internal Medicine and Surgery of the University Hospital in Rotterdam, The Netherlands, were studied. The diagnosis of CD and UC was based on accepted clinical, radiological and histological criteria (Lennard-Jones et al.(7) and Kirsner(8)), without knowledge of the results of the agglutination tests. Patients in whom the differential diagnosis between CD and UC could not be made were not included. Thus, a group of 164 patients with CD and a group of 46 patients

Table 1. Distribution of the Dutch training cases, the Dutch validation cases and the international validation cases over diagnostic categories.

	Crohn's Disease CD	Ulcerative Colitis UC	Diseased Controls DC	Healthy Subjects HS	Non-Crohn's $\bar{CD}$	Total
Dutch training cases	164 (41%)	46 (12%)	0 (0%)	188 (47%)	234 (59%)	398 (100%)
Dutch validation cases	61 (19%)	29 (9%)	226 (72%)	0 (0%)	255 (81%)	316 (100%)
International validation cases	317 (41%)	232 (30%)	67 (9%)	157 (20%)	456 (59%)	773 (100%)

with UC was collected. The 'control' group consisted of 188 healthy subjects (HS), which were volunteers of the Red Cross Blood Transfusion Service Rotterdam. For each patient (CD or UC) and each healthy subject the results of the four agglutination tests as described in the introduction section were obtained.

Polychotomous logistic regression analysis (see e.g. Anderson (9)) is applied for modelling the probabilities of belonging to one of the three diagnostic categories CD, UC and HS, given the results of the four agglutination tests $x = (x_1, x_2, x_3, x_4) = (\text{Me44}, \text{C18}, \text{Me46}, \text{Me47})$. Using the healthy subjects as baseline diagnostic category, this model assumes the following relationships:

$$\log P(UC|x)/P(HS|x) \overset{\text{def}}{=} \log P_1/P_0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \overset{\text{def}}{=} z_1$$

$$\log P(CD|x)/P(HS|x) \overset{\text{def}}{=} \log P_2/P_0 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 \overset{\text{def}}{=} z_2 \qquad (2.1)$$

With the results of the agglutination reactions x_1 to x_4 of the 164 patients with CD, 46 patients with UC, and 188 healthy subjects, the estimates of the parameters β_k, γ_k (k=0,1,2,3,4) and their 95% confidence limits are obtained using the method indicated in the appendix and summarized in Table 2.

Table 2. Individualized logistic regression for the Dutch training cases: Ulcerative Colitis (UC) versus Healthy Subjects and Crohn's Disease (CD) versus Healthy Subjects.

Tests	Diagnostic Category	
	UC	CD
	$\hat{\beta}$ (95% conf.limits)	$\hat{\gamma}$ (95% conf.limits)
constant	-1.56(-1.94,-1.18)	-1.78(-2.18,-1.38)
x_1(Me44)	-0.16(-0.70, 0.38)	0.60(0.25, 0.95)
x_2(C18)	0.35(-0.21, 0.91)	0.69(0.31, 1.07)
x_3(Me46)	1.94(0.63, 3.26)	2.43(1.23, 3.64)
x_4(Me47)	-0.73(-2.18, 0.72)	1.13(0.53, 1.73)

Calculations for the posterior probabilities P(CD |x), P(UC |x) and P(HS |x) are implicitly based on prior probabilities proportional to the sample sizes:

P(CD)= 164/398 = 0.41, P(UC)= 46/398 = 0.12, P(HS)= 188/398 = 0.47 (2.2)

For each case of the Dutch training cases, the three posterior probabilities were calculated and set out in a triangular diagram (figures 1.a, 1.b and 1.c),

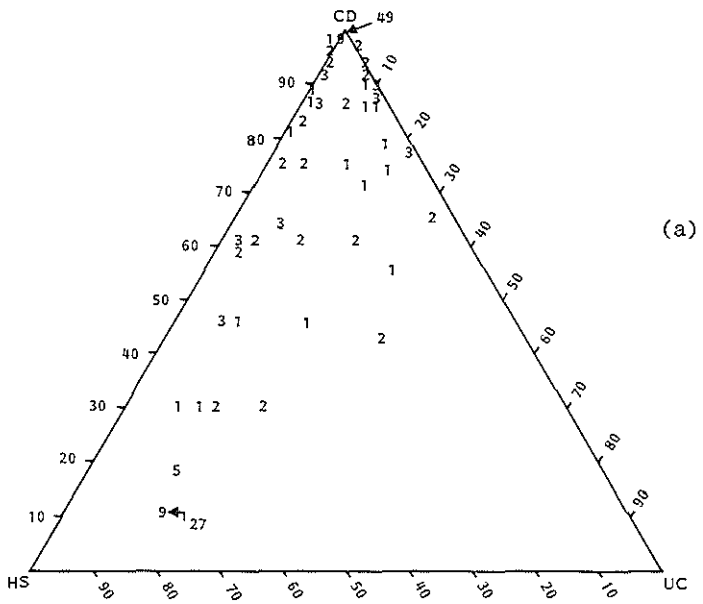


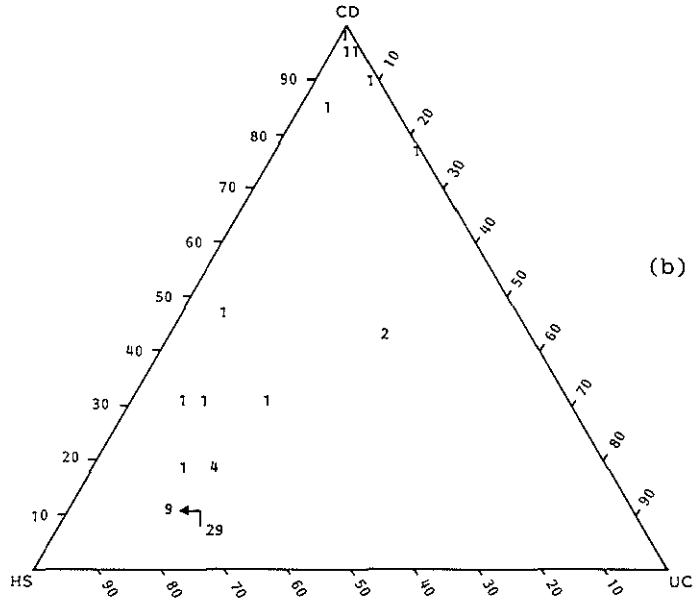
Figure 1. Posterior probabilities for the Dutch training cases, with prior probabilities proportional to the number of training cases: P(CD)=.41, P(HS)=.47, P(UC)=.11.

(a) 164 patients with Crohn's disease

(b) 46 patients with ulcerative colitis

(c) 188 healthy subjects

(9 means 9 or more points: indicated with an arrow)



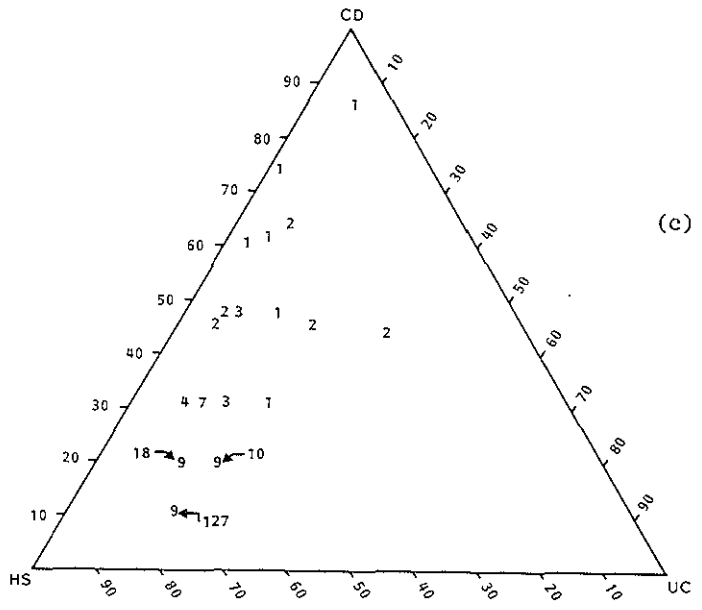


Figure 1. (Continued)

see Habbema et al. (12). The triangle is equilateral with one vertex for each diagnostic category (CD, UC and HS) and has altitude 1. Each case with probabilities (P(CD|x), P(UC|x), P(HS|x)) is a point in this diagram whose distance to the sides equals the disease probability concerned, e.g. in fig.1.a: the distance to the side opposite vertix HS (it is the side CD-UC) indicates the posterior probability P(HS|x). Analogously, posterior probabilities based on

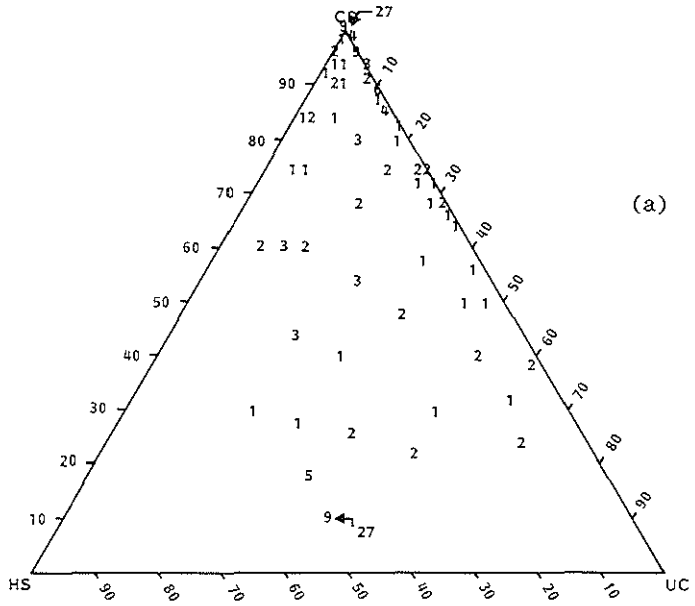


Figure 2. Posterior probabilities for the Dutch training cases, with prior probabilities equal: P(CD) = P(HS) = P(UC) = .33.
(a) 164 patients with Crohn's disease
(b) 46 patients with ulcerative colitis
(c) 188 healthy subjects
(9 means 9 or more points: indicated with an arrow)

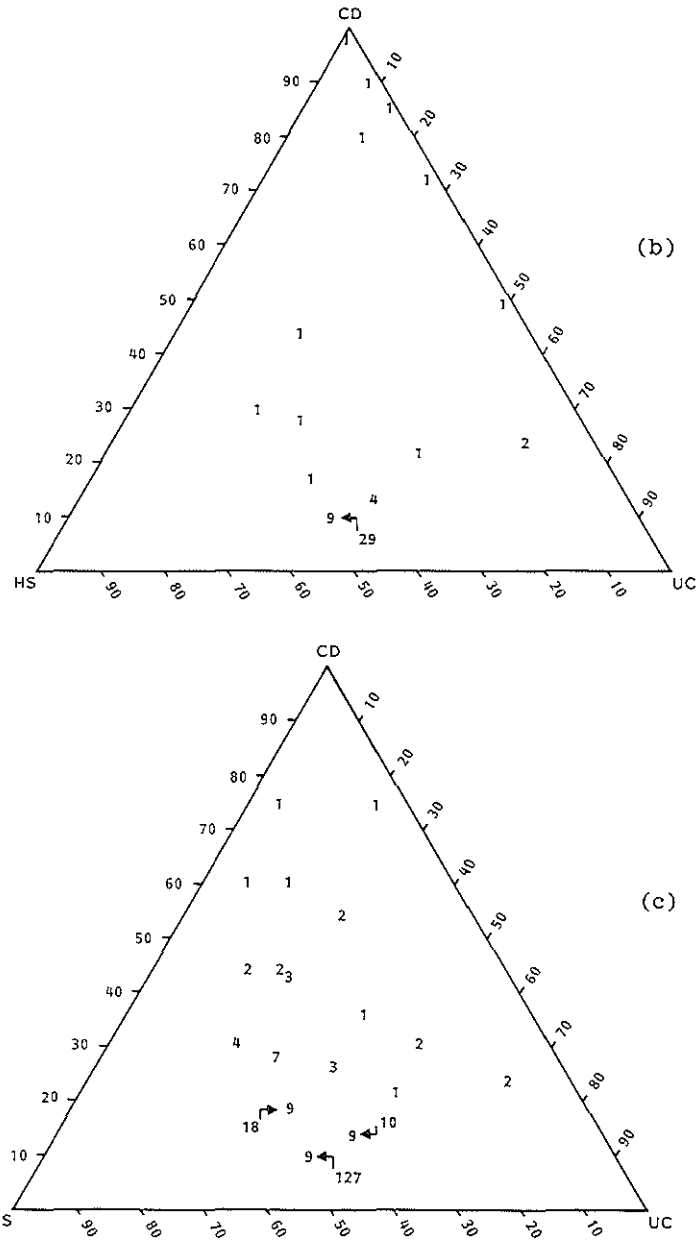


Figure 2. (Continued)

equal prior probabilities 0.33 have been set out in figures 2.a, 2.b and 2.c. Posterior probabilities for prior probabilities other than (2.2) can be calculated using (2.1) and appropriate adaption of the estimated constant terms $\hat{\beta}_0$ and $\hat{\gamma}_0$ (see appendix 1).
Thus, with our data, changing from 'size' priors to 'equal' priors, we find $\hat{\beta}_0 = -0.15$ and $\hat{\gamma}_0 = -1.64$.

A simple allocation rule for classification of the subjects with help of the posterior probabilities might be: assign each subject to the diagnostic category for which the estimated posterior probability is highest. Using this rule, it is immediately seen from figures 1.a-c that allocation to UC is never made, not even for UC patients. Of course, this may partly be due to the relative low prior of UC (12%). Therefore, more insight concerning the discriminatory ability of our diagnosis-support rule may be gained from figures 2.a-c where equal prior probabilities are assumed. Compared with figures 1.a-c the probabilities are pulled towards vertex UC. Also, based on the allocation rule and equal priors, an allocation matrix was constructed, see Table 3. Again, the overall picture is that differentation of UC from CD or HS is hardly possible. The positive predictive value for UC (i.e. the percentage of patients allocated as UC that actually has UC) is only 21%.

To solve this diagnostic problem, three possibilities may be viewed. First, we might want to exclude patients with UC before applying the agglutination tests, just as we do if carcinoma of large bowel or cirrhosis of liver is present. However, such an exclusion does not correspond with clinical

Table 3. Allocation matrix based on equal priors and assignment to the category with highest probability.

		Population of origin			
		CD	UC	HS	Total
Population of allocation	CD	119 (72%)	6 (13%)	13 (7%)	138
	UC	11 (7%)	8 (17%)	19 (10%)	38
	HS	34 (21%)	32 (70%)	156 (83%)	222
	Total	164 (100%)	46 (100%)	188 (100%)	398

practice in which differential diagnosis of inflammatory bowel diseases (CD and UC) is usually made only after exclusion of all other causes of intestinal disease. Therefore, a second approach could be considered: utilizing the agglutination tests for differentiation between IBD (unspecified) and healthy subjects. This seems to be clinically attractive, especially because further differentiation of IBD into CD or UC can be made in a subsequent diagnostic stage. Clamp et al. (13) developed a scoring system derived from clinical and investigational data. Allocations, made in 85% of the cases in their dataset, proved to be 95% accurate. However, the predictive value for IBD will be substantially less than that for CD. This can be seen from the relative frequencies of the test outcomes: the frequency distribution for patients with UC resembles more that for healthy subjects than that for patients with CD. Because a high positive predictive value of our diagnostic system with agglutination tests is a prerequisite, we finally preferred to use the tests for differentiation between CD and 'non-CD'. Here, 'non-CD' is a mixture of healthy subjects and patients with UC. A simple dichotomous logistic model for the discrimination of CD and non-CD($\bar{C}\bar{D}$) is studied in section 2.2.

2.2 Crohn's Disease and Diseased Controls.

With the results of the agglutination reactions of 164 patients with CD and 234 non-Crohn's ($\bar{C}\bar{D}$, consisting of 46 patients with UC and 188 healthy subjects) a dichotomous logistic regression model was applied:

$$\log P(CD|x)/P(\bar{C}\bar{D}|x) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \qquad (2.3)$$

with $x_k = 0$ (negative result) or 1,2,3 (positive results according to strength of agglutination), k=1,2,3,4. After inspection of the log likelihood ratios for each agglutination test separately: $\log P(x_k|CD)/P(x_k|\bar{C}\bar{D})$, $k = 1,2,3,4$, in order to check if the relationships are monotonous in x_k, two tests (x_3 = Me46 and x_4 = Me47) were recoded into binary variables: $x_k = 0$ for negative outcomes and $x_k = 1$ for positive outcomes (k=3,4). The coding of x_1 = Me44 and x_2 = C18 was maintained. Table 4 summarizes the estimates of the parameters b_k and their 95% confidence limits. The constant term $\hat{b}_0$ in (2.3) is for equal priors (instead of proportional priors P(CD) = 164/398 = 0.41, $P(\bar{C}\bar{D})$ = 234/398 = 0.59) modified to $\hat{b}_0^* = \hat{b}_0 + \log(234/164) = -1.50$.

Table 4. Dichotomous logistic regression for the Dutch training cases: Crohn's Disease versus non-Crohn's Disease. Prior probabilities are proportional to sample sizes (P(CD)=41%)

Test	$\hat{b}$ (95% conf.limits)
constant	-1.86(-2.23, -1.50)
x_1(Me44)	0.53(0.23, 0.83)
x_2(C18)	0.55(0.24, 0.87)
x_3(Me46)	2.44(1.66, 3.22)
x_4(Me47)	1.58(0.83, 2.34)

For each case of the Dutch training dataset, the posterior probabilities of Crohn's disease were calculated with (2.3) and priors 0.5. The results for the subjects, for patients with CD and for patients without CD separately, are given in Table 5, column (a). The posterior probabilities are classified into 10 deciles of probability, based on fixed cutpoints (0-<0.1), 0.1-<0.2, 0.2-<0.3, etc.). For this classification Hosmer and Lemeshow (14) used a goodness-of-fit statistic Hg* for the probabilities derived from the logistic model (appendix 2). Calculation of this test-statistic gives: $\hat{Hg}^* = 7.08$ (p=0.53), so model (2.3) has an excellent fit. Indeed, higher-order (quadratic and interaction) terms of the test variables x_k appeared to be not significant.

Based on the posterior probabilities, several allocation rules may be developed (see also van de Merwe et al. (15)). A simple one uses a cut-off point of 0.5: a subject with agglutination reactions x is classified as CD if $P(CD|x) > 0.50$ and as $\overline{CD}$ otherwise. With this rule, from Table 5, column (a), follow some elementary performance characteristics: PV^+ (positive predictive value) = 118/138 = 86%, PV^- (negative predictive value) = 214/260 = 82%, ERR (error rate) = 66/398 = 17%. A high positive predictive value of our decision rule may be reached by choosing a higher cut-off point. If we decide to classify to CD only if $P(CD|x) \geq 0.90$ for example, the positive predictive value is increased to 63/66 = 95%. However, the negative predictive value is then decreased to 231/332 = 70%. A more elaborative interpretation of the posterior probabilities derived from the Dutch training cases is considered in section 5.

Table 5. Frequency distribution of posterior probabilities P(CD|x) and performance at prior probabilities proportional to sizes in patients with or without CD for the Dutch training cases.
(a) Logistic discrimination LOG (b) Fishers linear discriminant analysis LDA (c) Independence model IND.

	(a) LOG		(b) LDA		(c) IND	
P(CD|x)	Crohn's Disease	No Crohn's Disease	Crohn's Disease	No Crohn's Disease	Crohn's Disease	No Crohn's Disease
0-<0.10	0	0	27	156	27	156
0.10-<0.20	27	156	7	35	5	33
0.20-<0.30	5	33	4	13	6	8
0.30-<0.40	5	13	8	10	3	11
0.40-<0.50	9	12	6	8	0	0
0.50-<0.60	6	8	9	6	12	13
0.60-<0.70	14	6	9	0	1	4
0.70-<0.80	12	0	11	1	4	2
0.80-<0.90	23	3	24	1	14	1
0.90- 1	63	3	59	4	92	6
Total	164	234	164	234	164	234
Error rate	0.17		0.16		0.17	
MLS	0.09		0.09		0.09	
Hg*	7.08		24.85		61.05	
p-value	0.53		0.0017		<0.001	
REL	-0.0008		0.02		0.04	

3. ALTERNATIVE STATISTICAL DECISION-SUPPORT MODELS AND THEIR PERFORMANCE.

3.1 Alternative Statistical Models for Discrimination.

Before studying the validity of the logistic discrimination procedure (section 4), two other questions will be dealt with. First, which are the appropriate evaluation criteria for assessment of performance of our discrimination model (section 3.2)? Second, is the logistic discrimination procedure a good choice from the statistical decision-support models which are at our disposal (section 3.3)?

The logistic model (LOG) directly estimates the posterior probabilities as a function of the explanatory or predicting varibales. Most other discriminant analysis methods are based on Bayes' theorem which expresses the posterior probabilities as a function of the priors and the densities in each population. Differences between methods concern different ways of estimating these densities. A classical approach is Fisher's linear discriminant analysis (LDA) which assumes multinormal densities with equal covariance matrices. The simplest model for the densities is to assume statistical independence of all variables, the independence model (IND). There are other methods, but we will limit us to these three popular procedures: LOG, LDA and IND. Simplicity, availability and interpretability are important aspects in the selection of a method in practice (Begg, (16)). Both LOG, LDA and IND possess these characterizations.

3.2 Performance of Decision-Support Models.

For evaluation of the performance of decision-support models, a spectrum of measures exists. The measures we will consider may be divided into three groups. The first group consists of measures independent of a priori probabilities of the diagnostic categories: sensitivity, specificity and the ROC (receiver operating characteristic)-curve. The other two classes of evaluation measures contain quantities which depend on prior probabilities. In the first of them, discriminatory ability is measured. The second class of priors-dependent measures considers a different characteristic of a decision-support model: the goodness-of-fit of the diagnostic probabilities. We will address to these three aspects of performance.

Two measures which do not depend on the prior probabilities are the sensitivity and specificity of a diagnostic procedure. They are defined as the percentage of correct allocations within a certain diagnostic category. Usually, sensitivity is used within the disease category (here CD), specificity within the normal or control category (here $\bar{C}\bar{D}$). For different choices of the cut-off point in the allocation rule, sensitivity may be set out on the vertical axis versus 1 - specificity on the horizontal axis. By connecting these points a ROC-curve is obtained. We will apply this procedure for the discriminant analysis methods LOG, LDA and IND with the Dutch training cases. It must be remarked that the results (see section 3.3) are calculated from the tables with posterior probabilities, but the same curve is obtained from the likelihood ratios, so it is independent of the priors.

Statistical approaches to the analysis of ROC-curves were recently reviewed by McNeil and Hanley (17). Use of ROC-curves as clinical performance measure of differential diagnostic methods in the field of gastroenterology is illustrated among others by De Dombal and Horrocks (18) and De Dombal et al. (19).

Measures calculated from posterior probabilities assess several characteristics. By far the most important characteristic is how well the diagnostic categories are separated in a probabilistic sense by the discrimination method. If a method does this well, the method can be said to possess a good discriminatory ability. For an extensive discussion of these measures see Hilden et al. (20). We will use two measures for discriminatory ability: the error rate ERR and the modified logarithmic score MLS. The error rate is the proportion of elements not assigned to their actual diagnostic category when allocating according to maximal posterior probability. It may be viewed as a weigthed average of the individual misclassification rates within each diagnostic category separately. For the weigths the prior probabilities are chosen. A measure which takes account of posterior probabilities in a continuous way, and which therefore may be more sensitive for detecting differences between methods than the error rate, is the modified logarithmic score MLS (appendix 2). The MLS varies between 0 and 1, while low values indicate good performance.

Quite another aspect of performance in probabilistic diagnosis is the reliability or goodness-of-fit of the diagnostic probabilities. Whenever a disease is assigned a probability P, this disease must occur with frequency P. The most thoroughly way of checking goodness-of-fit of the posterior probabilities would be to obtain a large number of elements for each possible x-value and see what

proportions arise from each diagnostic category. These proportions should be close to the predicted probabilities afforded by the discriminant analysis method. This procedure is out of question because of the very large number of cases involved. One alternative approach we apply is the goodness-of-fit statistic $\hat{H}g^*$, used by Hosmer and Lemeshow (14), calculated from the posterior probabilities classified into deciles of probability (see section 2.2). A second goodness-of-fit or reliability measure compares the discrimination which is actually obtained by the method, with the expected discrimination if the probabilities given by the discriminant analysis method were perfectly reliable. Such a measure equals the difference between a discrimination measure and its expectation under the null hypothesis of perfect reliability (see Hilden et al. (21)). The reliability measure REL based on the modified logarithmic score MLS is defined in appendix 2. Near zero-values of REL indicate good reliability, large values indicate unreliability. A negative value of REL indicates overconfident behaviour of the method (the method expects itself to discriminate better than it really does), a positive value indicates diffident behaviour (the method underestimates its own discriminatory ability).

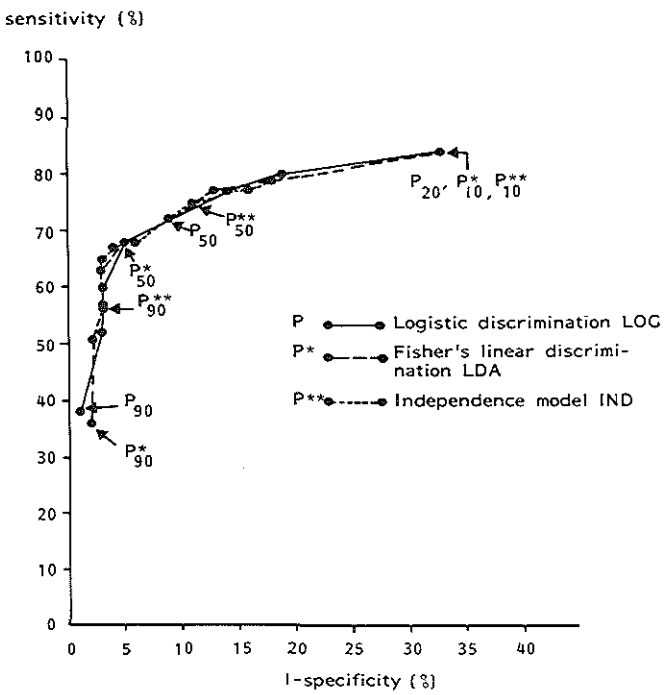


Figure 3. ROC-curves of logistic discrimination, linear discrimination and an independence model, applied to the Dutch training cases.

3.3 Application to the Dutch Training Cases.

The three statistical models for discrimination described in section 3.1, LOG, LDA and IND, were applied to the Dutch training cases: four agglutination tests x_k (k=1,...,4), with x_k= 0,1,2,3 if k=1,2 and x_k=0,1 if k=3,4, are available for 164 patients with Crohn's disease and 234 patients without Crohn's disease. A frequency distribution of the posterior probabilities P(CD|x) obtained with the logistic model has been described already in section 2.2 and is summarized in Table 5, together with the distributions obtained with LDA and IND. Taking the 10-deciles 0.1, 0.2, etc. of the posterior probabilities as cut-off points for allocation, for each of the methods LOG, LDA and IND a ROC-curve was constructed, see Figure 3. It appears that the three curves show a striking similar shape, although their ranges are not equal. This is because a fixed cut-off point of the posterior probabilities for all methods does not correspond with one point in the curve. An example is presented in Table 6 for the cut-off points P(CD|x)=0.50 and P(CD|x)=0.90. Especially for the more extreme point 0.90 remarkable differences in sensitivity and thus in the error rate exist between LOG and LDA at one side and IND at the other side. This reflects the pitfall of comparing methods with respect to discriminatory ability by only calculating error rates from posterior probabilities. IND has a lower error rate (0.20) than LOG, and LDA (0.26 and 0.27 respectively), but Figure 3 suggests that the discriminatory abilities are much more similar.

Table 6. Sensitivity, specificity, error rate and predictive values (in percentages) for LOG, LDA and IND using two different cut-off points for allocation. Results are based on the Dutch training cases with priors proportional to sample sizes.

	Allocate to CD if					
	P(CD|x)≥ 0.50			P(CD|x)≥ 0.90		
	LOG	LDA	IND	LOG	LDA	IND
Sensitivity	72	68	75	38	36	56
Specificity	91	95	89	99	98	97
Pos.predictive value	86	90	83	95	94	94
Neg.predictive value	82	81	84	70	69	76
Error rate	17	16	17	26	27	20

The modified logarithmic score MLS for LOG, LDA and IND is almost equal for the Dutch training cases (see Table 5). This confirms our impression of completely comparable discriminatory ability of the three methods. With respect of goodness-of-fit, however, striking differences are observed, both in measure $\hat{H}g^*$ and in measure REL (Table 5). The less reliable method is IND, but also LDA has a rather bad fit: both show a diffident behaviour. Especially for the independence method, this looks in contradiction with what would be expected. A problem with independence models is that the variables are frequently dependent. As a consequence, by multiplication of likelihoodratios from correlated variables, the posterior probabilities will be too close to 0 or 1, so an overconfident performance would result. An explanation of the diffident performance of IND in our data (REL positive) is the very skew distribution of the posterior probabilities caused by the frequent occurence of the outcome 'all agglutination tests zero'.

From these comparisons it is concluded that the logistic discrimination model is a good choice from the simple and good interpretable methods LOG, LDA and IND. Especially with respect to reliability the logistic model will be preferred. Whether this model derived from the Dutch training cases will perform satisfactory with independent new data is studied in the next section.

4. VALIDATION OF THE LOGISTIC MODEL.

4.1 Approaches to Validation of Diagnostic Methods.

The performance of the logistic model (2.3) is evaluated in section 3 by simple resubstitution of the training cases. Such an assessment is usually too optimistic (Lachenbruch and Mickey (22) and Schmitz et al. (23)). One should like to test the discrimination model on independent sources of data. This testing is generally referred to as cross-validation, or shortly validation.

With cross-validation by the jackknife technique, several subsets are created from the original data. The discriminant analysis method is developed on one subset and tested on another. With the leaving-one-out technique (Lachenbruch and Mickey (22)) the allocation rule is constructed from all training sample elements except the one that has to be allocated. Repeated application for each training sample element results in 'leaving-one-out' estimates of performance. For the logistic method this approach will generally

Table 7. Resubstitution and cross-validation evaluation of the logistic model for the Dutch training cases and validation results for the Dutch and the international cases.

	SENS	SPEC	PV+	PV-	ERR	MLS	Hg*(p-value)	REL
Dutch training cases (P(CD)= .41)								
- Resubstitution	.72	.91	.86	.82	.17	.09	7.08(.53)	-0.008
- Cross-validation 1	.76	.90	.84	.84	.16	.09	19.84(.01)	+.01
- Cross-validation 2	.66	.92	.86	.79	.19	.10	10.38(.24)	-0.005
-Average cross-valid.	.71	.91	.85	.82	.18	.10		+0.005
Dutch validation cases (P(CD)=.19)	.33	.98	.80	.86	.15	.09	19.17(0.1)	+0.02
International validation cases (P(CD)=.41)	.55	.80	.66	.72	.30	.15	69.46(<.001)	+0.03

take too much computer time. A less time-consuming approach is the random splitting of the training cases into two halves: one for construction of the allocation rule and one for evaluation (and vice-versa). The split-sample results obtained with the Dutch training cases are shown in Table 7. It appears that the discriminatory ability measures (ERR and MLS) remain approximately constant in this cross-validation, while the fit of the model ($\hat{Hg}^*$ and REL) clearly deteriorates.

Another and preferred approach is cross-validation with new data from the same or from independent populations. In section 4.2 an validation of the logistic model is studied with new Dutch cases obtained during the period 11.1.1983 to 7.31.1984. In section 4.3 agglutination test results from an international dataset were used. With these data the possibilities of transferability of our clinical decision-aid will be studied.

4.2 Validation with the Dutch Validation Cases.

The logistic model (2.3) was calculated from agglutination test results obtained from patients during the period 1975 to 1980. The 164 patients with Crohn's disease and 46 patients with ulcerative colitis visited the Departments of Internal Medicine and Surgery of the University Hospitals in Rotterdam and Leiden, while the 188 healthy subjects were volunteers of the Red Cross Blood Transfusion Service, Rotterdam. Because UC could not be distinguished as a separate group, we amalgamated UC and HS into one group: non-Crohn's.

In order to assess the diagnostic value of the discriminant function, constructed with these data, all agglutination test results obtained during the period 11.1.1983 - 7.31.1984 in the laboratory at the Department of Medical Microbiology at Rotterdam were collected. These tests were done with sera of patients in the Netherlands, and were ordered by Dutch specialists from several disciplines. In September, 1984 these specialists were requested to allocate the patients whose sera were tested into one of the following five diagnostic categories:

certain diagnosis:
(1) Crohn's disease (CD)
(2) ulcerative colitis (UC)
(3) other than CD or UC: diseased controls (DC)
uncertain diagnoses:
(4) CD or UC
(5) CD or UC or other

For this allocation clinical, radiological, endoscopical and histological information was used. Thus, 437 patients were divided into 61 CD, 29 UC, 226 DC, 28 category (4) and 93 category (5). The patients allocated to the uncertain diagnoses were excluded from further evaluation analyses. The patients with UC and with DC were amalgamated into a new category: non-Crohn's ($\overline{CD}$).

The logistic formula (2.3) was evaluated with the 61 patients with CD and the 29 + 226 = 255 patients without CD. The results, using prior probability P(CD) = 61/(61+255) = 19%, are shown in Table 7. The discriminatory ability, ERR=.15 and MLS=.09, is quite comparable with the training data results (ERR=.17, MLS=.09). However, the comparability is difficult because of the differences in the prior probabilities of CD in the training data (P(CD)=.41)

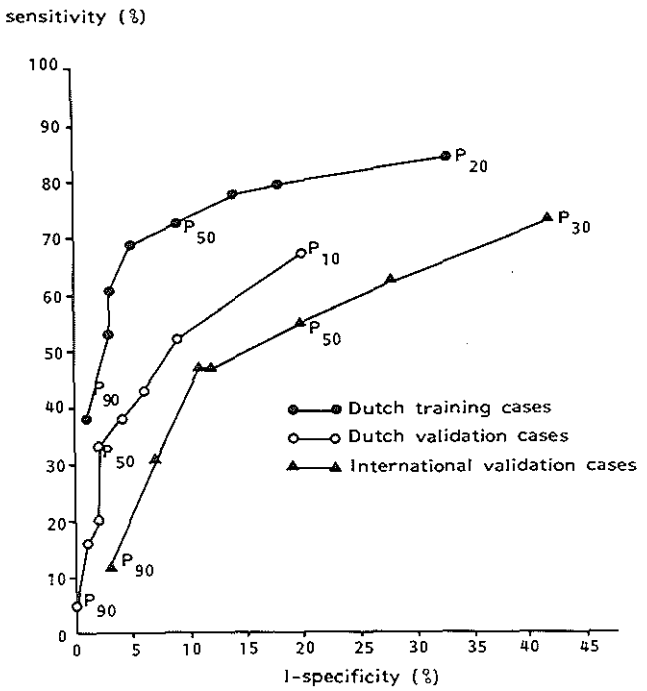


Figure 4. ROC-curves of logistic discrimination derived from the Dutch training cases and applied to three datasets: the Dutch training cases, the Dutch validation cases and the international validation cases.

and the validation data P(CD)=.19). More insight is gained from an ROC-curve: see Figure 4. It clearly appears that the discriminatory ability is worse for the Dutch validation cases than for the training cases. The rather small effect on the error rate depends largely upon the relatively small prior probability of CD in the validation cases. Actually, the ERR-value of .15 has to be compared with the ERR-value obtained by minimising the expression

$$ERR = P(CD).(1\text{-sensitivity}) + P(\overline{CD}).(1\text{-specificity})$$

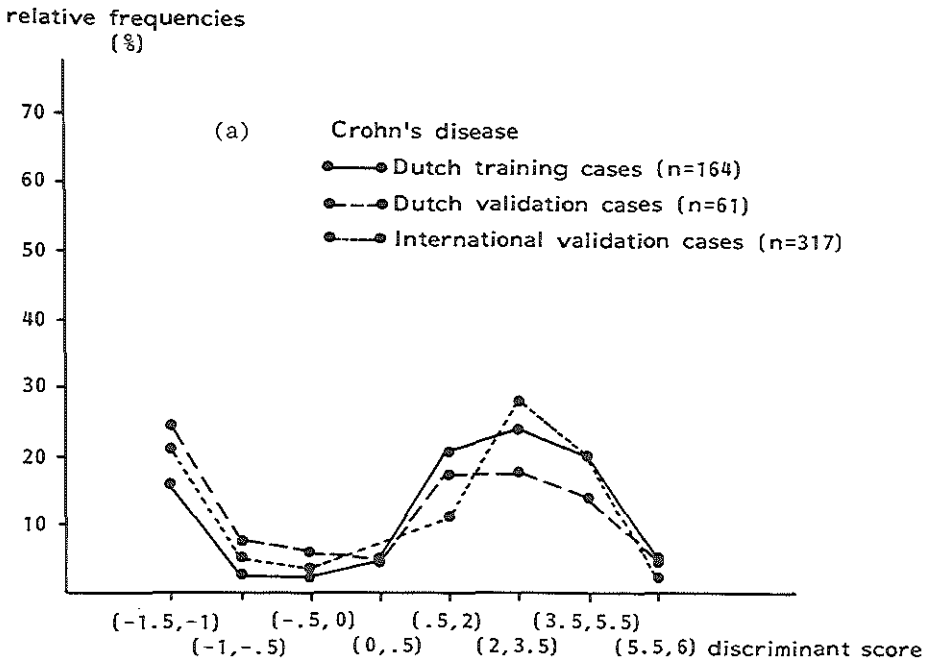
over the possible sensitivity-specificity combinations from the ROC-curve for the training data and substituting P(CD)=.19. This results in: sensitivity=.68, 1-specificity=.05, ERR=.10. So, a fall of 5 percent misclassification is the consequence of applying our logistic discrimination rule to the independent new Dutch cases.

4.3 Validation with International Cases.

The world-wide occurrence of agglutination antibodies to the four coccoid anaerobes Me44, Me47, Me46 and C18 was investigated in sera from 773 patients suffering from Crohn's disease, ulcerative colitis, various other diseases and from healthy controls. A list of the 19 participating centres from all over the world and preliminary results were described by Wensinck et al. (24). Participants were requested to send coded serum samples of about 20 patients with CD, 20 with UC and of 20 control subjects to Rotterdam for agglutination reactions, and to send the decodings to Cardiff. The criteria for diagnosis were those usually applied at each of the centres. No specifications were given for composition of groups with regard to site of disease activity, surgical history, medical treatment nor the way subjects were selected from the hospital patients population.

Test results were obtained from 317 patients with CD, 232 patients with UC, 157 healthy subjects HS, and 67 patients with various diseases other than CD or UC, say diseased controls DC. For validation of the logistic model derived from the Dutch training cases we amalgamated the 232+157+67 = 456 non-Crohns into an overall control category $\overline{CD}$. The validation results, using prior P(CD) = 317/773 = 0.41 are shown in Table 7. Both discriminatory ability (ERR, MLS and the ROC-curve in Figure 4) and goodness-of-fit are remarkably poor.

Several possible explanations for this poor performance on the world-wide data will be considered here briefly. First, a variation in the prior probability of CD could be a possible cause. However, the prior P(CD) in the international dataset appeared to be equal to that in the Rotterdam training cases (although this was by chance) Second, the conditional probability distributions, that is the distribution of the logistic discriminant score z, may be different between datasets. Indeed, the relative frequency distributions of z for the three samples with patients with Crohn's disease, are not completely similar (Fig.5.a), and especially between the non-Crohn groups (Figure 5.b) high (low) values of the discriminant score are more (less) frequently in the international sera than in those from the Netherlands (training and validation cases). The different composition of the control subjects could be an explanation. However, also between more or less homogeneous control groups (Figure 5.c and 5.d) the different distribution of the discriminant score for the international cases remains. The effect for UC-cases is somewhat stronger, so this may reinforce



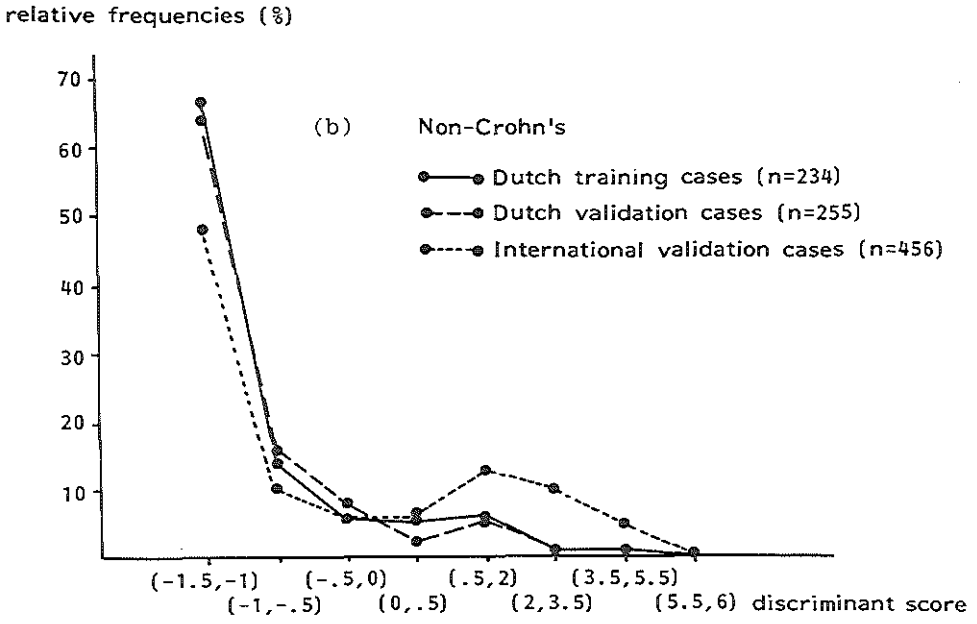
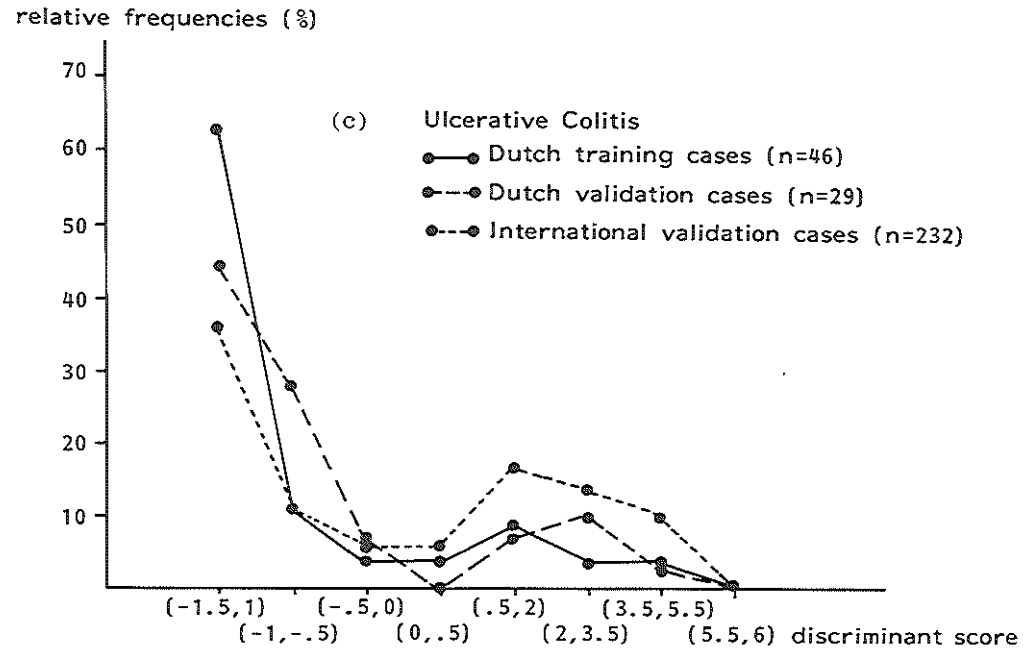


Figure 5. Distribution of the logistic discriminant score for the Dutch training and validation cases and the international validation cases.
(a) Patients with Crohn's disease (b) The total non-Crohn group
(c) Patients with ulcerative colitis (d) Control groups other than UC



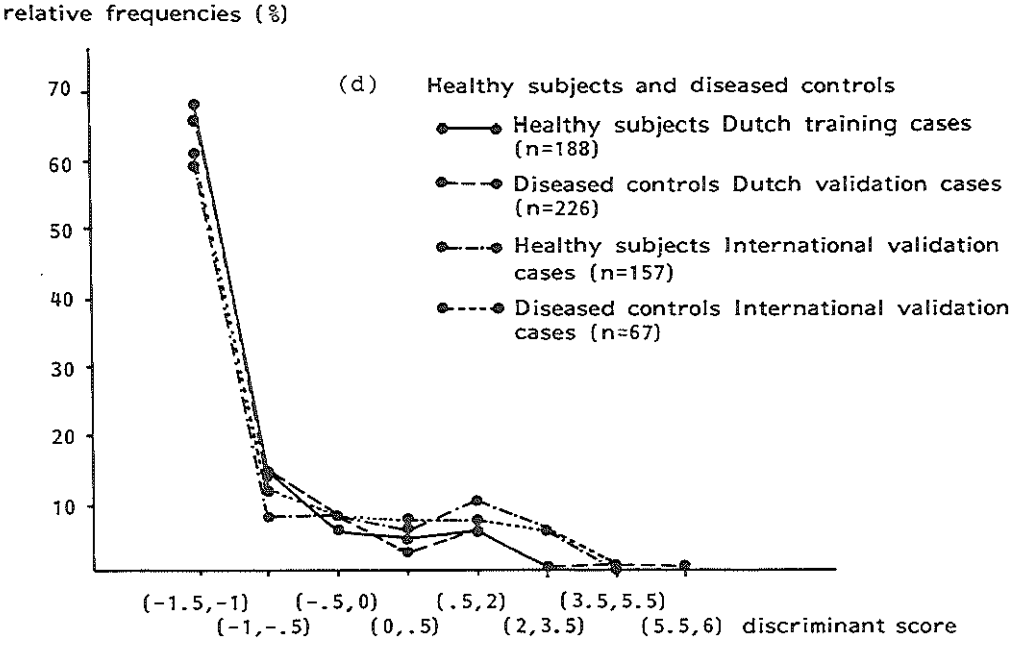


Figure 5. (Continued)

the poor result because UC-prevalence among non-Crohn's is much higher (232/456= 51%) in the international than in the Dutch cases (20% and 11%).

5. IMPLEMENTATION IN CLINICAL PRACTICE

The calculation of the posterior probability of Crohn's disease may be done with a simple calculator or microcomputer, using the expression derived in section 2.2 (see Table 4):

$$P(CD\,|x_1,x_2,x_3,x_4) = 1/(1+\exp(-(-1.50+0.53x_1+0.55x_2+2.44x_3+1.58x_4))) \qquad (5.1)$$

for equal priors. For prior probability P(CD) other than 0.5, the term -1.50 in (5.1) is modified into $-1.50 + \log(P(CD)/(1-P(CD)))$. Because the possible test outcomes possess only a limited number of values (4x4x2x2=64), a tabulation of these posterior probabilities for five different priors (0.01, 0.10, 0.25, 0.40 and 0.50) may be helpful (Table 8).

After obtaining the estimated posterior probabilities of CD, several allocation rules can be used. For example, a subject with posterior probability $p = P(CD\,|x_1,x_2,x_3,x_4)$ is classified as

non-CD if $0 \leq p < 0.5$
CD if $0.5 \leq p < 1$.

Other thresholds than 0.5, the use of a 'region of doubt' and dividing of the posterior probabilities into more than three classes, each with its own interpretation, has been considered by van de Merwe et al. (15). An ad hoc assessment of the value of our 'combined' agglutination test is to use three thresholds for the posterior probability, actually based on likelihood ratios. For prior P(CD) = 0.5 and the following interpretation:

no support for CD if $0 \leq p < 0.80$
suspected CD if $0.8 \leq p < 0.95$
probable CD if $0.95 \leq p \leq 1$,

the frequency distribution of posterior probabilities is shown in Table 9. It is left to the physician which orders the test how to use the result in further patient management. Assessment of the thresholds is based on rather subjective grounds.

Table 8. Posterior probabilities of Crohn's Disease (in %) for five different prior probabilities.

test	prior probability					test	prior probability				
outcome	0.01	0.10	0.25	0.40	0.50	outcome	0.01	0.10	0.25	0.40	0.50
0000	00	02	07	13	18	2000	01	07	18	30	39
0001	01	11	27	42	52	2001	03	26	51	68	76
0010	03	22	46	63	72	2010	07	45	71	83	88
0011	11	58	81	89	92	2011	27	80	92	96	97
0100	00	04	11	21	28	2100	01	11	27	43	53
0101	02	17	38	56	65	2101	05	38	64	78	84
0110	04	33	60	75	82	2110	12	59	81	90	93
0111	18	70	88	94	96	2111	39	87	95	98	98
0200	01	07	18	31	40	2200	02	18	39	56	66
0201	03	27	52	68	77	2201	09	51	76	86	90
0210	07	46	72	84	88	2210	18	71	88	94	96
0211	27	81	93	96	97	2211	52	92	97	99	99
0300	01	11	28	44	54	2300	03	27	53	69	77
0301	05	38	65	79	85	2301	14	64	84	92	94
0310	12	60	82	90	93	2310	28	81	93	96	98
0311	40	88	96	98	98	2311	65	95	98	99	99
1000	00	04	11	20	28	3000	01	11	27	42	52
1001	02	17	38	55	65	3001	05	37	64	78	84
1010	04	33	59	74	81	3010	11	58	81	89	93
1011	18	70	88	93	96	3011	38	87	95	98	98
1100	01	07	18	30	40	3100	09	17	39	56	66
1101	03	26	52	68	76	3101	09	51	75	86	90
1110	07	46	72	83	88	3110	18	71	88	94	96
1111	27	80	92	96	97	3111	52	92	97	99	99
1200	01	11	28	43	53	3200	03	27	52	69	77
1201	05	38	65	79	85	3201	14	64	84	91	94
1210	12	59	81	90	93	3210	28	81	93	96	97
1211	39	88	95	98	98	3211	65	95	98	99	99
1300	02	18	40	57	66	3300	05	39	65	79	85
1301	09	52	76	86	91	3301	22	76	90	95	97
1310	19	72	88	94	96	3310	40	88	96	98	99
1311	53	92	97	99	99	3311	76	97	99	99	100

The linear function of agglutination reactions in (5.1) may be viewed as an approximately continuous diagnostic test. In clinical practice, the usefulness of a test depends on its potential to alter patient management with regard to further diagnostic and therapeutic actions. Ideally, its clinical value should be assessed by calculating expected utilities for each strategy of patient management. Before calculating these utilities, clinical decisions must be structured, usually with help of decision trees. The following ad hoc procedure for assessment of the clinical value of our 'test' uses a single threshold for the posterior probability. This threshold may be based on an implicit loss-structure, without explicit defining of utilities (or losses). Suppose, for example, that the test is used for differentiating between CD and non-CD. After performing the agglutination reactions, the implication of the results could be strategy 1 or strategy 2. Each strategy consists of a certain combination of diagnostic and therapeutic actions, not further specified here. But strategy 1 is the most appropriate for patients with CD and strategy 2 is the most appropriate for non-Crohn cases. The relative values of the consequences of these actions, depend on the presence or absence of CD, and may be quantified as losses l_{12} and l_{21} (a 'loss' is a negative 'utility'). A simple decision tree for this approach is shown in Figure 6. The branch STRATEGY 1 is chosen if this minimises the expected loss:

Table 9. Frequency distribution of posterior probabilities P(CD|x) in three classes of posterior probabilities in patients with CD and in patients without CD. Prior P(CD) = 0.50.

P(CD|x)	Interpretation	Number (%) of CD	Number (%) of $\bar{CD}$
0-<0.80	No support for CD	66(40)	228(98)
0.80-<0.95	Suspected CD	43(26)	3(1)
0.95-1	Probable CD	55(34)	3(1)
Total		164(100)	234(100)

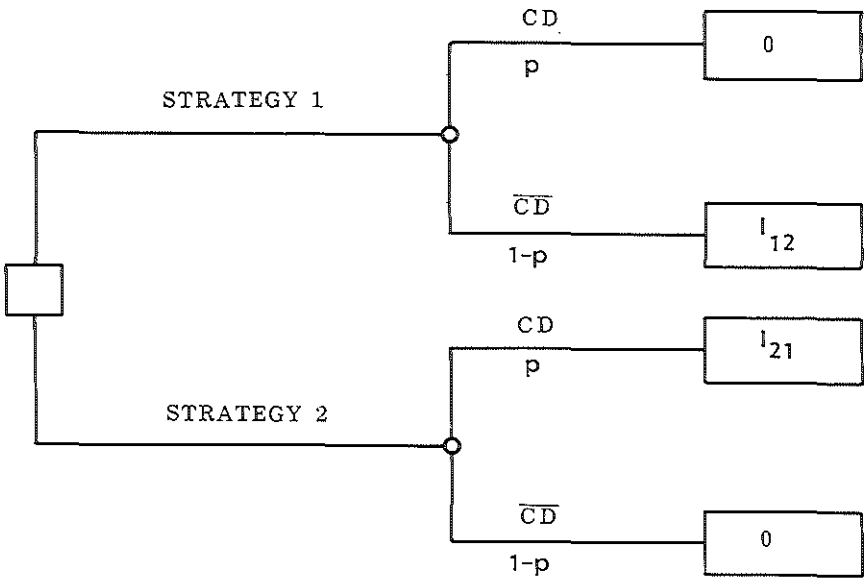


Figure 6. Decision tree for patients for which agglutination tests are performed.

$$(1-p)l_{12} < p\, l_{21}\ ,\ \text{thus:}\ \ p/(1-p) > l_{12}/l_{21} \qquad (5.2)$$

$$(p \overset{\text{def}}{=} P(CD \mid x_1,x_2,x_3,x_4))$$

For the loss-structure $l_{12} = l_{21} = 1$, (5.2) becomes: $p>0.5$.
A more realistic loss structure could be: $l_{12} = 10.l_{21}$ (it is ten times as harmful to follow strategy 1 for patients without CD, than to follow strategy 2 for patients with CD). The branch STRATEGY 1 (Fig. 6) is chosen now if the posterior probability is greater than $10/11 = 0.91$.

In order to illustrate the dependency of the posterior probability of CD on the prior probability of CD, a so-called post test - pre test graph may be helpful, see Figure 7. For each of the 64 possible test outcomes a figure may be drawn. In Fig.7 the 'lowest' graph (a) (for $x_1=x_2=x_3=x_4=0$) and the 'highest' graph (b) (for $x_1=x_2=3$, $x_3=x_4=1$) have been plotted. For the loss-structure considered, so with threshold $p=P(CD \mid x_1,x_2,x_3,x_4)=0.91$, it can be seen that only for very low

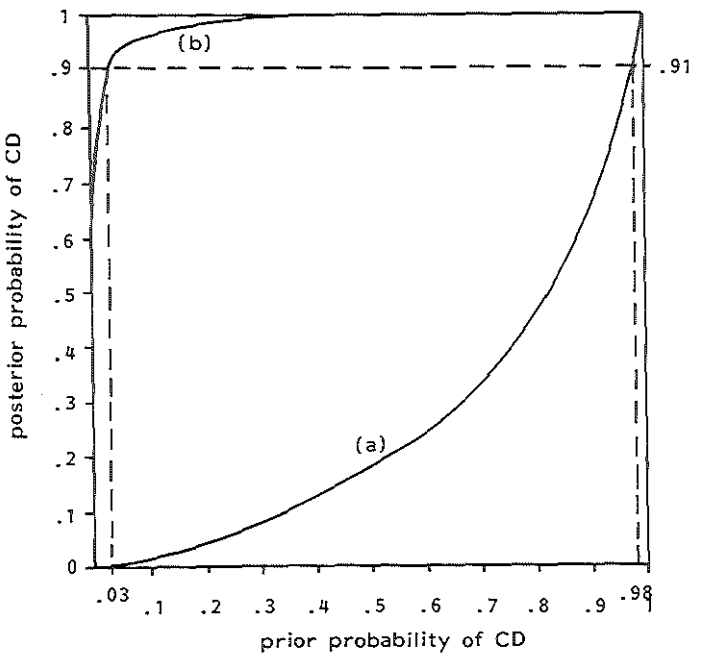


Figure 7. Post test - pre test graph for two possible combinations of outcomes of the agglutination tests.
(a) $x_1=x_2=x_3=x_4=0$, (b) $x_1=x_2=3$, $x_3=x_4=1$.

priors (P(CD)<0.03) ordering agglutination tests will never result in choosing the STRATEGY 1 branch, and only for very high priors (P(CD)>0.98) each test outcome should result in choosing the STRATEGY 1 branch. So,there remains a wide range of priors for which test ordering may alter further patient management.

6. DISCUSSION

Comparison of discriminant analysis methods has been studied extensively before using evaluation criteria which measure discriminatory ability (Schmitz et al.(23),(25),(26)), and criteria which measure reliability

(Titterington et al. (27)). From the three discriminant analysis methods considered in the present study, the logistic method had a better reliability. Logistic regression is nowadays straightforward to carry out, utilizing one of the available computer packages, but problems arise with marginal null frequencies (Anderson, (28)) or considerable missing data (Spiegelhalter, (29)). An independence model may be a good alternative in those situations. Also, it may be remarked that variable selection is usually more important than the selection of a model (Begg, (16)). None of the possible problems mentioned (null frequencies, missing data, selection of variables) played a part in the construction of our discrimination rule.

The relative strong deterioration of the performance of the logistic model in both the Dutch validation cases and the international cases may have several causes. Selection biases, such as referral bias (Begg and Greenes, (30)), will certainly be not absent because the way of gathering the cases is rather different in the datasets considered. The most plausible reason of the poor validity, however, is the lack of precise definitions for the several diagnostic categories in especially the validation data. It should have been preferred if for the definitions of Crohn's disease and ulcerative colitis criteria such as from Lennard-Jones et al. (7) had been strictly applied, or if the similar approach of Clamp et al. (13), that is using a 'working' diagnosis based on a simple scoring system, had been utilized. Then, it could be expected that a more valid result was obtained.

In clinical practice in the Netherlands there are several situations in which agglutination tests are ordered. For example, for patients from an outpatient department with abdominal discomfort or diarrhoea the agglutination reactions are used as a screening test, so the prior probability of Crohn's disease will be low. In other situations, such as when patients have been admitted to hospital and when other diagnostic investigations do not give a decisive answer, the prior probability may be approximately 50%. As long as the prior is in the test informative range, ordering agglutination tests may be meaningful. As no other simple diagnostic test for Crohn's disease is available, the 'composed' test, summarizing the results of the four agglutination tests in a logistic score, has a place in the clinical diagnostic methods for Crohn's disease.

APPENDIX 1. Polychotomous logistic regression analysis.

It is easily seen from (2.1) that for a given agglutination test result $x = (x_1, x_2, x_3, x_4)$ the three diagnostic probabilities which should add to one, are given by:

$$P(HS|x) = P_0 = 1/(1+\exp(z_1)+\exp(z_2))$$
$$P(UC|x) = P_1 = \exp(z_1)/(1+\exp(z_1)+\exp(z_2))$$
$$P(CD|x) = P_2 = \exp(z_2)/(1+\exp(z_1)+\exp(z_2)) \qquad (a.1)$$

As with the dichotomous logistic model in section 1, estimates of the unknown parameters β_k and γ_k (k=0,1,...,4) can be obtained using the method of maximum likelihood. Wijesinha et al. (10) show that the estimates in the polychotomous regression model also follow from the alternative approach of individualized regression. This is a simplified method in which each diagnostic category is individually compared with a 'normal' baseline category using the dichotomous logistic model. Begg and Gray (11) have shown that the asymptotic relative efficiences of the individual parameter estimates are generally high. The conditional probabilities in the individualized logistic regressions are:

$$\theta_1 = P(UC|x, \text{ UC or HS}), \quad \theta_2 = P(CD|x, \text{ CD or HS}) \qquad (a.2)$$

The individualized logistic regressions are modelled by:

$$\log \theta_1/(1-\theta_1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = z_1$$

$$\log \theta_2/(1-\theta_2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 = z_2 \qquad (a.3)$$

With Bayes' s theorem follows:

$$\theta_1(1-\theta_1) = P_1/P_0 \quad \text{and} \quad \theta_2/(1-\theta_2) = P_2/P_0.$$

Consequently, the resulting parameter estimates $\hat{\beta}_k$, $\hat{\gamma}_k$ from the separate simple logistic analyses (a.3) may be substituted in (2.1).
Suppose generally, π_0, π_1 and π_2 are the 'old' prior probabilities of HS, UC and CD respectively, and π_0^*, π_1^* and π_2^* are the 'new' priors, calculation of the posterior probabilities for new priors is performed as before, using (2.1) and (a.1), only with replacement of the constant terms $\hat{\beta}_0$ and $\hat{\gamma}_0$ by:

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log((\pi_1^* \pi_0)/(\pi_0^* \pi_1^*))$$
$$\hat{\gamma}_0^* = \hat{\gamma}_0 + \log((\pi_2^* \pi_0)/(\pi_0^* \pi_1^*)) \qquad (a.4)$$

(see Anderson (9)).

APPENDIX 2. Performance measures.

The goodness-of-fit statistic for the logistic model (2.3) based on the approach of Hosmer and Lemeshow (14) is:

$$\hat{Hg}^* = \sum_{k=0}^{1} \sum_{m=0}^{10} (o_{km} - e_{km})^2 / e_{km} \qquad (a.5)$$

with $o_{lm} = \sum_{i \in D_m} y_i$, $o_{om} = \sum_{i \in D_m} (1-y_i)$, $e_{lm} = \sum_{i \in D_m} P(x_i)$, $e_{om} = \sum_{i \in D_m} (1-P(x_i))$,

y_i is the disease indicator variable: $y_i=1$ if patient has CD, $y_i=0$ if patient has not CD, $P(x_i) = P(CD \mid x_{1i}, x_{2i}, x_{3i}, x_{4i})$, $D_m (m=1,2,\ldots,10)$ denotes the set of cases in the mth decile of probability: D_1 = 0-<0.1, D_2 = 0.1-<0.2, etc.
$\hat{Hg}^*$ has a chi-square distribution with 8 degrees of freedom.
The modified logarithmic score MLS (Hilden et al. (20)) is to a close approximation equal to:

$$MLS = \left(\frac{1}{n} \sum_i - \ln (P(i) + 0.02)/3.93 + 0.005, \right. \qquad (a.6)$$

with P(i) the probability assigned to the disease category of origin, and summation over all cases in the training or validation data.
The reliability measure REL (Hilden et al. (21) based on the modified logarithmic score MLS is:

$$REL = MLS - E(MLS) = \left(\frac{1}{n} \sum_i - \ln(P(i)+0.02) - \frac{1}{n} \sum_i (P(i)(-\ln(P(i)+0.02)) \right.$$

$$\left. + (1-P(i)) (-\ln(1-P(i)+0.02)))\right)/3.93 \qquad (a.7)$$

REFERENCES

(1) Kirsner,J.B., Shorter,R.G.: Recent developments in 'nonspecific' inflammatory bowel disease. N.Engl.J.Med.306(1982)775-785,837-848.

(2) Johnson,W.D., Roth,J.LA.: Diagnosis and differential of chronic ulcerative colitis and Crohn's colitis. In J.B.Kirsner and R.G.Shorter (Eds.): Inflammatory Bowel Disease, pp.201-224 (Philadelphia: Lea and Febiger 1975).

(3) Dyer,N.H. and Dawson,A.M.: Diagnosis of Crohn's Disease. A continuing Source of Error. Brit.Med.J.1(1970)735-737.

(4) Brandes,J.W., Eulenburg,F.: Der lange Weg zur Diagnose Morbus Crohn. Zeitschrift für Gastroenterologie 14(1976)400.

(5) Mekhjian,H.S., Switz,D.M., Melnyk,C.S., Rankin,G.B., Brooks,R.K.: Clinical features and natural history of Crohn's disease. Gastroenterology 77(1979)898-906.

(6) Wensinck, F., van de Merwe,J.P.: Serum agglutinins to Eubacterium and Peptostreptococcus species in Crohn's and other diseases. J.Hyg. 87 (1981)13-24.

(7) Lennard-Jones,J.E., Lockhart-Mummery,H.E., Morson,B.C.: Clinical and pathological differentiation of Crohn's disease and proctocolitis. Gastroenterology 54(1968) 1162.

(8) Kirsner,J.B.: Problems in the differentiation of ulcerative colitis and Crohn's disease of the colon: the need for repeated diagnostic evaluation. Gastroenterology 68(1975)187.

(9) Anderson,J.A.: Separate sample logistic discrimination. Biometrika 59 (1972)19-35.

(10) Wijesinha,A., Begg,C.B., Funkenstein,H.H., McNeil,B.J.: Methodology for the Differential Diagnosis of a Complex Data Set. A Case Study Using Data from Routine CT Scan Examinations. Med.Decis.Making 3(1983)133-154.

(11) Begg,C.B., Gray,R.: Calculation of polychotomous logistic regression parameters using individualized regressions. Biometrika 71(1984)11-18.

(12) Habbema,J.D.F., Hilden,J., Bjerregaard,B.: The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. Meth.Inf.Med.17(1978) 217-226.

(13) Clamp,S.E., Myren,J., Bouchier,I.A.D., Watkinson,G., de Dombal,F.T.: Diagnosis of inflammatory bowel disease: an international multicentre scoring system. Br.Med.J.284(1982)91-95.

(14) Hosmer,D.W., Lemeshow,S.: Goodness of fit tests for the multiple logistic regression model. Commun.Stat.A9(1980)1043-1069.

(15) van de Merwe,J.P., Schmitz,P.I.M., Wensinck,F.: Antibodies to Eubacterium and Peptostreptococcus species and the estimated probability of Crohn's disease. J.Hyg.87(1981)25-33.

(16) Begg,C.B.: Statistical methods in medical diagnosis (1985).In preparation.

(17) McNeil,B.J., Hanley,J.A.: Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med.Dec.Making 4(1984) 137-150.

(18) De Dombal,F.T., Horrocks,J.C.: Use of receiver operating characteristic (ROC) curves to evaluate computer confidence threshold and clinical performance in the diagnosis of appendicitis. Meth.Inf.Med.17(1978)157-161.

(19) De Dombal,F.T., Staniland,J.R., Clamp,S.E.: Geographical variation in disease presentation. Does it constitute a problem and can information science help? Med.Decis.Making 1(1981)59-69.

(20) Hilden,J., Habbema, J.D.F., Bjerregaard,B.: The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Meth.Inform.Med.17(1978)238-246.

(21) Hilden,J., Habbema,J.D.F., Bjerregaard,B.: The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Meth.Inform.Med. 17(1978)227-237.

(22) Lachenbruch,P.A., Mickey,M.R.: Estimation of error rates in discriminant analysis. Technometrics 10(1968)1-10.

(23) Schmitz,P.I.M., Habbema,J.D.F., Hermans,J.: The performance of logistic discrimination on myocaridal infarction data, in comparison with some other discriminant analysis methods. Stat.in Med. 2(1983)199-205.

(24) Wensinck,F., van de Merwe,J.P., Mayberry,J.F.: An international study of agglutination to Eubacterium, Peptostreptococcus and Coprococcus species in Crohn's disease, ulcerative colitis and control subjects. Digestion 27(1983)63-69.

(25) Schmitz,P.I.M., Habbema,J.D.F., Hermans,J., Raatgever,J.W.: Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables. Commun. Statist.-Simula.Computa 12 (1983)727-751.

(26) Schmitz,P.I.M., Habbema,J.D.F., Hermans,J.: A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables. J.Stat.Comput.Simula. 16(1985). In press.

(27) Titterington,D.M., Murray,G.D., Murray,L.S., Spiegelhalter,D.J., Skene,A.M., Habbema,J.D.F., Gelpke,G.J.: Comparison of discrimination techniques applied to a complex data set of head injured patients. J.Roy.Stat.Soc.Series A 144(1981)145-175.

(28) Anderson,J.A.: Diagnosis by logistic discriminant function: further practical problems and results. Appl.Stat.23(1974)397-404.

(29) Spiegelhalter,D.J.: Statistical aids in clinical decision-making. The Statistician 31(1982)19-36.

(30) Begg,C.B., Greenes,R.A.: Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 39(1983)207-215.

CHAPTER 3

ANTIBODIES TO EUBACTERIUM AND PEPTOSTREPTOCOCCUS SPECIES AND THE ESTIMATED PROBABILITY OF CROHN'S DISEASE

Antibodies to Eubacterium and Peptostreptococcus Species and the Estimated Probability of Crohn's Disease. J. P. van de Merwe, P. I. M. Schmitz and F. Wensinck, J. Hyg., Camb. 87 (1981) 25 - 33.

J. Hyg., Camb. (1981), **87**, 25
Printed in Great Britain

Antibodies to *Eubacterium* and *Peptostreptococcus* species and the estimated probability of Crohn's disease

By J. P. VAN DE MERWE
Department of Medical Microbiology, Erasmus University, Rotterdam, The Netherlands

P. I. M. SCHMITZ
Institute of Biostatistics, Erasmus University, Rotterdam, The Netherlands

AND F. WENSINCK
Department of Medical Microbiology, Erasmus University, Rotterdam, The Netherlands

(*Received* 2 *January* 1980)

SUMMARY

Anaerobic coccoid rods belonging to species of *Eubacterium* and *Peptostreptococcus* agglutinate more frequently with sera from patients with Crohn's disease than with sera from patients suffering from other diseases and from healthy subjects. Results of agglutination tests with four strains of coccoid anaerobes were used to estimate the probability that a patient suffers from Crohn's disease. The data on healthy subjects and patients with Crohn's disease were subjected to logistic discriminant analysis. With the methods and interpretation described, 52 % of the patients with Crohn's disease were recognized as 'definite' or 'probable' Crohn's disease and 14 % as 'suspected'. Only 1 % of the healthy subjects were classified as 'suspected' and none as 'definite' or 'probable' Crohn's disease.

INTRODUCTION

Recently, antibodies to anaerobic gram-positive coccoid rods belonging to species of *Eubacterium* and *Peptostreptococcus* were found in sera of a considerable percentage of patients with Crohn's disease (Wensinck, 1975, 1976; Wensinck & Van De Merwe, 1981). In healthy subjects and patients with other diseases than Crohn's disease (CD) these antibodies were found less frequently.

Agglutination of four strains of coccoid anaerobes was used to establish the diagnosis of CD. Tests with these strains showed different sensitivities and specificities and a simple interpretation of the agglutination results was not possible. The results with sera from standard groups of patients with CD and of healthy subjects were therefore subjected to discriminant analysis. The methods used and the interpretation of the results in terms of probability that a patient suffers from CD are reported in this paper.

MATERIALS AND METHODS

Agglutination reactions

The techniques used to demonstrate agglutinins to *Eubacterium contortum* (strains Me_{44} and Me_{47}), *E. rectale* (strain Me_{46}) and *Peptostreptococcus productus* (strain C_{18}) were described by Wensinck & Van De Merwe (1981). The results of agglutination reactions were scored as negative (0) or positive (1, 2 or 3, according to strength).

Patients and control subjects

Between 1 October 1975 and 1 February 1978, all consecutive patients with CD in the Departments of Internal Medicine and Surgery were studied. The data obtained at the first presentation of the patient were used. The diagnosis of CD* was based on generally accepted clinical, radiological and histological criteria (Lennard-Jones, Lockhart-Mummery & Morson, 1968; Kirsner, 1975). Patients in whom the differential diagnosis between CD and ulcerative colitis could not be made in the period mentioned were not included. A group of 114 patients with CD was thus collected, 43 men with a median age of 33 years (range 14–68) and 71 women with a median age of 29 years (range 17–74). Forty-three had ileal disease, 25 ileocolonic and 46 colonic disease. Fifty-three patients had undergone intestinal resections. Nineteen patients were taking salicylazosulphapyridine, ten were on corticosteroids and eight on corticosteroid enemas.

The group of healthy subjects consisted of 95 volunteers of the Red Cross Blood Transfusion Service, Rotterdam, 53 male with a median age of 38 years (range 23–64) and 42 female with a median age of 35 years (range 28–62). The agglutination reactions of patients and controls are given in Table 1.

Statistical methods

The interpretation of any test result rests upon three factors: the initial likelihood that a disease is present on the basis of the clinical evidence at hand (*a priori* probability) and the likelihood that a given test result occurs in the population with the disease and in that without the disease, respectively. Combining these likelihoods, usually calculated by applying Bayes' theorem, results in the *a posteriori* probability of the disease. For our results logistic discriminant analysis (Anderson, 1972) was used for discrimination between patients with CD and healthy subjects (HS) on the basis of the agglutination reactions. The *a posteriori* probability of CD can be written as:

$$\Pi(\mathrm{CD}|x_1, x_2, x_3, x_4) = \frac{1}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4)}. \quad (1)$$

In this formula, x_1–x_4 are the results of the agglutination reactions with strains Me_{44}, C_{18}, Me_{46} and Me_{47}, respectively. The expression is known as the multivariate logistic function. The coefficients β_0–β_4 were estimated with the maximum

* Diagnoses were established by M. Van Blankenstein and J. Dees, Department of Internal Medicine II (Professor Dr M. Frenkel), University Hospital Dijkzigt, Erasmus University, Rotterdam.

Table 1. *Frequency distribution of agglutination reactions in* 114 *patients with Crohn's disease (CD) and* 95 *healthy subjects (HS)*

Agglutination*	CD	HS	Agglutination*	CD	HS
0 0 0 0	21	75	2 0 0 0	1	5
0 0 0 1	0	1	2 0 0 1	1	0
0 0 0 2	2	1	2 0 0 2	2	0
0 0 0 3	1	0	2 0 1 0	1	0
0 0 1 0	1	0	2 0 3 0	1	0
0 0 2 0	1	0	2 1 3 0	1	0
0 0 3 0	3	0	2 1 3 2	1	0
0 0 3 1	1	0	2 2 0 0	3	1
0 0 3 2	1	0	2 2 3 0	2	0
0 1 0 0	0	2	2 3 0 1	1	0
0 1 0 1	0	1	2 3 3 2	1	0
0 1 0 2	1	0	3 0 0 0	3	2
0 1 1 0	1	0	3 0 0 1	1	0
0 1 3 2	1	0	3 0 0 2	1	0
0 2 0 0	2	0	3 0 1 0	2	1
0 2 3 0	1	0	3 0 3 0	2	0
0 3 0 0	1	1	3 1 0 2	2	0
0 3 0 1	2	0	3 1 2 0	1	0
0 3 0 2	1	0	3 1 2 3	1	0
0 3 1 0	1	0	3 1 3 0	2	0
0 3 3 0	1	0	3 1 3 3	2	0
1 0 0 0	3	4	3 2 0 2	1	0
1 0 0 2	0	1	3 2 0 3	1	0
1 0 2 0	1	0	3 2 3 0	1	0
1 2 3 2	1	0	3 3 0 0	3	0
1 3 0 0	2	0	3 3 0 1	1	0
1 3 0 2	1	0	3 3 0 3	1	0
1 3 1 0	1	0	3 3 2 3	1	0
1 3 2 2	1	0	3 3 3 0	9	0
1 3 3 0	1	0	3 3 3 1	1	0
1 3 3 1	1	0	3 3 3 2	4	0
1 3 3 2	1	0	3 3 3 3	3	0

* Agglutination reactions with strains Me_{44}, C_{18}, Me_{46} and Me_{47}, respectively, according to strength.

likelihood method (Cox, 1970) and the estimates are denoted by b_0–b_4, respectively. The estimated *a posteriori* probability of CD can thus be written as:

$$P(\mathrm{CD}|x_1, x_2, x_3, x_4) = \frac{1}{1+\exp\,(b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4)}. \tag{2}$$

The estimates b_0–b_4 are obtained from the reference samples (n_1 patients with CD and n_2 healthy subjects) with *a priori* probabilities $P(\mathrm{CD})$ and $P(\mathrm{HS})$ proportional to the sample sizes n_1 and n_2:

$$P(\mathrm{CD}) = \frac{n_1}{n_1+n_2}; \quad P(\mathrm{HS}) = \frac{n_2}{n_1+n_2}. \tag{3}$$

If *a posteriori* probabilities have to be calculated from other *a priori* probabilities $Q(\mathrm{CD})$ and $Q(\mathrm{HS})$, b_0 is replaced by d_0:

$$d_0 = b_0 - \ln \frac{n_2}{n_1} + \ln \frac{Q(\mathrm{HS})}{Q(\mathrm{CD})}. \tag{4}$$

Allocation rules and interpretation

After calculation of the estimated *a posteriori* probability of CD, several allocation rules can be used. It could be stated, for example, that a subject with agglutination reactions x $(= x_1, x_2, x_3, x_4)$ is classified as:

$$\left.\begin{aligned} &\text{non-CD if } 0 \leqq P(\mathrm{CD}|x) < 0{\cdot}5, \\ &\text{CD if } 0{\cdot}5 \leqq P(\mathrm{CD}|x) \leqq 1. \end{aligned}\right\} \tag{5}$$

When we classify as 'CD' when the *a posteriori* probability exceeds a certain value, say 0·8, and classify as 'non-CD' when the probability is below another value, say 0·2, we have an allocation rule with a region of doubt. Thus:

$$\left.\begin{aligned} &\text{non-CD if } 0 \leqq P(\mathrm{CD}|x) < 0{\cdot}2, \\ &\text{no decision (doubt) if } 0{\cdot}2 \leqq P(\mathrm{CD}|x) < 0{\cdot}8, \\ &\text{CD if } 0{\cdot}8 \leqq P(\mathrm{CD}|x) \leqq 1. \end{aligned}\right\} \tag{6}$$

On the basis of the results of the analysis (see Results), we divided the possible results of *a posteriori* probabilities in four classes with the following interpretation:

$$\left.\begin{aligned} &\text{no support for CD if } 0 \leqq P(\mathrm{CD}|x) < 0{\cdot}8, \\ &\text{suspected CD if } 0{\cdot}8 \leqq P(\mathrm{CD}|x) < 0{\cdot}95, \\ &\text{probable CD if } 0{\cdot}95 \leqq P(\mathrm{CD}|x) < 0{\cdot}99, \\ &\text{definite CD if } 0{\cdot}99 \leqq P(\mathrm{CD}|x) \leqq 1. \end{aligned}\right\} \tag{7}$$

RESULTS

Estimation of the coefficients and application of the allocation rules

With the results of the agglutination reactions x_1 (Me_{44}), x_2 (C_{18}), x_3 (Me_{46}) and x_4 (Me_{47}) of 114 patients with CD and 95 healthy subjects, the estimates b_0–b_4 of the coefficients β_0–β_4 in (1) were:

$$b_0 = 1{\cdot}33. \tag{8}$$

$$\left.\begin{aligned} b_1 &= -0{\cdot}45, \\ b_2 &= -0{\cdot}95, \\ b_3 &= -2{\cdot}32, \\ b_4 &= -1{\cdot}04. \end{aligned}\right\} \tag{9}$$

The coefficient b_0 is connected with *a priori* probabilities:

$$P(\mathrm{CD}) = 114/209 = 0{\cdot}55; \quad P(\mathrm{HS}) = 0{\cdot}45.$$

For equal *a priori* probabilities $Q(\mathrm{CD}) = Q(\mathrm{HS}) = 0{\cdot}5$, coefficient b_0 is replaced by d_0 according to (4), thus:

$$d_0 = 1{\cdot}51. \tag{10}$$

The other coefficients remain unchanged. The 95 % confidence intervals for these coefficients are:

$$\left.\begin{aligned} -0{\cdot}86 &< \beta_1 < -0{\cdot}05 \\ -1{\cdot}51 &< \beta_2 < -0{\cdot}38 \\ -4{\cdot}23 &< \beta_3 < -0{\cdot}41 \\ -1{\cdot}77 &< \beta_4 < -0{\cdot}31 \end{aligned}\right\}. \tag{11}$$

From (11) it is concluded that coefficients β_1–β_4 differ significantly from 0 ($\alpha = 0{\cdot}05$). With the estimated coefficients (9) and (10) *a posteriori* probabilities were calculated for all possible agglutination reactions (Table 2). The results for the elements from the two samples of patients with CD and healthy subjects are given in Table 3. When allocation rule (5) is used, classifications are obtained as summarized in Table 4. If allocation (6) is used, a classification is obtained as given in Table 5. From this table it is evident that about 20 % of the elements in both samples are not classified.

With the use of allocation rule (7) *a posteriori* probabilities are obviously interpreted more realistically (Table 6). In this situation, 42 % of the patients with CD are classified as 'definite CD', 10 % as 'probable CD', 14 % as 'suspected CD' and 34 % as 'no support for CD'. Of the healthy subjects, 99 % are classified in this latter category, whereas only 1 % is classified as 'suspected CD' and none as 'probable' or 'definite CD'.

Verification

The results of allocation were presented for the elements, which were also used for estimation of the coefficients of the logistic function. It would be better to evaluate the classification in new samples from the patients with CD and healthy subjects. As these samples were not available, a split-sample method was used. The sample consisting of 114 patients with CD was randomly divided into two groups of 57 subjects. The sample consisting of 95 healthy subjects was divided into a group of 48 and one of 47 subjects. With the agglutination reactions of the first subsample of 57 patients with CD and 48 healthy subjects, new coefficients were estimated. The results, for equal *a priori* probabilities were:

$$\left.\begin{aligned} d_0 &= 1{\cdot}75, \\ b_1 &= -0{\cdot}56, \\ b_2 &= -0{\cdot}94, \\ b_3 &= -2{\cdot}01, \\ b_4 &= -1{\cdot}66. \end{aligned}\right\} \tag{12}$$

The *a posteriori* probabilities for the elements of both groups, as determined with these coefficients, are given in Table 7. From this table it is seen that the results

Table 2. A. posteriori *probabilities* $P(CD|x)$ *of CD at* a priori *probability of* 0·5 *for all possible agglutination results with strains* Me_{44}, C_{18}, Me_{46} *and* Me_{47}

Agglutination*	$P(CD\|x)$	Agglutination	$P(CD\|x)$	Agglutination	$P(CD\|x)$	Agglutination	$P(CD\|x)$
0 0 0 0	0·18	1 0 0 0	0·26	2 0 0 0	0·35	3 0 0 0	0·46
0 0 0 1	0·38	1 0 0 1	0·50	2 0 0 1	0·61	3 0 0 1	0·71
0 0 0 2	0·64	1 0 0 2	0·73	2 0 0 2	0·81	3 0 0 2	0·87
0 0 0 3	0·83	1 0 0 3	0·89	2 0 0 3	0·92	3 0 0 3	0·95
0 0 1 0	0·69	1 0 1 0	0·78	2 0 1 0	0·85	3 0 1 0	0·90
0 0 1 1	0·86	1 0 1 1	0·91	2 0 1 1	0·94	3 0 1 1	0·96
0 0 1 2	0·95	1 0 1 2	0·97	2 0 1 2	0·98	3 0 1 2	0·99
0 0 1 3	0·98	1 0 1 3	0·99	2 0 1 3	0·99	3 0 1 3	0·99
0 0 2 0	0·96	1 0 2 0	0·97	2 0 2 0	0·98	3 0 2 0	0·99
0 0 2 1	0·98	1 0 2 1	0·99	2 0 2 1	0·99	3 0 2 1	1
0 0 2 2	0·99	1 0 2 2	1	2 0 2 2	1	3 0 2 2	1
0 0 2 3	1	1 0 2 3	1	2 0 2 3	1	3 0 2 3	1
0 0 3 0	1	1 0 3 0	1	2 0 3 0	1	3 0 3 0	1
0 0 3 1	1	1 0 3 1	1	2 0 3 1	1	3 0 3 1	1
0 0 3 2	1	1 0 3 2	1	2 0 3 2	1	3 0 3 2	1
0 0 3 3	1	1 0 3 3	1	2 0 3 3	1	3 0 3 3	1
0 1 0 0	0·36	1 1 0 0	0·47	2 1 0 0	0·58	3 1 0 0	0·69
0 1 0 1	0·62	1 1 0 1	0·72	2 1 0 1	0·80	3 1 0 1	0·86
0 1 0 2	0·82	1 1 0 2	0·88	2 1 0 2	0·92	3 1 0 2	0·95
0 1 0 3	0·93	1 1 0 3	0·95	2 1 0 3	0·97	3 1 0 3	0·98
0 1 1 0	0·85	1 1 1 0	0·90	2 1 1 0	0·93	3 1 1 0	0·96
0 1 1 1	0·94	1 1 1 1	0·96	2 1 1 1	0·98	3 1 1 1	0·98
0 1 1 2	0·98	1 1 1 2	0·99	2 1 1 2	0·99	3 1 1 2	0·99
0 1 1 3	0·99	1 1 1 3	1	2 1 1 3	1	3 1 1 3	1
0 1 2 0	0·98	1 1 2 0	0·99	2 1 2 0	0·99	3 1 2 0	1
0 1 2 1	0·99	1 1 2 1	1	2 1 2 1	1	3 1 2 1	1
0 1 2 2	1	1 1 2 2	1	2 1 2 2	1	3 1 2 2	1
0 1 2 3	1	1 1 2 3	1	2 1 2 3	1	3 1 2 3	1
0 1 3 0	1	1 1 3 0	1	2 1 3 0	1	3 1 3 0	1
0 1 3 1	1	1 1 3 1	1	2 1 3 1	1	3 1 3 1	1
0 1 3 2	1	1 1 3 2	1	2 1 3 2	1	3 1 3 2	1
0 1 3 3	1	1 1 3 3	1	2 1 3 3	1	3 1 3 3	1
0 2 0 0	0·60	1 2 0 0	0·70	2 2 0 0	0·78	3 2 0 0	0·85
0 2 0 1	0·81	1 2 0 1	0·87	2 2 0 1	0·91	3 2 0 1	0·94
0 2 0 2	0·92	1 2 0 2	0·95	2 2 0 2	0·97	3 2 0 2	0·98
0 2 0 3	0·97	1 2 0 3	0·98	2 2 0 3	0·99	3 2 0 3	0·99
0 2 1 0	0·94	1 2 1 0	0·96	2 2 1 0	0·97	3 2 1 0	0·98
0 2 1 1	0·98	1 2 1 1	0·99	2 2 1 1	0·99	3 2 1 1	0·99
0 2 1 2	0·99	1 2 1 2	0·99	2 2 1 2	1	3 2 1 2	1
0 2 1 3	1	1 2 1 3	1	2 2 1 3	1	3 2 1 3	1
0 2 2 0	0·99	1 2 2 0	1	2 2 2 0	1	3 2 2 0	1
0 2 2 1	1	1 2 2 1	1	2 2 2 1	1	3 2 2 1	1
0 2 2 2	1	1 2 2 2	1	2 2 2 2	1	3 2 2 2	1
0 2 2 3	1	1 2 2 3	1	2 2 2 3	1	3 2 2 3	1
0 2 3 0	1	1 2 3 0	1	2 2 3 0	1	3 2 3 0	1
0 2 3 1	1	1 2 3 1	1	2 2 3 1	1	3 2 3 1	1
0 2 3 2	1	1 2 3 2	1	2 2 3 2	1	3 2 3 2	1
0 2 3 3	1	1 2 3 3	1	2 2 3 3	1	3 2 3 3	1
0 3 0 0	0·79	1 3 0 0	0·86	2 3 0 0	0·90	3 3 0 0	0·94
0 3 0 1	0·92	1 3 0 1	0·94	2 3 0 1	0·96	3 3 0 1	0·98
0 3 0 2	0·97	1 3 0 2	0·98	2 3 0 2	0·99	3 3 0 2	0·99
0 3 0 3	0·99	1 3 0 3	0·99	2 3 0 3	1	3 3 0 3	1
0 3 1 0	0·97	1 3 1 0	0·98	2 3 1 0	0·99	3 3 1 0	0·99

Table 2 (*cont.*)

0 3 1 1	0·99	1 3 1 1	0·99	2 3 1 1	1	3 3 1 1	1
0 3 1 2	1	1 3 1 2	1	2 3 1 2	1	3 3 1 2	1
0 3 1 3	1	1 3 1 3	1	2 3 1 3	1	3 3 1 3	1
0 3 2 0	1	1 3 2 0	1	2 3 2 0	1	3 3 2 0	1
0 3 2 1	1	1 3 2 1	1	2 3 2 1	1	3 3 2 1	1
0 3 2 2	1	1 3 2 2	1	2 3 2 2	1	3 3 2 2	1
0 3 2 3	1	1 3 2 3	1	2 3 2 3	1	3 3 2 3	1
0 3 3 0	1	1 3 3 0	1	2 3 3 0	1	3 3 3 0	1
0 3 3 1	1	1 3 3 1	1	2 3 3 1	1	3 3 3 1	1
0 3 3 2	1	1 3 3 2	1	2 3 3 2	1	3 3 3 2	1
0 3 3 3	1	1 3 3 3	1	2 3 3 3	1	3 3 3 3	1

* Agglutination reactions with strains Me_{44}, C_{18}, Me_{46} and Me_{47} respectively.

Table 3. *Frequency distribution of* a posteriori *probabilities* $\hat{P}(CD|x)$ *of CD at* a priori *probability of* 0·5 *in patients with CD and healthy subjects*

$P(CD\|x)$	Crohn's disease	Healthy subjects
0–< 0·10	0	0
0·10–< 0·20	21	75
0·20–< 0·30	3	4
0·30–< 0·40	1	8
0·40–< 0·50	3	2
0·50–< 0·60	0	0
0·60–< 0·70	6	2
0·70–< 0·80	5	3
0·80–< 0·90	9	0
0·90–< 0·95	7	1
0·95–< 0·99	11	0
0·99–1	48	0
Total	114	95

Table 4. *Classification matrix for allocation rule* (5)

		Population of origin	
		CD	HS
Population of allocation	CD	86	6
	HS	28	89
	Total	114	95

Table 5. *Classification matrix for allocation rule* (6)

		Population of origin	
		CD	HS
Population of allocation	CD	75	1
	Doubt	18	19
	HS	21	75
	Total	114	95

Table 6. *Frequency distribution of* a posteriori *probabilities* $P(CD|x)$ *in four classes for allocation rule* (7) *in patients with CD and healthy subjects (HS)*

$P(CD|x)$	Interpretation	Number (%) of CD	Number (%) of HS
0–< 0·80	No support for CD	39 (34)	94 (99)
0·80–< 0·95	Suspected CD	16 (14)	1 (1)
0·95–< 0·99	Probable CD	11 (10)	0 (0)
0·99–1	Definite CD	48 (42)	0 (0)
Total		114 (100)	95 (100)

Table 7. *Frequency distribution of* a posteriori *probabilities of CD,* $P(CD|x)$ *for the elements from the first sub-sample of* 57 *patients with CD and* 48 *HS and the second sub-sample of* 57 *patients with CD and* 47 *HS, with use of coefficients* (12)

	Sub-sample 1		Sub-sample 2	
$P(CD|x)$	CD	HS	CD	HS
0–< 0·10	0	0	0	0
0·10–< 0·20	9	38	12	37
0·20–< 0·30	0	3	3	1
0·30–< 0·40	1	3	0	4
0·40–< 0·50	1	1	2	2
0·50–< 0·60	2	0	1	0
0·60–< 0·70	0	0	0	0
0·70–< 0·80	4	2	3	1
0·80–< 0·90	5	1	2	2
0·90–< 0·95	5	0	5	0
0·95–< 0·99	6	0	6	0
0·99–1	24	0	23	0
Total	57	48	57	47

of both groups are similar. It is concluded, therefore, that the coefficients (9) and (10) obtained with the original samples n_1 and n_2 may be used.

A posteriori *probability of CD at* a priori *probability* $\neq 0{\cdot}5$

The estimations of probabilities of CD and the corresponding interpretations in this paper are based on *a priori* probabilities of 0·5. For application at other *a priori* probabilities for all individuals, coefficient d_0 as well as the interpretations of the *a posteriori* probabilities have to be adjusted. For situations with considerable individual variation of *a priori* probabilities, the coefficient d_0 cannot be adjusted individually because a reliable interpretation of the *a posteriori* probabilities can be made only on the basis of an evaluation of the results. With the present material this is not possible.

DISCUSSION

With the methods described, 52% of patients with known CD could be recognized with the agglutination reactions as 'definite CD' or 'probable CD'. None of the healthy subjects was classified in these categories. Moreover, 14% of the patients with CD were classified as 'suspected CD' compared to only 1% of

healthy subjects. About one-third of the patients with CD were classified as 'no support for CD' compared to 99% of healthy subjects. From these results we conclude that application of the agglutination reactions in combination with the interpretation given yields an improvement in discrimination of individuals into groups of patients with CD and healthy subjects. If this system is used for diagnostic purposes, it should discriminate between 'CD' and 'non-CD'. The extrapolation from the group 'healthy subjects' to 'non-CD' is only allowed when there are no relevant differences in agglutination reactions between these groups. It has been shown previously that this is true for a large number of 'control' diseases, but in cirrhosis of liver and coeliac disease the results require a different interpretation (Wensinck & Van De Merwe, 1981).

It should be noted that the test is evaluated with the samples which were also used for estimation of the coefficients of the discriminant function. This difficulty is partly overcome by verification via the split-sample method, but ideally the predictive value of a diagnostic test for clinical use should be evaluated via a comparison of the patient population with a similar population that only differs by the absence of the particular disease. The final classification of patients as 'CD' or 'non-CD', however, requires long-term studies as the average period of time between onset of symptoms of Crohn's disease and diagnosis is 4 years (Dyer & Dawson, 1970; Brandes & Eulenburg, 1976; Mekhjian *et al.* 1979).

REFERENCES

Anderson, J. A. (1972). Separate logistic discrimination. *Biometrika* **59**, 19.

Brandes, J.-W. & Eulenburg, F. (1976). Der lange Weg zur Diagnose Morbus Crohn. *Zeitschrift für Gastroenterologie* **14**, 400.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Dyer, N. H. & Dawson, A. M. (1970). Diagnosis of Crohn's disease. A continuing source of error. *British Medical Journal* **i**, 735.

Kirsner, J. B. (1975). Problems in the differentiation of ulcerative colitis and Crohn's disease of the colon: the need for repeated diagnostic evaluation. *Gastroenterology* **68**, 187.

Lennard-Jones, J. E., Lockhart-Mummery, H. E. & Morson, B. C. (1968). Clinical and pathological differentiation of Crohn's disease and proctocolitis. *Gastroenterology* **54**, 1162.

Mekhjian, H. S., Switz, D. M., Melnyk, C. S., Rankin, G. B. & Brooks, R. K. (1979). Clinical features and natural history of Crohn's disease. *Gastroenterology* **77**, 898.

Wensinck, F. (1975). The faecal flora of patients with Crohn's disease. *Antonie van Leeuwenhoek* **41**, 214.

Wensinck, F. (1976). Faecal flora of Crohn's patients. Serological differentiation between Crohn's disease and ulcerative colitis. In *The Management of Crohn's Disease*, pp. 103–5. Amsterdam: Excerpta Medica.

Wensinck, F. & Van De Merwe, J. P. (1981). Serum agglutinins to *Eubacterium* and *Peptostreptococcus* species in Crohn's and other diseases. *Journal of Hygiene* **87**, 13.

CHAPTER 4

THE PERFORMANCE OF LOGISTIC DISCRIMINATION ON MYOCARDIAL INFARCTION DATA, IN COMPARISON WITH SOME OTHER DISCRIMINANT ANALYSIS METHODS.

The Performance of Logistic Discrimination on Myocardial Infarction Data, in Comparison with some other Discriminant Analysis Methods. P. I. M. Schmitz, J. D. F. Habbema and J. Hermans. Stat. in Med. 2 (1983) 199 - 205.

STATISTICS IN MEDICINE, VOL. 2, 199–205 (1983)

THE PERFORMANCE OF LOGISTIC DISCRIMINATION ON MYOCARDIAL INFARCTION DATA, IN COMPARISON WITH SOME OTHER DISCRIMINANT ANALYSIS METHODS

P. I. M. SCHMITZ

Institute of Biostatistics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

J. D. F. HABBEMA

Institute of Public Health and Social Medicine, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

AND

J. HERMANS

Department of Medical Statistics, University of Leiden, Wassenaarseweg 80, 2333 AL Leiden, The Netherlands

SUMMARY

The Imminent Myocardial Infarction study Rotterdam (IMIR), concerns patients who visit their general practitioners and have complaints suspected to be of cardiac origin. The study aims to develop a protocol for diagnosing myocardial infarction without laboratory assistance. We use the IMIR data, consisting of continuous and binary variables, to compare the diagnostic performance of logistic discrimination with some other discriminant analysis techniques. We discuss which characterizations of mixed data sets may give indications for an appropriate choice from among the alternative methods for discrimination.

KEY WORDS Discriminant analysis Myocardial infarction Mixtures of variables Logistic discrimination Kernel model Linear discrimination Quadratic discrimination

1. INTRODUCTION

We consider the performance of four methods of discriminant analysis for mixed continuous and discrete data; we restrict attention to comparisons with only two populations. We describe an empirical dataset, to which all our discriminant analysis methods will be applied, in Section 2, the discriminant analysis methods in Section 3, and the performance measures in Section 4. We present the results in Section 5, and a discussion in Section 6.

In a two-groups discriminant analysis, we use the observation of a random variable X on a sample element to answer the question, 'to which one of two distinct populations, Π_1 or Π_2, does the element belong?'. Generally, the probability structure of the problem contains three basic

0277–6715/83/020199–07$01.00 *Received September 1982*

components:

(i) the population priors $P(\Pi_1)$ and $P(\Pi_2) = 1 - P(\Pi_1)$
(ii) the distribution of X conditional on Π_i: $P(X|\Pi_i)$, $i = 1, 2$
(iii) the probability of being a member of Π_i given X: $P(\Pi_i|X)$, $i = 1, 2$.

Basically, one solves the problem by evaluation of the two conditional probabilities $P(\Pi_i|X)$, also called posterior probabilities. The probability $P(\Pi_1|X)$ is, according to Bayes theorem,

$$P(\Pi_1|X) = \frac{P(X|\Pi_1)P(\Pi_1)}{P(X|\Pi_2)P(\Pi_2)+P(X|\Pi_1)P(\Pi_1)}$$

and $P(\Pi_2|X) = 1 - P(\Pi_1|X)$.

For the calculation of $P(\Pi_1|X)$ one needs, in addition to the priors $P(\Pi_i)$, the distributions $P(X|\Pi_i)$. In applications one has to choose these distributions, which describe the variability of X within each population. One estimates the parameters of these distributions from training samples from the two populations. Different choices for $P(X|\Pi_i)$ lead to different discriminant analysis methods. All methods studied, except logistic discrimination (see section 3), reduce to the estimation of the densities $P(X|\Pi_i)$. The logistic method concerns the direct estimation of the posterior probability $P(\Pi_1|X)$. For a general introduction to discriminant analysis, see Reference 1.

2. DESCRIPTION OF THE DATA

In the Imminent Myocardial Infarction Rotterdam (IMIR) study, data have been gathered from patients who visit their general practitioners and have complaints suspected to be of cardiac origin.[2] Special interest focused on the question of whether diagnosis of acute myocardial infarction (AMI) is possible without laboratory assistance. To quantify the diagnostic uncertainty in a particular case, we used Fisher's linear discriminant analysis in the IMIR study. For a series of 92 definite cases of AMI and for 1211 non-cases drawn from patients seen by some general practitioners for coronary-type symptoms, information was available for several characteristics. We chose the following variables for use in the diagnostic rule:

x_1, sex (binary)
x_2, age, in years (continuous)
x_3, chest pain within 48 hours (binary)
x_4, duration of chest pain of more than 30 minutes (binary)
x_5, stabbing chest pain (binary)
x_6, chest pain primary symptom (binary)
x_7, clammy skin (binary)
x_8, heart rate, in beats/min (continuous)
x_9, premature beats more than 3/min (binary)
x_{10}, systolic minus diastolic blood pressure, in mmHg (continuous)
x_{11}, systolic blood pressure less than 110 mmHg (binary)
x_{12}, chest pain indicator (binary)

The data structure consists of a mixture of nine binary and three continuous variables. We excluded some cases with incomplete observations, and chose a random subset from the non-cases. Thus, the data studied in this paper consist of 84 infarction cases and 392 non-infarction cases. The prior probability of infarction, $P(\text{AMI}) = P(\Pi_1)$ was either fixed as 0·5 (equal) or as 0·1 (roughly

proportional to the number of AMI and non-AMI cases in the study). In calculating the performance measures (see Section 4) for each of the applied discriminant analysis techniques (see Section 3), we used either all cases both as training and as test data (this means that we estimated posterior probabilities by the resubstitution method), or we randomly split the cases into a training group and an independent test group (of equal size), with the evaluation based on the posterior probabilities of the test data.

3. THE DISCRIMINANT ANALYSIS METHODS USED

We applied four discrimination methods to the IMIR data, two of them, the logistic discrimination (LOG) and the kernel approach (KER), are meant for mixed data structures, whereas the other two models, the classical linear discriminant analysis of Fisher (LDA) and the quadratic discriminant analysis (QDA), are based on continuous distributions.

Linear and quadratic discriminant analyses assume multinormal densities with equal or unequal covariance matrices. With estimation of the mean vectors and the covariance matrices by their usual maximum likelihood estimates, the estimated density for QDA becomes

$$P_{\mathrm{QDA}}(X|\Pi_i) = (2\pi)^{-1/2p}|S_i|^{-1/2}\exp\{-\tfrac{1}{2}(X-\bar{x}_i)^{\mathrm{T}}S_i^{-1}(X-\bar{x}_i)\}$$

with $\bar{x}_i$ the sample mean and S_i the sample covariance matrix.[3] We obtain the density $P_{\mathrm{LDA}}(X|\Pi_i)$ by replacing S_i with the pooled sample covariance matrix S.

Logistic discriminant analysis assumes the following expression for the posterior probabilities:

$$P_{\mathrm{LOG}}(\Pi_1|X) = \frac{1}{1+\exp\{-(\beta_0+\beta^{\mathrm{T}}X)\}}, \quad \beta^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_p)$$

The logistic method[4,5] covers a fairly large class of distributions. One estimates the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ by maximum likelihood.

The kernel method uses a kernel function K for mixed data, proposed by Habbema *et al.*[6]

$$P_{\mathrm{KER}}(X|\Pi_i) = \frac{1}{n_i}\sum_{k=1}^{n_i}\prod_{j=1}^{p}K_{\mathrm{type}(j)}(X_j; X_{ikj})$$

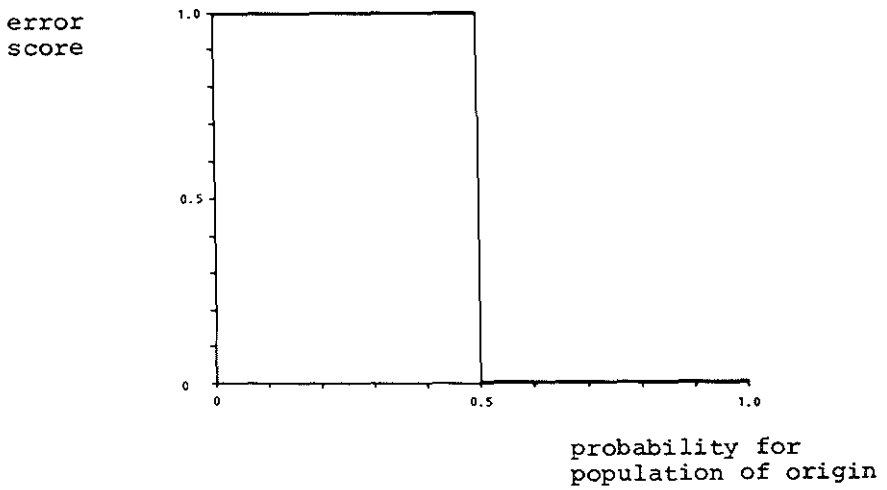
Kernel components $K_{\mathrm{type}(j)}$ are defined for each type of variable j; type may be continuous, ordinal, nominal or binary.

4. MEASURES OF PERFORMANCE FOR EVALUATION OF THE DISCRIMINANT ANALYSIS METHODS

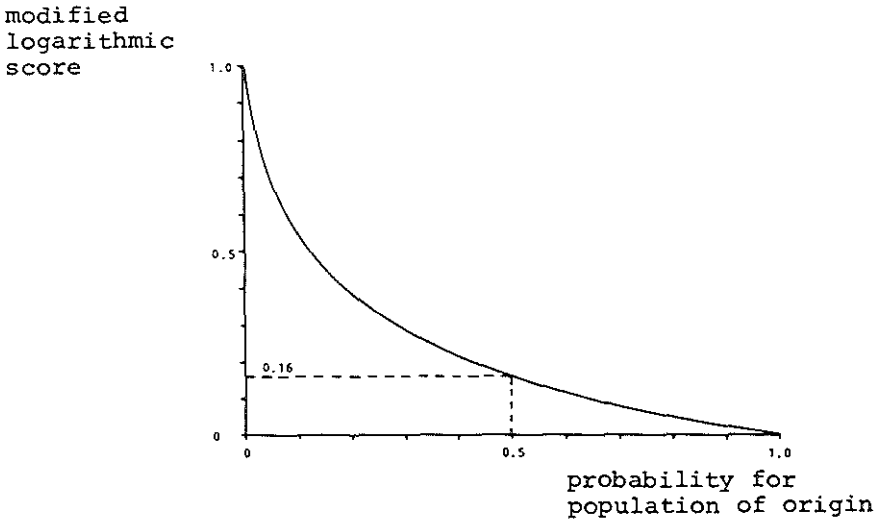
An important aspect of evaluating a discriminant analysis method is how well it separates the two populations. We use two measures of discriminatory ability in the present paper. (For a discussion of these and other measures of performance, see References 7–11.)

The first performance measure is based on the very common error rate scoring rule, which assigns a score of 1 for an incorrect allocation, and a score of 0 for a correct allocation (see Figure 1(a)).

Contrary to the error rate scoring rule, the other scoring rule, the modified logarithmic scoring rule, decreases continuously with increasing probability assigned to the actual population. To a close approximation, the modified logarithmic rule equals $-\ln\{P(i)+\varepsilon\}$, with $P(i)$ the probability assigned to the population of origin. (See Reference 9 for the exact formulation, and Figure 1(b) for a graphical representation.) In the present study we take $\varepsilon = 0{\cdot}02$.



(a)



(b)

Figure 1. Two measures of performance for discriminant analysis methods: (a) the error rate scoring rule; (b) the modified logarithmic scoring rule

By averaging the scores of the patients, we obtain the two measures of performance ERR (the error rate) and MLS (the modified logarithmic score). As one can easily see from the definition of the error scoring rule, the error rate is the proportion of elements not assigned to their actual population when allocating according the maximal posterior probability. Although the error rate is a much used measure, one may prefer the MLS because it takes into account differences in posterior probabilities in a continuous way.

5. RESULTS

Table I indicates the values for the performance measures on the IMIR data. The comparison of results of 'resubstitution' with the 'independent test sample' show (of course) a worse performance for the 'independent test sample'. An especially large difference occurs for the kernel method with priors 0·5. This is related to the data-adaptive character of the kernel method. Each training data observation contributes one term to the sum which determines the density estimate. In particular, probability mass is spread about this training observation. Calculation of the posterior probabilities with the resubstitution method will then be based for each observation on a density which is (relatively) too peaked for its group of origin. This leads to an overly optimistic picture of the performance of the method, when using a resubstitution method. Although all methods suffer more or less from this phenomenon, the kernel method is more sensitive to it. The effect can be avoided to a large extent with the 'leaving-one-out' technique in calculating the posterior probabilities for training data, as indeed is done by Lachenbruch and Mickey.[12] We did not use this technique, however, because our LDA, QDA and LOG algorithms did not have this facility. We will not further discuss the resubstitution results, but only the independent test data results. Three methods, LOG, LDA and KER, are nearly identical on both discriminatory performance measures and always better than QDA. The overall conclusion from this data set is that LOG, LDA and KER are preferable to QDA for these selected variables of the IMIR study. For classifying new cases, LOG and LDA are identical in regard to discriminatory ability, and simpler than KER, since they require only the evaluation of a linear function, whereas KER requires computer use.

Table I. Performance of four discriminant analysis methods applied to the IMIR data

	Discriminant Analysis method†	Performance measure‡ ERR	MLS
Resubstitution	LDA	0·14	0·07
Priors = 0·5	LOG	0·17	0·08
	QDA	0·13	0·08
	KER	0·04	0·03
Resubstitution	LDA	0·10	0·06
Priors = 0·1	LOG	0·09	0·06
	QDA	0·09	0·06
	KER	0·09	0·04
Independent	LDA	0·20	0·10
test samples	LOG	0·20	0·10
Priors = 0·5	QDA	0·22	0·15
	KER	0·21	0·11
Independent	LDA	0·15	0·09
test samples	LOG	0·13	0·08
Priors = 0·1	QDA	0·17	0·13
	KER	0·14	0·10

† LDA = Linear discriminant analysis, LOG = Logistic discrimination, QDA = Quadratic discriminant analysis, KER = Kernel method.
‡ ERR = Error rate, MLS = Modified logarithmic score.

6. DISCUSSION

This paper sought to gain insight into the performance of some discrimination techniques applied to mixed data. Most published comparative studies of discriminant analysis concern populations with either continuous or discrete data. Only a few, such as that of Titterington *et al.*,[13] deal with this topic. But, as Knill–Jones[14] remarks, 'it would be unwise to overgeneralize from the results of such a study until other discrimination problems have been examined in this way'. Indeed, it would be unwise to conclude only from application to the IMIR data, that for mixed data quadratic discrimination (QDA) performs worse than linear (LDA) or logistic (LOG) discrimination or the kernel approach (KER). There are, however, a few other mixed data studies which support the impression that QDA is never the best of these methods, and quite often is the poorest.[15] Also Knoke[16] does not recommend the use of QDA as an 'omnibus method'. If first-order interactions are present between the discrete variables and the two groups, QDA exhibits increased efficiency compared to LDA; but in such situations KER has approximately the same performance as QDA,[15] whereas the increased efficiency of QDA disappears with higher-order interactions.[16] This is consistent with our findings, because we did not detect any substantial interaction in the IMIR data, and QDA had the poorest performance. These results suggest a good strategy of making a further choice from LDA, LOG and KER.

In agreement with earlier results (for instance, Reference 17) LDA and LOG are almost similar in discriminatory ability. Some studies report a marginally better result for LOG, but, in our opinion, choice between these two methods in practice will generally depend on factors other than discriminatory ability. An advantage of LDA is its appropriateness for use in a stepwise selection approach of variables within a reasonable amount of computer time. LOG, moreover, fails to produce a classification rule in some situations (known as 'zero marginal proportions'), but may produce somewhat more 'reliable' estimates of the discriminant function coefficients. If one has to choose between LDA and LOG, it appears wise to start with LDA, to use it if desired for selection of variables, to add some of the main interaction terms if necessary, and finally to compare the performance of the best fitting LDA model with that of the corresponding LOG model.

Next, consider the comparison of LDA (LOG) with KER. Application to the IMIR data shows only marginal differences. Other studies which use the kernel method with mixed data show contradictory results. Titterington *et al.*[13] mention as a general conclusion the lack of success of some kernel methods. Vlachonikolis and Marriott[17] also report that kernel methods applied to their data sets were less effective than the parametric methods studied. However, Schmitz *et al.*[15] found that in many of the situations studied in their simulation study KER belongs to the methods with the best performance. It is not known at present which of the two approaches (LDA–LOG versus KER) generally shows the best discriminatory ability in a specific situation (with mixed data). It needs further study, by which a good characterization of the data (interaction and correlation structure, distance between populations, dimension, sample size, etc.) is a first requisite.

Investigations have shown that the linear methods LDA and LOG perform very poorly in the presence of interaction between the discrete variables and groups.[16–18] A location model[18] may be a better choice for those situations, but one can achieve the same improvement in performance by augmenting the linear discriminant function by the appropriate interaction terms. For that reason, many authors consider the LDA as a flexible and efficient method, whereas, if interaction is present, the inclusion of appropriate terms will suffice to reach an optimal performance. However, especially with a large number of variables, detection of interaction terms may require much effort. One may therefore prefer a kernel method, which automatically takes account of interaction structures, to the 'augmented' LDA technique. If the kernel method KER as described in this paper really can cope with such structures, it merits further study.

ACKNOWLEDGEMENT

The IMIR study group (head Dr. J. Lubsen) is acknowledged for making the IMIR tape available.

REFERENCES

1. Lachenbruch, P. A. *Discriminant analysis*, Hafner, New York, 1975.
2. Does, E. van der, and Lubsen, J. 'The Imminent Myocardial Infarction Rotterdam study', *Thesis*, Erasmus University Rotterdam, 1979.
3. Anderson, T. W. *Introduction to Multivariate Analysis*, Wiley, New York, 1958.
4. Day, N. E. and Kerridge, D. F. 'A general maximum likelihood discriminant', *Biometrics*, **23**, 313–323 (1967).
5. Anderson, J. A. 'Separate sample logistic discrimination', *Biometrika*, **59**, 19–35 (1972).
6. Habbema, J. D. F., Hermans, J. and Remme, J. 'Variable kernel density estimation in discriminant analysis', *Compstat Proceedings*, Physica Verlag, Wien, 178–185 (1978).
7. Habbema, J. D. F., Hilden, J. and Bjerregaard, B. 'The measurement of performance in probabilistic diagnosis I. The problem, descriptive tools, and measures based on classification matrices', *Methods of Information in Medicine*, **17**, 217–226 (1978).
8. Hilden, J., Habbema, J. D. F. and Bjerregaard, B. 'The measurement of performance in probabilistic diagnosis II. Trustworthiness of the exact values of the diagnostic probabilities', *Methods of Information in Medicine*, **17**, 227–237 (1978).
9. Hilden, J., Habbema, J. D. F. and Bjerregaard, B. 'The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnostic probabilities', *Methods of Information in Medicine*, **17**, 238–246 (1978).
10. Habbema, J. D. F. and Hilden, J. 'The measurement of performance in probabilistic diagnosis IV. Utility considerations in therapeutics and prognostics', *Methods of Information in Medicine*, **20**, 80–96 (1981).
11. Habbema, J. D. F., Hilden, J. and Bjerregaard, B. 'The measurement of performance in probabilistic diagnosis V. General recommendations', *Methods of Information in Medicine*, **20**, 97–100 (1981).
12. Lachenbruch, P. A. and Mickey, M. R. 'Estimation of error rates in discriminant analysis', *Technometrics*, **10**, 1–10 (1968).
13. Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. and Gelpke, G. J. 'Comparison of discrimination techniques applied to a complex dataset of head injured patients', *Journal of the Royal Statistical Society Series A*, **144**, 145–175 (1981).
14. Knill-Jones, R. P. Discussion of Reference 13.
15. Schmitz, P. I. M., Habbema, J. D. F., Hermans, J. and Raatgever, J. W. 'Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables', submitted for publication.
16. Knoke, J. D. 'Discriminant analysis with discrete and continuous variables', *Biometrics*, **38**, 191–200 (1982).
17. Vlachonikolis, I. G. and Marriott, F. H. C. 'Discrimination with mixed binary and continuous data', *Applied Statistics*, **31**, 23–31 (1982).
18. Krzanowski, W. J. 'Discrimination and classification using both binary and continuous variables', *Journal of the American Statistical Association*, **70**, 782–790 (1975).

CHAPTER 5

COMPARATIVE PERFORMANCE OF FOUR DISCRIMINANT ANALYSIS METHODS FOR MIXTURES OF CONTINUOUS AND DISCRETE VARIABLES

Comparative Performance of Four Discriminant Analysis Methods for Mixtures of Continuous and Discrete Variables. P. I. M. Schmitz, J. D. F. Habbema, J. Hermans and J. W. Raatgever. Comm. Stat. -Simula. Computa. 12 (1983) 727 - 751.

COMPARATIVE PERFORMANCE OF FOUR DISCRIMINANT ANALYSIS METHODS FOR MIXTURES OF CONTINUOUS AND DISCRETE VARIABLES

P.I.M.Schmitz
Institute of Biostatistics, Erasmus University Rotterdam, P.O.Box 1738, 3000 DR Rotterdam, The Netherlands

J.D.F.Habbema
Institute of Public Health and Social Medicine, Erasmus University Rotterdam

J.Hermans
Department of Medical Statistics, University of Leiden, Wassenaarseweg 80, 2333 AL Leiden, The Netherlands

J.W.Raatgever
Institute of Biostatistics, Erasmus University Rotterdam

Key Words and Phrases: discriminant analysis; mixtures of variables; linear discrimination; logistic discrimination; quadratic discrimination; kernel model; discriminatory ability; simulation.

ABSTRACT

The present study investigates the performance of four classification rules with respect to discriminatory ability for data consisting of a mixture of continuous and discrete variables. The four discriminant analysis methods are Fisher's linear discrimination, logistic discrimination, quadratic discrimination and a kernel model. Four measures of performance for evaluation of the classification rules are used: the error rate, the quadratic scoring rule, the modified logarithmic scoring rule and a doubt-based scoring rule. The mixed

data are obtained by generating from the fourdimensional normal distribution. Three of these variables were discretized. The results show that Fisher's linear discrimination and logistic discrimination have an almost similar performance. In most of the situations studied a choice from linear discrimination and the kernel model seems to be appropriate as far as discriminatory ability is concerned.

1. INTRODUCTION

Data from most medical and biological experiments consist of both continuous and discrete variables. In this respect, it is surprising that only very few studies for comparison of discriminant analysis methods are based on mixtures of discrete and continuous data. Therefore, the present study is concentrated on models which can take account of mixed data structures in discriminant analysis. More specific, the topic of the present paper is comparison of the performance of discriminant analysis methods for problems involving mixed data from two populations.

Three distinct approaches to the treatment of mixtures of continuous and discrete variables in discriminant analysis are described in recent literature.

(1) Day and Kerridge (1967) and Anderson (1972,1974) advocate the use of logistic discrimination, in which the probability of group membership is assumed to be a logistic function of the observed variables.

(2) Aitchison and Aitken (1976), Habbema, Hermans and Remme (1978a) and Titterington (1980) use the method of kernel density estimation with an appropriate kernel for mixed data.

(3) Krzanowski (1975,1977,1980) proposes a method based on the so-called location model.

Beside these methods for mixed data two other approaches are often applied:

(a) Discriminant analysis methods for continuous variables: especially Fisher's linear discriminant analysis, but also quadratic discrimination. In this approach, as in the logistic one, scores are assigned to binary and ordinal variables; nominal variables are replaced by dummy variables.

(b) Discriminant analysis methods originally developed for discrete variables, like the simple independence model. Here, continuous variables have to be converted into discrete ones by categorization.

The main interest is a comparison between the logistic discrimination and the kernel method for mixed data. Although the method of Krzanowski, based on the location model, would have a proper place in the study, it will not be considered further, mainly because it would be rather complicated to implement the location model in a simulation study.

In addition to logistic discrimination (LOG) and the kernel approach (KER), two other methods from category (a) have been used in this study: Fisher's linear discriminant analysis (LDA) and quadratic discrimination (QDA). Details about the four methods are given in section 2.

For LDA and QDA the estimative and not the predictive method was used (Aitchison et al.,1977). The predictive approach has a superior performance for small sample sizes. However for the sample sizes and dimension in the present study the differences in performance of the predictive and the estimate approach are very small (Hermans and Habbema, 1975).

A brief summary of results from earlier studies on mixed data is given in section 3. Section 4 deals with the measures of performance used in the present study. The design of the simulation study is described in section 5. The scores by the four methods on the measures of performance, as defined in section 4 are always calculated on independent test samples. These scores are used to evaluate the differences between the four discriminant analysis methods. The results of the simulation study are described in section 6. Finally, in section 7, the main conclusions will be formulated.

2. THE FOUR DISCRIMINANT ANALYSIS METHODS USED

In a two-groups discriminant analysis the observation of a random variable X on a sample element is used for answering the question to which one of two distinct populations the element belongs. More formally: associated with each sample element is a random vector (X,Y). Y is an index variable indicating the population of origin of the element (e.g. Y = 1 for population Π_1 , Y = 2 for population Π_2), and X is a p-dimensional feature vector that can be observed on the element.

Generally, the probability structure of the problem contains three basic components:

- the population priors $P(\Pi_1) = P(Y = 1)$ and $P(\Pi_2) = P(Y = 2) = 1 - P(\Pi_1)$,
- the distribution of X conditional on Π_i : $P(X|\Pi_i)$, i = 1,2,

and

- the probability of being a member of Π_i given X: $P(\Pi_i|X)$, $i = 1,2$.

Basically, the problem is solved by the evaluation of the two conditional probabilities $P(\Pi_i|X)$, also called posterior probabilities. The probability $P(\Pi_1|X)$ is, according to Bayes theorem,

$$P(\Pi_1|X) = \frac{P(X|\Pi_1)P(\Pi_1)}{P(X|\Pi_2)P(\Pi_2) + P(X|\Pi_1)P(\Pi_1)}$$

while $P(\Pi_2|X) = 1 - P(\Pi_1|X)$.

For the calculation of $P(\Pi_1|X)$ one needs, beside the priors $P(\Pi_i)$, the distributions $P(X|\Pi_i)$. In applications one has to choose these distributions $P(X|\Pi_i)$, which describe the variability of X within each population. Different choices for $P(X|\Pi_i)$ lead to different discriminant analysis methods. All four methods studied, except logistic discrimination, boil down to the estimation of the densities $P(X|\Pi_i)$. The logistic method is set up for the direct estimation of the posterior probability $P(\Pi_1|X)$. Linear and quadratic discriminant analysis (LDA and QDA respectively) assume multinormal densities with equal or unequal covariance matrices. Estimating the mean vectors and the covariance matrices by their usual maximum likelihood estimates, the estimated density for QDA becomes

$$P_{QDA}(X|\Pi_i) = (2\pi)^{-\frac{1}{2}p}|S_i|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(X-\bar{x}_i)'S_i^{-1}(X-\bar{x}_i)\}$$

with $\bar{x}_i$ the sample mean and S_i the sample covariance matrix (Anderson, 1958, chapter 6). Capital X indicates the stochastic variable; small x indicates the observation on X.

The density $P_{LDA}(X|\Pi_i)$ is obtained by replacing S_i by the pooled sample covariance matrix S. In logistic discriminant analysis (LOG) the following expression for the posterior probabilities is assumed:

$$P_{LOG}(\Pi_1|X) = \frac{1}{1 + \exp\{-(\beta_0 + \beta'X)\}}, \quad \beta' = (\beta_1, \beta_2, \ldots, \beta_p)$$

A fairly large class of distributions is covered by this logistic method (Day and Kerridge, 1967; Anderson, 1972). The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are estimated with the maximum likelihood method. An adapted version of the logistic regression program described by Lee (1974) has been used.

In the kernel estimation method, a probability mass (the so-called kernel function $K^{(p)}$) is placed around each training sample point x_{ij} from population Π_i (i=1,2; j=1,2,...,n_i). The density is estimated by the average of the n_i kernel functions (see Habbema et al., 1978a):

$$P_{KER}(X|\Pi_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} K^{(p)}(X;x_{ij},u_i).$$

A kernel method is fully defined by two specifications:
(a) The choice for the functional shape of the kernel $K^{(p)}$.
(b) The method of estimating the smoothness parameter or window size u_i.
For the situation of mixed data, the only feasible approach is to define the multivariate kernel function the product of p kernel components:

$$K^{(p)}(X;x_{ij},u_i) = \prod_{k=1}^{p} K_{type(k)}(X_k;x_{ijk},u_i)$$

The index type (k) indicates the type of variable k: 'type'= continuous, ordinal, nominal or binary. The choice for the general form of the kernel function components is, according to Habbema et al.(1978a):

$$K_{type(k)}(X_k;x_{ijk},u_i) =$$

$$= \frac{1}{C_{type(k)}(u_i)} \cdot u_i^{\,d^2_{type(k)}(X_k,x_{ijk})}$$

Thus, each type of variable has its own normalizing factor C and distance measure d^2, see Appendix A.
Having defined C and d^2, the functional shape of the kernel is fully defined.

Next, a method of estimating the smoothness parameter has to be chosen. Maximization of the likelihood function

$$L(u_i) = \prod_{j=1}^{n_i} P_{KER}(x_{ij}|\Pi_i)$$

results in the value zero for u_i. Therefore, this straight-forward likelihood function has been modified (see Habbema et al.,1974) according to the

leaving-one-out method, into:

$$L^{*}(u_i) = \prod_{j=1}^{n_i} \frac{1}{n_i - 1} \sum_{\substack{t=1 \\ t \neq j}}^{n_i} K^{(p)}(x_{ij}; x_{it}, u_i)$$

This method leads to effective u_i estimates (Hand, 1982).

3. LITERATURE REVIEW

Several studies about the performance of discriminant analysis methods have been published. Most of these studies concern populations with either only continuous or only discrete data (see e.g. Remme, Habbema and Hermans(1980)). What follows is a short review of some recent studies which consider populations with mixed continuous and discrete data.

The objective of the study described by Titterington et al.(1981), is a discussion of the application of several methods of discriminant analysis to a large dataset. All four methods as used in the present study (and more techniques than these) were used. The dataset is a series of 1000 patients with severe head injury. Four measures of separation and two measures of reliability were calculated. Among the conclusions from this case study are the robustness of LDA, the lack of success of KER, and the comparibility of the results from LOG and LDA.

Knoke (1982) compares several approaches, among which LDA, QDA, LOG and an 'augmented LDA-approach', i.e. a LDA augmented with appropriate higher-order terms (interaction terms; squared variables). The methods are applied among others on a three-dimensional problem consisting of two mutually independent binary variables and one normally distributed continuous variable with different means within each cell of the two 2x2-tables defined by the binary variables. In such a way, situations with interactions among these variables and between these variables and groups can be studied. Important conclusions are: (a) If interactions are not present, LDA is a remarkably robust method with a good performance. (b) If first-order interactions are present, QDA is to be preferred; however it loses its advantage when higher-order interactions exist. Therefore, Knoke does not recommend the use of QDA as an omnibus method. The augmented LDA-approach (possibly improved by using the LOG-model corresponding with the selected augmented-LDA) is recommended if interactions exist.

Vlachonikolis and Marriott (1982) applied LDA, LOG, some KER-methods and various modifications of LDA on two sets of data with binary and continuous variables. The modified LDA has the advantage of compensating for interactions due to the binary variables. They mention as the main conclusion that 'LDA is flexible and efficient, and where interaction is present, the inclusion of appropriate terms can remove its most serious disadvantage'. LOG performed marginally better than LDA. The mixed kernel methods were less effective than the parametric methods.

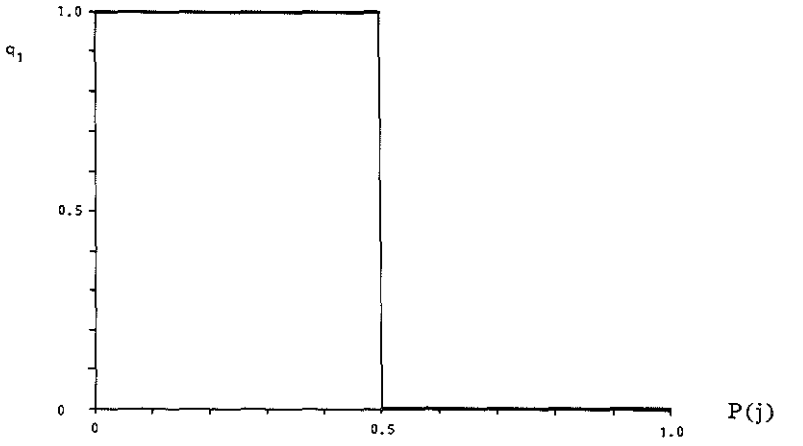
4. MEASURES OF PERFORMANCE FOR EVALUATION OF THE DISCRIMINANT ANALYSIS METHODS

When trying to assess a discriminant analysis method, the most important aspect is how well the populations are separated (in a probabilistic sense) by the method. If a method does this well, the method can be said to possess a good discriminatory ability. For a wide discussion of measures of performance, including the measures used in the present paper, see Habbema, Hilden and Bjerregaard (1978b,1981), Hilden, Habbema and Bjerregaard (1978a,1978b) and Habbema and Hilden (1981).

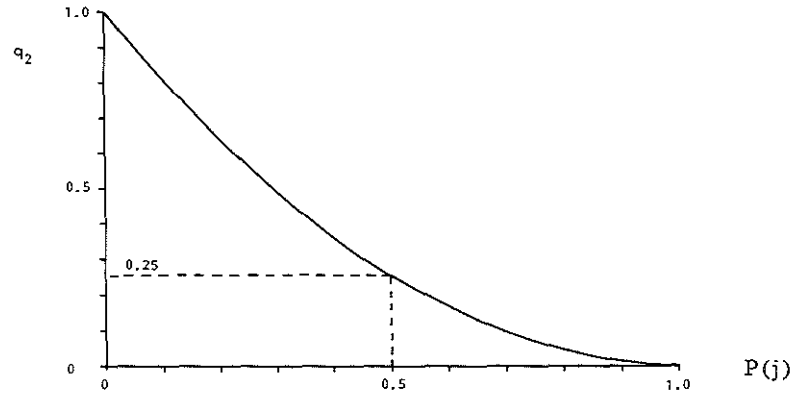
Let $P(\Pi_i|x_j)$ be the posterior probability assigned by a discriminant analysis method to population Π_i ($i = 1,2$) to a test element j with observation x_j, $j=1,2,\ldots,n$. Define $P(j)$ as the posterior probability on the population of origin for element j, so $P(j) = P(\Pi_1|x_j)$ if x_j originates from population Π_1 and $P(j) = P(\Pi_2|x_j) = 1 - P(\Pi_1|x_j)$ if x_j originates from population Π_2. Let $q_m = q_m\{P(j)\}$ be a 'scoring rule', i.e. a function of the assigned probability $P(j)$ for element j. Index m indicates the particular chosen measure for the scoring rule. Then the average performance is just the average score of all test elements on this measure:

$$Q_m = \frac{1}{n} \sum_{j=1}^{n} q_m\{P(j)\} \qquad (4.1)$$

The function q_m will be defined such that a higher score indicates worse performance (it is a 'penalty' score). Each discriminant analysis method gets a score on each performance measure for each situation simulated. Four different scoring rules for measuring discriminatory ability will now be defined (for graphical representation, see Figure 1).

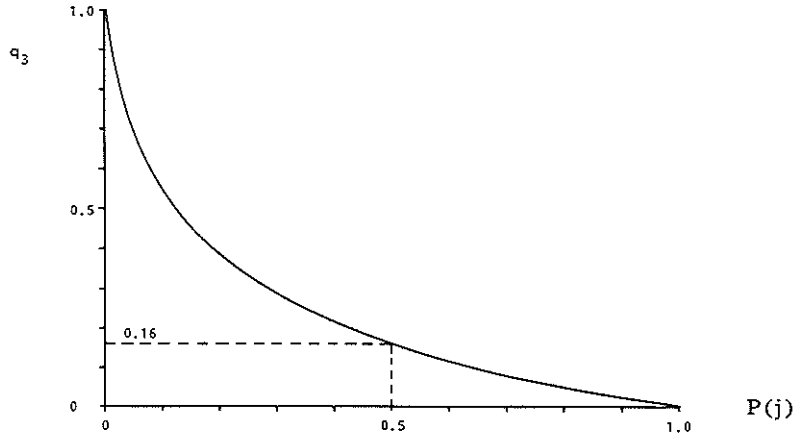


(a) The error scoring rule $q_1(P(j))$

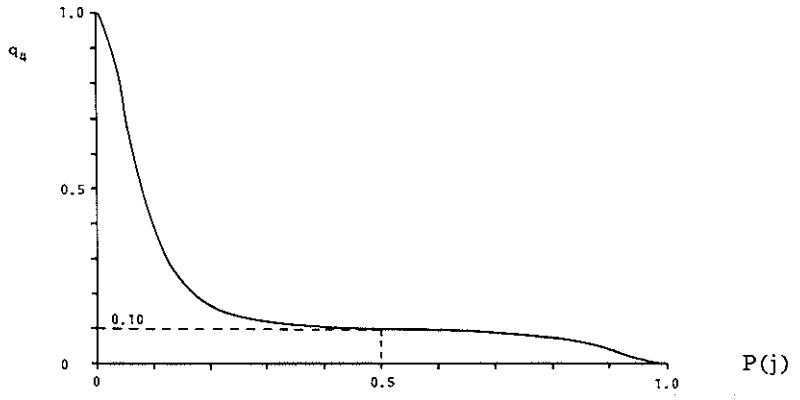


(b) The quadratic scoring rule $q_2(P(j))$

Figure 1. Four scoring rules for measuring discriminatory ability



(c) The modified logarithmic scoring rule $q_3(P(j))$



(d) The doubt-based scoring rule $q_4(P(j))$

Figure 1. (Continued)

Best known is the error scoring rule:

$$q_1\{P(j)\} = I\{P(j) < 0.5\}\ ,\ \text{with } I(\text{true}) = 1 \text{ and } I(\text{false}) = 0.$$

(Values of the posterior probability of exactly 0.5 are penalized with $\frac{1}{2}$). Contrary to q_1, the other three scoring rules all decrease in a continuous way with increasing probability assigned to the actual population.

The quadratic scoring rule penalizes deviations from 100% probability assignment to the actual population in a quadratic way:

$$q_2\{P(j)\} = (1 - P(j))^2$$

The modified logarithmic scoring rule penalizes mis-assignment of probability in a logarithmic way, but modified in order to avoid an infinite penalty for zero probabilities:

$$q_3\{P(j)\} = -\{\ln\{P(j) + \varepsilon(1-P(j))\} + \varepsilon\ln\{P(j) + (1-P(j))/\varepsilon\}\}\ /\ \{(\varepsilon - 1)\ln\varepsilon\}$$

Throughout the present study we take $\varepsilon = 0.02$.

At last, a so-called 'doubt-based' scoring rule is used. This scoring rule in principle makes a distinction between three probability ranges, i.e. very high, very low and in between ('doubt') posterior probabilities for the actual population. The formula is:

$$q_4\{P(j)\} = \frac{P(j)^{-b-1} + \delta^{-b}}{\{P(j)^{-b} + (1-P(j))^{-b} + \delta^{-b}\}^{1+1/b}}$$

For this study we take $\delta = 0.10$ and $b = 2$.

All four rules are standardized to run from 0 (the best possible score) to 1 (the worst possible score). Using these four scoring rules and expression (4.1), four measures of discriminatory ability are obtained:

1. The error rate ERR = Q_1.
2. The quadratic score QSC = Q_2.
3. The modified logarithmic score LSC = Q_3.
4. The doubt-based score DSC = Q_4.

As is easily seen from its definition, the error rate is the proportion of elements not assigned to their actual population when allocating according the maximal posterior probability. Although the error rate is much used, the other three

As is easily seen from its definition, the error rate is the proportion of elements not assigned to their actual population when allocating according the maximal posterior probability. Although the error rate is much used, the other three measures may be preferable, because of their taking account of posterior probabilities in a continuous way.

5. DESIGN OF THE SIMULATION STUDY

The study is restricted to discrimination between two populations. Only the dimension p = 4 of the distributions is used in the present simulations. Initially, the data are generated from two four-dimensional normal distributions:

population Π_1 : $N(\mu_1,\Sigma_1)$, with $\mu_1' = (0,0,0,0)$, and
population Π_2 : $N(\mu_2,\Sigma_2)$, with $\mu_2' = \mu(1,1,1,1)$.

Five values of the distance parameter μ and nine values of the covariance structures (Σ_1,Σ_2) are used in our study, see Table 1. For each combination of μ and (Σ_1,Σ_2) values, 16 runs were made with training samples of size 50 and test samples of size 50 in each population. Furthermore, 9 runs were made with training and test sample sizes of 100.

The value μ = 2.0 of the distance parameter was used for runs with sample sizes of 100 only. So, a total of 81 situations have been studied: 4x9 (for sample size 100) plus 5x9 (for sample size 50).

In order to generate pseudo-random N(0,1)-numbers, the NAG-subroutine G055DDF is used (see the NAG FORTRAN Library Manual, Mark 7; NAG Central Office, 7 Banbury Road, Oxford OX2 6NN, United Kingdom). For each run, for a certain set of parameter values, and for each sample (2 training, 2 test samples) a new starting value for the generator was used by means of a NAG-subroutine G05CBF. From the N(0,I)-distributed random vector $(x_1^*,x_2^*,x_3^*,x_4^*)$ a $N(\underline{\mu},\Sigma)$-distributed random vector (x_1,x_2,x_3,x_4) was obtained by a transformation based on Crout factorization of Σ (see e.g. Newman and Odell, 1971, section 5.1).

After generating the vector (x_1,x_2,x_3,x_4), the variables x_2,x_3 and x_4 were discretized into four, three and two categories respectively:

TABLE I

VALUES OF μ AND (Σ_1,Σ_2) OF THE TWO UNDERLYING FOUR-DIMENSIONAL NORMAL DISTRIBUTIONS $N(0,\Sigma_1)$ AND $N(\mu_2,\Sigma_2)$ WITH $\mu_2'=\mu(1,1,1,1)$

- Values for the distance parameter μ:
  0.4 ; 0.7 ; 1.0 ; 1.3 ; 2.0
- Values for the underlying covariance matrices (Σ_1,Σ_2):
  (I,I), $(I,4I)$, $(I,16I)$,
  $(\Sigma_{.5},\Sigma_{.5})$, $(\Sigma_{.9},\Sigma_{.9})$, $(I,\Sigma_{.5})$,
  $(I,\Sigma_{.9})$ $(\Sigma_{.5},\Sigma_{-.5})$, $(\Sigma_{.9},\Sigma_{-.9})$
- Explanation of covariance matrices:

$$\Sigma_{.5}=\begin{bmatrix}1.0 & 0.5 & 0.5 & 0.5\\ 0.5 & 1.0 & 0.5 & 0.5\\ 0.5 & 0.5 & 1.0 & 0.5\\ 0.5 & 0.5 & 0.5 & 1.0\end{bmatrix} \qquad \Sigma_{-.5}=\begin{bmatrix}1.0 & -0.5 & 0.5 & -0.5\\ -0.5 & 1.0 & -0.5 & 0.5\\ 0.5 & -0.5 & 1.0 & -0.5\\ -0.5 & 0.5 & -0.5 & 1.0\end{bmatrix}$$

Structures of $\Sigma_{.9}$ and of $\Sigma_{-.9}$ can be obtained by replacing 0.5 by 0.9 in structures $\Sigma_{.5}$ and $\Sigma_{-.5}$ respectively.

(1) $x_4' = 0$ if $x_4 < \delta_1$
$x_4' = 0$ if $x_4 \geq \delta_1$

(2) $x_3' = 0$ if $x_3 < \delta_2$
$x_3' = 1$ if $\delta_2 \leq x_3 < \delta_3$
$x_3' = 2$ if $x_3 \geq \delta_3$

(3) $x_2' = 0$ if $x_2 < \delta_4$
$x_2' = 1$ if $\delta_4 \leq x_2 < \delta_5$
$x_2' = 2$ if $\delta_5 \leq x_2 < \delta_6$
$x_2' = 3$ if $x_2 \geq \delta_6$.

The cut-off points $\delta_1,\ldots,\delta_6$ were chosen as follows:

$$\delta_1 = \tfrac{1}{2}\mu,\quad \delta_2 = 0,\quad \delta_3 = \mu,\quad \delta_4 = -\tfrac{1}{4}\mu,\quad \delta_5 = \tfrac{1}{2}\mu,\quad \delta_6 = 1\tfrac{1}{4}\mu.$$

In this way a 4-dimensional vector (x_1,x_2',x_3',x_4') was created with:

x_1 = normally distributed continuous variable
x_2' = ordinal variable with 4 categories, assigned scores 0, 1, 2, 3.
x_3' = variable with 3 categories (labels 0, 1, 2), to be treated as a nominal variable (see below)
x_4' = binary variable (outcomes 0 or 1).

As mentioned before, x_3' is considered a nominal variable: so considering the

labels 0, 1 and 2 as scores when applying LDA, QDA or LOG would not be correct. Therefore, x_3' was replaced by two dummy variables x_5 and x_6. LDA, QDA and LOG are now applied to $(x_1, x_2', x_5, x_6, x_4')$.

In each run the four discriminant analysis methods (section 2) were applied to the data of the two training samples. The performance measures (section 4) were calculated for the posterior probabilities of the elements of the independently generated test samples, created in each run. Means and standard deviations of the performance measures were calculated from the 16 (training-test sample sizes = 50) or 9 (training-test sample sizes = 100) runs for each set of values of the distance parameter and covariance structure.

6. RESULTS OF THE SIMULATION STUDY

The four discriminant analysis methods were compared on their discriminatory ability by applying a rank-order analysis of the scores on the four performance measures (the error rate ERR, the quadratic score QSC, the modified logarithmic score LSC, and the doubt-based score DSC). For each performance measure and each situation, the method with the best (i.e. lowest) score will get rank number 1, and the worst method will get rank number 4. Thus each method gets in total 4 x 81 = 324 rank numbers (4 measures, 81 situations). Taking the average over all 324 rank numbers we get as the average rank numbers:

LDA	LOG	QDA	KER
2.9	2.4	2.8	1.9

This suggests that the kernel approach is superior to the other methods, and that the linear (LDA), quadratic (QDA) and logistic (LOG) analyses perform about equally well.

This is of course a very crude first assessment, and deviations from this overall picture will arise by studying the distinct performance measures, sample sizes, underlying covariance structures and distances separately. It appeared that only slight deviations from the overall picture were obtained when dependence of average rank numbers on performance measure, on sample size, and on distance was analyzed. The only major source of deviation from the overall picture is the underlying covariance structure. The relevant results are given in Table II. Some striking conclusions from this table are:

TABLE II

INFLUENCE OF UNDERLYING COVARIANCE STRUCTURE ON THE AVERAGE RANK NUMBER FOR THE FOUR DISCRIMINANT ANALYSIS METHODS. THE AVERAGE REFERS TO 36 RANK NUMBERS: 9 SITUATIONS FOR EACH COVARIANCE STRUCTURE TIMES 4 PERFORMANCE MEASURES.

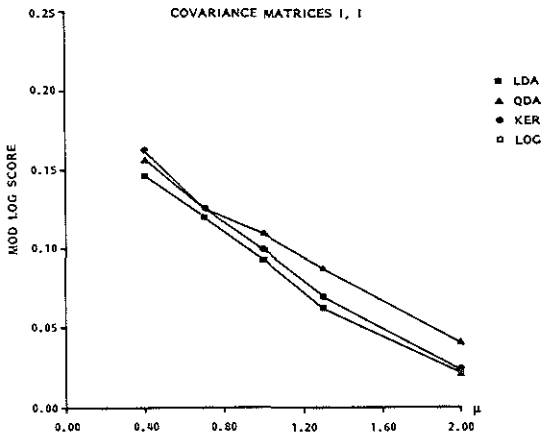
	LDA	LOG	QDA	KER
(I,I)	1.8	1.2	3.8	3.1
(I,4 I)	3.4	2.8	2.6	1.2
(I,16 I)	3.5	3.4	2.0	1.1
$(\Sigma_{.5}, \Sigma_{.5})$	1.9	1.3	3.7	3.1
$(\Sigma_{.9}, \Sigma_{.9})$	1.9	1.2	3.9	2.9
$(I, \Sigma_{.5})$	2.9	2.2	3.0	1.9
$(I, \Sigma_{.9})$	3.7	3.3	1.9	1.1
$(\Sigma_{.5}, \Sigma_{-.5})$	3.5	2.8	2.2	1.6
$(\Sigma_{.9}, \Sigma_{-.9})$	3.3	3.5	2.0	1.2

(a) LDA and LOG have a quite comparable performance, with LOG nearly always slightly better than LDA. Best ranks are obtained for equal underlying covariance structures.

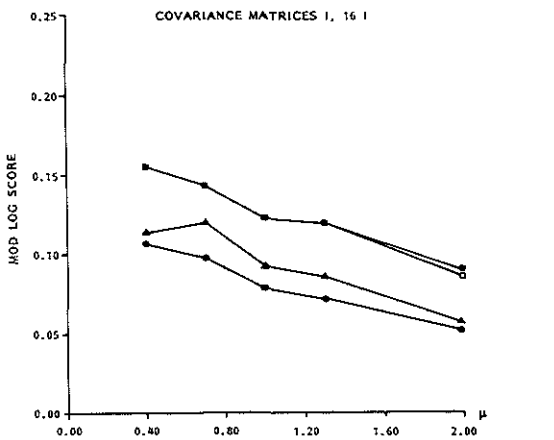
(b) Rank numbers of KER are always superior to QDA.

Further analysis of the relation between performance measure and average rank number, revealed that the modified logarithmic score LSC is the most representative measure, i.e. closest to the other three measures. Therefore, LSC has been used for some graphical illustrations of the results, see Figure 2. The score LSC, averaged over runs, is plotted in Figure 2 against the values of the distance parameter μ. There are only small differences in performance between sample sizes 50 and 100, as should be the case, because a factor 2 difference in sample size is not large. Therefore, only figures for sample size 100 are presented.

Figure 2 shows the results for five underlying covariance structures: (I,I), (I,16 I), $(I,\Sigma_{.9})$, $(\Sigma_{.9},\Sigma_{.9})$ and $(\Sigma_{.9},\Sigma_{-.9})$. The results for structures (I,4 I), $(\Sigma_{.5},\Sigma_{.5})$, $(I,\Sigma_{.5})$ and $(\Sigma_{.5},\Sigma_{-.5})$ were somewhere in between those for (I,I) and respectively (I,16 I), $(\Sigma_{.9},\Sigma_{.9})$, $(I,\Sigma_{.9})$ and $(\Sigma_{.9},\Sigma_{-.9})$.

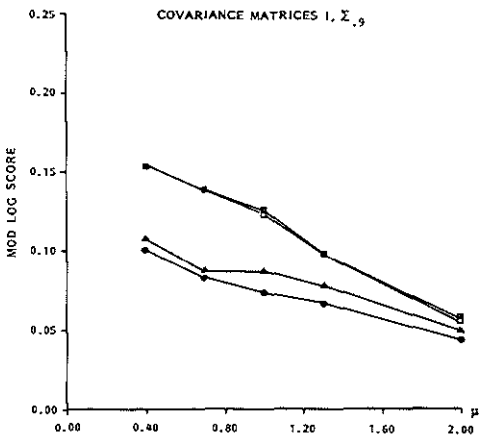


(a)

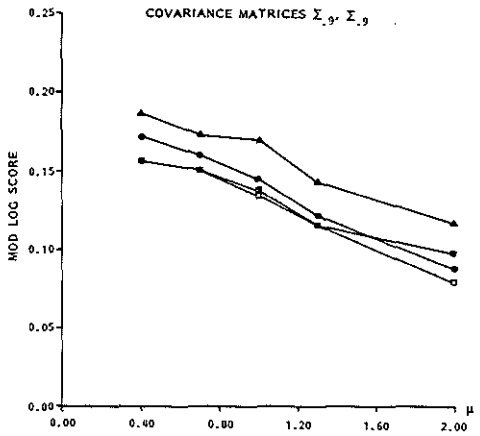


(b)

Figure 2. Modified logarithmic score LSC against distance parameter μ for training-test sample sizes 100 and five different covariance structures.

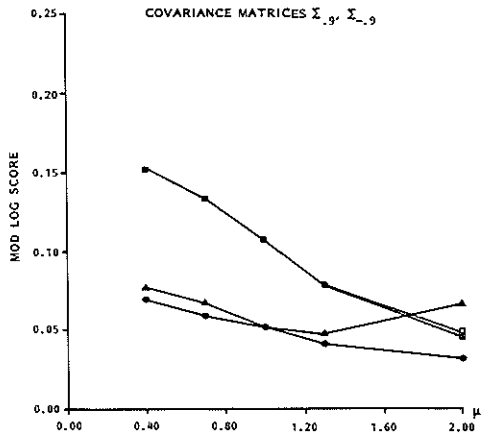


(c)



(d)

Figure 2. (Continued)



(e)

Figure 2. (Continued)

The performance improves in general (i.e. LSC decreases) with increasing distance, as should be the case. As an exception, LSC increases for QDA for $\mu = 2.0$ in Figure 2(e). This can be explained by the fact that only 2 instead of 9 runs were made for QDA in this situation. The estimated covariance matrices for QDA became singular in the other 7 runs. (In other situations these singularity did not occur).

The order in performance of the four discriminant analysis methods does not change with different μ-values. An exception is structure $(I, \Sigma_{.9})$ (Figure 2(c)), where LDA and LOG have a rather bad performance for small values of μ, but a better one for large values, comparible with QDA and KER.

From Figure 2 the following conclusions may be drawn:
(1) QDA performed worse than LDA and LOG for equal underlying covariance structures, but better for unequal covariance matrices. Except in a few situations (small distances, equal covariance matrices), KER was better than QDA.

(2) The differences between LDA and LOG are generally negligible. Only for large distances and equal covariance matrices (with non-zero covariances), LOG is somewhat better.
(3) Except for small distances and covariance structure ($\Sigma_{\cdot 9}, \Sigma_{-\cdot 9}$), KER showed the best discriminatory ability for unequal covariance structures. For equal covariance structures LDA and LOG were generally somewhat better than KER.
(4) In most situations KER shows the best separation. When KER is not the best, it is always reasonably close to the best discriminant analysis method for any underlying covariance structure.

7. CONCLUSIONS AND DISCUSSION

The simulation studies for discriminant analysis published thus far are based on either only continuous or only discrete data. The present study compares the performance of four classification methods (LDA, LOG, QDA and KER) for data consisting of a mixture of continuous and discrete variables. We used simulated data, generated from a 4-dimensional normal distribution; three of the four variables were discretized. The discriminatory ability of the methods is measured by the error rate ERR, the quadratic score QSC, the modified logarithmic score LSC and the doubt-based score DSC.

Several conclusions from the simulation results can be deduced:
(a) The differences in performance between training sample sizes 50 and 100 are very small. Also, the order in performance of the four methods remains nearly always unchanged with increasing distance between the populations. By far the largest source of deviation from the overall average performance was the underlying covariance structure. Therefore, the following conclusions relate to the nine distinct covariance structures.
(b) In agreement with earlier results from continuous and for discrete data, LDA and LOG were almost similar in discriminatory ability. Because LOG fails to produce a classification rule in some situations (indicated as 'zero marginal proportions'), use of LDA instead of LOG may have a certain preference.
(c) Next, the selection of a method from the remaining three 'best' ones: LDA, QDA and KER. The results for these methods can be summarized roughly by the following scheme:
('$\geq$' indicates a better than or approximately equal performance)

- unequal covariance structures: KER $\geq$ QDA > LDA, and
- equal covariance structures : LDA > KER $\geq$ QDA.

KER is always better than or approximately equal to QDA, so a final choice will be from KER and LDA.

A few final remarks have to be made.

(1) Our conclusions are based on a study of a necessary limited number of situations. Only moderate sample sizes and a rather low dimension p = 4 were considered in the present study. Caution is required in extrapolation of the results to other situations.

(2) As a criticism on the design of the present simulation study, it may be argued that the effect of discretization of multinormal data, rather than genuine mixtures of continuous and discrete variables is considered. However, it can be shown that if all the four normally distributed continuous variables are dichotomized, this completely corresponds with a Bahadur model (see Appendix B). Likewise, discretizing only a part of the normally distributed variables results in genuine mixtures of discrete and continuous variables. Two limitations can be mentioned here. (i) The characterization of the generated discretized data, other than in terms of the 'latent' normally distributed variables, is not a simple matter. (ii) By using an underlying normal distribution, three- or more-factor interactions cannot be generated.

(3) It must be noticed that the performance of LDA and LOG may be improved upon in practical applications. Addition of one or more interaction terms (simply the cross products of the original variables) in the linear discriminant function may give a substantial improvement in separation (Knoke,1982; Vlachonikolis and Marriott,1982).

Our main conclusion, based on this simulation study and recent literature (Knoke,1982; Vlachonikolis and Marriott,1982) is that a choice between LDA, if necessary augmented by appropriate interaction terms, and KER still has to be made. As mentioned before, the location model is not considered in this study. This method will probably perform similar to the augmented LDA (or LOG) approach, because these two approaches only differ considerably in the method of estimation of the parameters.

Especially when the number of variables is large, the location model becomes intractable, while detection of the best fitting augmented LDA-model may require much effort. Possibly, KER automatically takes account of inter-

action structures. Whether the kernel method really can cope well with higher-order interaction structures with many variables needs further study.

APPENDIX A

The kernel function in the kernel density estimator for mixed data, used in the present study is defined by:

$$K^{(p)}(X;x_{ij},u_i) = \prod_{k=1}^{p} K_{type(k)}(X_k;x_{ijk},u_i) =$$

$$= \prod_{k=1}^{p} \frac{1}{C_{type(k)}(u_i)} \cdot u_i^{\,d^2_{type(k)}(X_k,x_{ijk})}$$

For discrete variables we will denote the possible outcome categories by $X_k(t)$, $t = 1,2,\ldots,T_k$. The normalizing factor C is determined by

$$\sum_{t=1}^{T_k} K(X_k(t);x_{ijk},u_i) = 1.$$

The distances d^2 are defined for each type separately.

Binary component

$$d^2_{binary}(X_k,x_{ijk}) = \frac{(X_k - x_{ijk})^2}{s^2_{ik}}$$

$$s^2_{ik} = \{ \sum_{j=1}^{n_i} (x_{ijk} - \bar{x}_{ik})^2 \} / (n_i - 1)$$

Choose scores such that $| X_k(1) - X_k(2) | = 0$.

Then, $d^2_{binary} = 0$ if $X_k = x_{ijk}$, and

$d^2_{binary} = 1/s^2_{ik}$ if $X_k \neq x_{ijk}$, while $C_{binary}(u_i) = 1+u_i^{1/s^2_{ik}}$

Nominal component

$$d^2_{nominal}(X_k, x_{ijk}) = 0 \quad \text{if } X_k = x_{ijk}$$

$$= 1/s^2_{ik} \quad \text{if } X_k \neq x_{ijk}.$$

$$s^2_{ik} = \{ n_i - \frac{1}{n_i} \sum_{t=1}^{T_k} N^2_{ik}(t) \} / 2(n_i - 1)$$

$N_{ik}(t)$ = the number of observations out of $x_{i1k}, x_{i2k}, \ldots, x_{in_ik}$ with outcome category t.

$$C_{nominal}(u_i) = 1 + (T_k - 1) \cdot u_i^{1/s^2_{ik}}.$$

Ordinal component

The distance parameter d^2 and s^2_{ik} are defined as in the binary case. The normalizing factor becomes:

$$C_{ordinal}(u_i) = \sum_{t=1}^{T_k} u_i^{(X_k(t) - x_{ijk})^2 / s^2_{ik}}$$

Continuous component

Specifying the distance d^2 like in the binary case, the normalizing factor turns out to be for $0 < u_i < 1$:

$$C_{continuous}(u_i) = \int u_i^{(X_k - x_{ijk})^2 / s^2_{ik}} \, dX_k = \sqrt{(-\pi s^2_{ik} / \ln(u_i))}.$$

APPENDIX B

Discretization of multinormal distributed variables into binary data completely corresponds with a Bahadur model. For example, consider a situation with two-dimensional normal distributions:

in population Π_1 : $N(\mu_1, \Sigma_1)$, $\mu_1' = (0,0)$, $\Sigma_1 = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

in population Π_2 : $N(\mu_2, \Sigma_2)$, $\mu_2' = (2,2)$, $\Sigma_2 = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

After dichotomizing the two normally distributed variables (by using the cut-off point 1), calculating bivariate normal integrals, and reparameterization, it can be shown that this example completely corresponds with a Bahadur-parameterization with $\theta_1 = \theta_2 = 0.16$, $\rho_{12}=0.7$ in population Π_1, and $\theta_1 = \theta_2 = 0.84$, $\rho_{12}=0.7$ in population Π_2.

ACKNOWLEDGEMENTS

The authors would like to thank E. Kasanmoentalib from the University of Leiden, The Netherlands, for doing much of the programming work necessary for this study.

BIBLIOGRAPHY

Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. Biometrika 63, 413-420.

Aitchison, J., Habbema, J.D.F. and Kay, J.W. (1977). A critical comparison of two methods of statistical discrimination. Applied Statistics 26, 15-25.

Anderson, J.A. (1972). Separate sample logistic discrimination. Biometrika 59, 19-35.

Anderson, J.A. (1974). Diagnosis by logistic discriminant function: Further practical problems and results. Applied Statistics 23, 397-404.

Anderson, T.W. (1958). Introduction to multivariate analysis. John Wiley, New York.

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

Habbema, J.D.F., Hermans, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. Compstat Proceedings, Physica-Verlag, 101-110.

Habbema, J.D.F. Hermans, J. and Remme, J. (1978a). Variable kernel density estimation in discriminant analysis. Compstat Proceedings, Physica-Verlag, 178-185.

Habbema, J.D.F. and Hilden, J. (1981). The measurement of performance in probabilistic diagnosis IV. Utility considerations in therapeutics and prognostics. Methods of Information in Medicine 20, 80-96.

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. Methods of Information in Medicine 17, 217-226.

Habbema, J.D.F., Hilden, J. and Bjerregaard, B. (1981). The measurement of performance in probabilistic diagnosis V. General Recommendations. Methods of Information in Medicine 20, 97-100.

Hand, D.J. (1982). Kernel Discriminant Analysis. New York: Research Studies Press.

Hermans, J. and Habbema, J.D.F. (1975). Comparison of five methods to estimate posterior probabilities. EDV Medizin Biol. 6, 14-19.

Hermans, J., Habbema, J.D.F., Kasanmoentalib, T.K.D. and Raatgever, J.W. (1983). Manual for the ALLOC80 discriminant analysis program. Technical Report, Department of Medical Statistics, University of Leiden, The Netherlands.

Hilden, J., Habbema, J.D.F. and Bjerregaard, B. (1978a). The measurement of performance in probabilistic diagnosis II. Trusworthiness of the exact values of the diagnostic probabilities. Methods of Information in Medicine 17, 227-237.

Hilden,J., Habbema, J.D.F. and Bjerregaard, B. (1978b). The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnosic probabilities. Methods of Information in Medicine 17, 238-246.

Knoke, J.D. (1982). Discriminant analysis with discrete and continuous variables. Biometrics 38, 191-200.

Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous variables. Journal of the American Statistical Association 70, 782-790.

Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics 19, 191-200.

Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. Biometrics 36, 493-499.

Lee, E.T. (1974). A computer program for linear logistic regression analysis. Computer Programs in Biomedicine 4, 80-92.

Newman, T.G. and Odell, P.L. (1971). The generation of random variates. Griffin, London.

Remme, J, Habbema, J.D.F. and Hermans J. (1980). A simulative comparison of linear, quadratic and kernel discrimination. Journal of Statistical Computation and Simulation 11, 87-106.

Titterington, D.M. (1980). A comparative study of kernel-based density estimates for categorical data. Technometrics 22, 259-268.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F. and Gelpke, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society Series A 144, 145-175.

Vlachonikolis, I.G. and Marriott, F.H.C. (1982). Discrimination with mixed binary and continuous data. Applied Statistics 31, 23-31.

Received by Editorial Board member December, 1982, Revised April, 1983.
Recommended by Chien-Pai Han, University of Texas at Arlington, Arlington, TX
Refereed by James D. Knoke, University of North Carolina, Chapel Hill, NC

CHAPTER 6

A SIMULATION STUDY OF THE PERFORMANCE OF FIVE DISCRIMINANT ANALYSIS METHODS FOR MIXTURES OF CONTINUOUS AND BINARY VARIABLES

A Simulation Study of the Performance of Five Discriminant Analysis Methods for Mixtures of Continuous and Binary Variables. P. I. M. Schmitz, J. D. F. Habbema and J. Hermans. J. Stat. Comput. Simul. 23 (1985) 69 - 95.

J. Statist. Comput. Simul., 1985, Vol. 23, pp. 69–95
0094-9655/85/2302-0069 $18.50/0

Printed in Great Britain

A Simulation Study of the Performance of Five Discriminant Analysis Methods for Mixtures of Continuous and Binary Variables

P. I. M. SCHMITZ

Institute of Biostatistics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

J. D. F. HABBEMA

Institute of Public Health and Social Medicine, Erasmus University Rotterdam, The Netherlands

and

J. HERMANS

Department of Medical Statistics, University of Leiden, The Netherlands

(*Received April 27, 1984; in final form July 22, 1985*)

The present study investigates the performance of five discrimination methods for data consisting of a mixture of continuous and binary variables. The methods are Fisher's linear discrimination, logistic discrimination, quadratic discrimination, a kernel model and an independence model. Six-dimensional data, consisting of three binary and three continuous variables, are simulated according to a location model. The results show an almost identical performance for Fisher's linear discrimination and logistic discrimination. Only in situations with independently distributed variables the independence model does have a reasonable discriminatory ability for the dimensionality considered. If the log likelihood ratio is non-linear with respect to its continuous and binary part, the quadratic discrimination method is substantial better than linear and

logistic discrimination, followed by the kernel method. A very good performance is obtained when in every situation the better one of linear and quadratic discrimination is used.

KEY WORDS: Discriminant analysis, mixtures of variables, linear discrimination, logistic discrimination, quadratic discrimination, kernel model, independence model, discriminatory ability, simulation.

1. INTRODUCTION

Most of the studies in which discrimination methods have been compared, deal either with continuous data or with discrete data. Only few studies consider mixtures of continuous and discrete variables.

The present paper describes the results of a simulation study in which five discriminant analysis methods are applied to two-group discrimination problems involving six-dimensional mixed data. Three of the variables are continuous, and three are binary. In two-group discriminant analysis, the observation of a random variable X on a sample element is used in order to infer to which one of two distinct populations, Π_1 and Π_2, the element belongs. Let $P(X|\Pi_1)$ and $P(X|\Pi_2)$ be the distribution of X in populations Π_1 and Π_2, respectively. Suppose that the prior probability is $P(\Pi_1)$ for Π_1, and $P(\Pi_2)=1-P(\Pi_1)$ for Π_2. The inference problem is solved by the evaluation of the two conditional or posterior probabilities $P(\Pi_1|X)$ and $P(\Pi_2|X)$. With Bayes theorem:

$$P(\Pi_1|X)=\frac{P(X|\Pi_1)P(\Pi_1)}{P(X|\Pi_1)P(\Pi_1)+P(X|\Pi_2)P(\Pi_2)} \tag{1.1}$$

and

$$P(\Pi_2|X)=1-P(\Pi_1|X). \tag{1.2}$$

Four of the five discriminant analysis methods under study are characterized by their way of estimating the densities $P(X|\Pi_i)$, $i=1,2$. The fifth method, logistic discriminant analysis, is based on direct estimation of $P(\Pi_1|X)$. The discrimination techniques are described in Section 2. The discriminatory ability of the methods

evaluated is assessed with two measures: the error rate and the modified logarithmic score, see Section 3.

The usefulness of a simulation study highly depends on the quality of its design. Therefore, great care has been taken in the formulation of the design. The design is described in Section 4. The data structure is specified in two steps. A 2^3-dimensional multinomial vector is specified for the three binary variables in terms of a loglinear model. A three-dimensional normally distributed vector is specified for each state of the two multinomial vectors. Thus, a mixture of three binary and three continuous variables is obtained. Fifty-nine different specifications of the data structure are used in the simulation study. The results are presented and analysed in Section 5 and discussed in Section 6.

2. THE DISCRIMINANT ANALYSIS METHODS

Five discrimination techniques are used in the present simulation study:

a) LDA, Linear Discriminant Analysis

b) QDA, Quadratic Discriminant Analysis

c) LOG, Logistic Discriminant Analysis

d) KER, a KERnel model

e) IND, an INDependence model.

In deriving the posterior probabilities (1.1) linear and quadratic discriminant analysis (LDA and QDA) assume multinormal densities with equal and unequal covariance matrices respectively.

The density estimate used in QDA is

$$P_{\mathrm{QDA}}(X|\Pi_i)=(2\pi)^{-1/2p}|S_i|^{-1/2}\exp\{-\tfrac{1}{2}(X-\bar{x}_i)^T S_i^{-1}(X-\bar{x}_i)\},\ i=1,2. \tag{2.1}$$

where $\bar{x}_i$ and S_i are the sample mean vectors and covariance matrices. The density $P_{\mathrm{LDA}}(X|\Pi_i)$ is obtained by replacing S_i by the pooled sample covariance matrix S (Anderson, 1958, Chapter 6). Although this use of LDA and QDA is theoretically based on continuous variables, the methods are also frequently applied to

discrete and mixed data. In these situations, scores are assigned to binary and ordinal variables; nominal variables have to be replaced by dummy variables.

The predictive analogues of LDA and QDA have not been used, although they have a better performance for small sample sizes (Aitchison *et al.*, 1977). For the sample sizes in the present study, differences between the predictive and the estimative approach will be very small (Hermans and Habbema, 1975).

In logistic discriminant analysis (LOG), the posterior probability $P_{\text{LOG}}(\Pi_1|X)$ is modelled according to

$$P_{\text{LOG}}(\Pi_1|X)=\frac{1}{1+\exp\{-(\beta_0+\beta' X)\}},\ \beta'=(\beta_1,\beta_2,\ldots,\beta_p). \quad (2.2)$$

A fairly large class of distributions is covered by this logistic method (Day and Kerridge, 1967; Anderson, 1972). The parameters $\beta_0,\beta_1,\beta_2,\ldots,\beta_p$ are estimated with the maximum likelihood method. This implicitly implies a proportionality of the prior probabilities to the sizes of the training samples (n_i from population Π_i; $i=1,2$):

$$P(\Pi_1)=\frac{n_1}{n_1+n_2}. \quad (2.3)$$

For other prior probabilities, the estimate $\hat{\beta}_0$ of the constant term β_0 has to be replaced by:

$$\hat{\beta}_0^*=\hat{\beta}_0+\ln\frac{P(\Pi_1)}{P(\Pi_2)}-\ln\frac{n_1}{n_2}. \quad (2.4)$$

In the kernel method (KER), a probability mass (the so-called kernel function $K^{(p)}$), is placed around each training sample point x_{ij} from population Π_i ($i=1,2$; $j=1,2,\ldots,n_i$).

The density is estimated by the average of the n_i kernel functions:

$$P_{\text{KER}}(X|\Pi_i)=\frac{1}{n_i}\sum_{j=1}^{n_i}K^{(p)}(X;x_{ij},u_i) \quad (2.5)$$

with u_i a smoothness parameter or window size.

A kernel method is fully defined by two specifications:

a) The functional shape of the kernel $K^{(p)}$.

b) The method of estimating the smoothness parameter u_i.

For the situation of mixed data, a feasible approach is to define the multivariate kernel function as the product of p one-dimensional kernel components:

$$K^{(p)}(X; x_{ij}, u_i) = \prod_{k=1}^{p} K_{\text{type}(k)}(X_k; x_{ijk}, u_i). \tag{2.6}$$

The index "type(k)" indicates the type of variable k: "type" = continuous, ordinal, nominal or binary. The general form of the one-dimensional kernel function components is chosen according to Habbema *et al.* (1978a):

$$K_{\text{type}(k)}(X_k; x_{ijk}, u_i) = \frac{1}{C_{\text{type}(k)}(u_i)} \cdot u_i d^2_{\text{type}(k)}(X_k, x_{ijk}). \tag{2.7}$$

Thus, each type of variable has its own normalizing factor C and distance measure d^2 (see Appendix A of Schmitz *et al.*, 1983b). By C and d^2, the functional shape of the kernel is fully defined. The smoothness parameter is estimated according to a leaving-one-out modification of the likelihood function (Habbema *et al.*, 1974):

$$L^*(u_i) = \prod_{j=1}^{n_i} \frac{1}{n_i - 1} \sum_{\substack{t=1 \\ t \neq j}}^{n_i} K^{(p)}(x_{ij}; x_{it}, u_i). \tag{2.8}$$

This method leads to effective u_i estimates (Hand, 1982).

The last method is the independence model (IND). The computer programme we used for this method calls for discretizing the continuous variables. The method uses a "zero-avoiding device":

$$P_{\text{IND}}(X|\Pi_i) = \prod_{j=1}^{p} \frac{n_i(x_j) + 1/(2c_j)}{n_i + \frac{1}{2}} \tag{2.9}$$

where c_j = number of categories of variable x_j, $n_i(x_j)$ = the number of sample elements with score x_j on variable j, and n_i = sample size.

Two other approaches for mixed data have not been used. The method based on a so-called location model (Krzanowski, 1975, 1977, 1980) would be rather complicated to implement in the simulation study. The same applies to the approach with modified linear functions (Vlachonikolis and Marriott, 1982).

A special purpose computer program has been written which contains all five methods. For the part for logistic discriminant analysis an adapted version of the logistic regression program described by Lee (1974) has been used. For the kernel method we used the relevant subroutines of the ALLOC80 program described by Hermans *et al.* (1983).

3. MEASURES OF PERFORMANCE

Two measures for discriminatory ability will be used in evaluating the performance of the five discriminant analysis methods. See the series of papers by Habbema *et al.* (1978), Hilden *et al.* (1978a), Hilden *et al.* (1978b), Habbema and Hilden (1981) and Habbema *et al.* (1981), for a discussion of these and other measures of performance.

The first performance measure is the error rate (ERR). This is the proportion of incorrect allocations when allocating according to maximal posterior probability. The second performance measure is the modified logarithmic score (MLS) (Hilden *et al.*, 1978b). This measure takes account of posterior probabilities in a continuous way, and therefore may be more sensitive for detecting differences between methods than the error rate. To a close approximation MLS equals $(1/n)\sum_i -\ln(P(i)+0.02)$, with $P(i)$ the probability assigned to the disease category of origin, and summation over all cases.

4. DESIGN OF THE SIMULATION STUDY

In a mixed data simulation study there is an abundance of possible distributions from which the data can be generated. The chosen distributions should reflect the important features of a mixed dataset when evaluating different discriminant analysis methods. Therefore, both the design of the simulation study and the characterization of a mixed dataset will receive much attention.

4.1 The underlying model

The data will be generated according to a particular choice of the location model, see Krzanowski (1977, 1983). The general form of this model for the two-population situation is as follows. Suppose that q discrete variables $X_1, X_2, \ldots, X_q$ and c continuous variables $X_{q+1}, X_{q+2}, \ldots, X_{q+c}$ (say $p=q+c$) are observed. The q discrete variables are assumed to define the multinomial vector $\mathbf{M}$ containing s possible states, and p_{im} is the probability of observing state m $(m=1,2,\ldots,s)$ in population Π_i $(i=1,2)$. Next, the c continuous variables are assumed to follow a distinct multivariate normal distribution $N(\boldsymbol{\mu}_m^{(i)}, \Sigma_m^{(i)})$ for each state m.

Many parameters have to be chosen. Therefore, one has to restrict the study in several aspects. The first limitation (as already mentioned before) will be in only considering dimension $p=6$. The second limitation is in the type of variables. A mixture of three binary and three continuous variables will be considered, so $q=3$, $c=3$, $p=6$, $s=2^q=8$. The further choices for the parameters p_{im}, $\boldsymbol{\mu}_m^{(i)}$ and $\Sigma_m^{(i)}$ will be made such that the most important characteristics of the mixed dataset can be studied.

These characteristics are:

a) The *interaction structure* of the binary variables within both populations.

b) The *correlation structure* between the continuous and the binary variables, especially the linearity or non-linearity of the "binary" and the "continuous" part of the log likelihood ratio.

c) The *distance* between the two populations.

Before describing these features, the *sample sizes* of the training samples (from which the five discrimination rules will be derived) and the test samples (which are used to estimate the performance) will be specified.

4.2 Sample sizes

Four training sample sizes have been chosen. These may be characterized as "moderate" (100, 100), "small" (25, 25), "large" (500, 500) and "unequal" (25, 100). Footnote a of Table I gives the test sample sizes and the number of runs associated with the training and test sample sizes.

TABLE I

Descriptions of simulations

Simulation number	Sample size[a]		Binary variables[b]		Continuous variables[c]		
	Pop 1	Pop 2	Pop 1	Pop 2	Type	Difference	Distance
1	100	100	F1	F2	I	0.0	0.38
2						0.5	0.56
3						1.0	0.85
4						1.5	1.10
5	100	100	F2	F3	I	0.0	0.69
6						0.5	0.78
7						1.0	0.98
8						1.5	1.16
9	100	100	F1	F4	I	0.0	0.98
10						0.5	1.03
11						1.0	1.14
12						1.5	1.25
13	100	100	F2	F3	I	0.0	0.34
14						0.5	0.53
15						1.0	0.84
16						1.5	1.09
17	100	100	F2	F3	I	1.0	0.84
18			S2	S2	L	1.0	0.80
19			T2	T2	L	1.0	0.80
20	25	25	F2	F3	I	1.0	0.84
21			S2	S2	L	1.0	0.80
22			T2	T2	L	1.0	0.80
23	500	500	F2	F3	I	1.0	0.84
24			S2	S2	L	1.0	0.80
25			T2	T2	L	1.0	0.80
26	25	100	F2	F3	I	1.0	0.84
27			S2	S2	L	1.0	0.80
28			T2	T2	L	1.0	0.80
29	100	100	S1	S3	N	±1.0	0.89
30			S2	S4			1.11
31			T1	T3			0.90
32			T2	T4			0.97
33	25	25	S1	S3	N	±1.0	0.89
34			S2	S4			1.11
35			T1	T3			0.90
36			T2	T4			0.97

TABLE I (continued)

Simulation number	Sample size[a]		Binary variables[b]		Continuous variables[c]		
	Pop 1	Pop 2	Pop 1	Pop 2	Type	Difference	Distance
37	500	500	S1	S3	N	± 1.0	0.89
38			S2	S4			1.11
39			T1	T3			0.90
40			T2	T4			0.97
41	25	100	S1	S3	N	± 1.0	0.89
42			S2	S4			1.11
43			T1	T3			0.90
44			T2	T4			0.97
45	100	100	S5	S5	L	1.0	0.80
46			S5	S6	N	± 1.0	1.11
47			T5	T5	L	1.0	0.80
48			T5	T6	N	± 1.0	0.97
49	100	100	G1	G1	L	1.0	0.79
50			G1	G2	L	1.0	0.80
51			G1	G3	N	± 1.0	1.08
52			G3	G4	N	± 1.0	1.09
53	100	100	F2	F3	N	± 1.0	0.84
54			S2	S2	N	± 1.0	0.79
55			T2	T2	N	± 1.0	0.79
56	100	100	S1	S3	L	1.0	0.89
57			S2	S4	L	1.0	1.11
58			T1	T3	L	1.0	0.90
59			T2	T4	L	1.0	0.97

[a]Test sample sizes are (200, 200). Exception: for training sample sizes (500, 500) the test sample sizes are (500, 500). The number of runs r is 9. Exceptions: for training sample sizes (25, 25) $r = 16$, for sizes (500, 500) $r = 2$.

[b]Loglinear parameterization in a 2^3-table according to (4.1) in text. F1–F4 are first-order models with $\alpha_1 = \alpha_2 = \alpha_3$ resp.: 0.7, 0.2, −0.2, −0.7. S1–S6 are second-order models with the following values of the parameters ($\alpha_1 = \alpha_2 = \alpha_3$, $\alpha_{12} = \alpha_{13} = \alpha_{23}$) respectively: (0.2, 0.2), (0.2, 0.6), (−0.2, −0.2), (−0.2, −0.6), (0.2, −0.6), (−0.2, 0.6). T1–T6 are third-order models with values for ($\alpha_1 = \alpha_2 = \alpha_3$, $\alpha_{12} = \alpha_{13} = \alpha_{23}$, α_{123}) respectively: (0.2, 0.2, 0.2), (0.2, 0.2, 0.5), (−0.2, −0.2, −0.2), (−0.2, −0.2, −0.5), (0.2, −0.2, 0.5), (−0.2, 0.2, −0.5). G1–G4 are "general" second-order models with the following specifications for ($\alpha_1, \alpha_2, \alpha_3, \alpha_{12}, \alpha_{13}, \alpha_{23}$) respectively: (0.2, 0.3, 0.4, 0.4, 0.5, 0.6), (0.3, 0.4, 0.2, 0.5, 0.6, 0.4), (0.2, −0.3, 0.4, −0.4, 0.5, −0.6), (−0.2, 0.3, −0.4, 0.4, −0.5, 0.6).

[c]Three types of distribution of continuous variables: I = independently distributed with difference-parameter $\mu_m^{(2)} - \mu_m^{(1)}$ defined according to (4.9) in text. L = linear likelihood ratio with difference parameter defined according to (4.12) in text. N = non-linear likelihood ratio with difference parameter defined according to (4.13) in text. For further explanation see Section 4.4.

4.3 Interaction structures of the binary variables

A cell (or state) m ($m=1,2,\ldots,8$) may be represented by the outcome (x_1,x_2,x_3) of the binary variables ($x_j=0$ or 1; $j=1,2,3$) in the following order: (0,0,0), (1,0,0), (0,1,0), etc. Then, the 8 state-probabilities in population Π_i $(p_{i1},p_{i2},\ldots,p_{i8})$ can be written as 8 cell-probabilities $p_i(x_1,x_2,x_3)$. Omitting for the moment the population index i, the following loglinear model is used for representing the interaction structure between the binary variables:

$$\log p(x_1,x_2,x_3)=\alpha+\sum_{j=1}^{3}(-1)^{x_j}\alpha_j+\sum_{j=1}^{2}\sum_{k=j+1}^{3}(-1)^{x_j+x_k}\alpha_{jk}$$

$$+(-1)^{x_1+x_2+x_3}\alpha_{123} \tag{4.1}$$

with α the overall mean, α_j the main effect of variable x_j, α_{jk} the two-factor effect (first-order interaction) between variables x_j and x_k, and α_{123} the three-factor effect (second-order interaction) between x_1, x_2 and x_3 (see e.g. Bishop *et al.*, 1976).

In order to reduce the number of parameters, in most situations considered the three binary variables are completely interchangeable. That is, both the main terms α_i and the two-factor terms α_{ij} are chosen equal:

$$\alpha_1=\alpha_2=\alpha_3, \tag{4.2}$$

and

$$\alpha_{12}=\alpha_{13}=\alpha_{23}. \tag{4.3}$$

The overall term α follows from the condition

$$\sum_x p(x_1,x_2,x_3)=1. \tag{4.4}$$

Several first-order models (F) with (4.2) specified, second-order models (S) with (4.2) and (4.3) specified, and third-order models (T) with (4.2), (4.3) and α_{123} specified were considered. Also some "general" second-order models (G) for which (4.2) and (4.3) do not hold, are simulated. See footnote b of Table I for an overview of the parameter sets chosen in the present study.

4.4 Correlation structures between the continuous and the binary variables

In the location model (see Section 4.1) the three continuous variables (X_4, X_5, X_6) are assumed to follow a three-dimensional normal distribution $N(\mu_m^{(i)}, \Sigma_m^{(i)})$ for each state (or cell) m in population Π_i $(m=1,2,\ldots,8;\ i=1,2)$. The marginal distribution of (X_4, X_5, X_6) is then mixed normal. We will consider appropriate choices for the parameters $\mu_m^{(i)}$ and $\Sigma_m^{(i)}$. Firstly, we take for all situations under study $\Sigma_m^{(i)}=I$ $(i=1,2;\ m=1,2,\ldots,8)$. This is not necessarily a strong limitation: a suitable transformation of (X_4, X_5, X_6) can be found such that the new variables are conditionally independent with unit variance (Krzanowski, 1977). The (unconditional) correlation structure between X_4, X_5 and X_6 and between the continuous and the binary variables is governed by the choice for $\mu_m^{(i)}$. Without loss of generality, we take:

$$\mu_m^{(i)} = \mu_m^{(i)} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{4.8}$$

The choice for the 16 parameters $\mu_m^{(i)}$ remains to be made. In a number of situations it is assumed that all variables are independently distributed (labelled with I in Table I). In order to study the effect of distance between the two populations (see also Section 4.5) in those situations, we choose then:

$$\mu_m^{(1)}=0,\ \mu_m^{(2)}=\mu,\ m=1,2,\ldots,8,\ \mu=0,0.5,1,1.5. \tag{4.9}$$

See also Table I, footnote c.

For other situations, where the binary and/or the continuous variables are not independently distributed, we choose for $\mu_m^{(1)}$, according to the ideas of Krzanowski (1975), the expectation of the mth order statistic from a sample of size $2^3=8$ taken from a standard normal distribution. These 8 expectations are (see Pearson and Hartley, 1972, Table 9):

$$\pm 1.42360,\ \pm 0.85222,\ \pm 0.47282,\ \pm 0.15251. \tag{4.10}$$

The means will thus differ between cells, so the continuous variables will be correlated with the binary ones. In order to keep the generalized distance (see Section 4.5) as constant as possible, we will choose $\mu_m^{(2)}$ such that always:

$$\mu_m^{(2)} = \mu_m^{(1)} \pm 1, \quad m = 1, 2, \ldots, 8. \tag{4.11}$$

Two different situations will be distinguished. First we take:

$$\mu_m^{(2)} = \mu_m^{(1)} + 1, \quad m = 1, 2, \ldots, 8. \tag{4.12}$$

In the second situation we take:

$$\mu_m^{(2)} = \mu_m^{(1)} + 1, \quad m = 1, 3, 5 \text{ or } 7,$$

$$\mu_m^{(2)} = \mu_m^{(1)} - 1, \quad \text{if } m = 2, 4, 6 \text{ or } 8. \tag{4.13}$$

It is easily seen that (4.12) and (4.13) correspond with linearity and non-linearity, respectively, of the log likelihood ratio LLR with respect to its "binary" part LLR1 and its "continuous" part LLR2. Generally, $\mathrm{LLR}(X_1, X_2, \ldots, X_6) = \mathrm{LLR1}(X_1, X_2, X_3) + \mathrm{LLR2}(X_4, X_5, X_6 | X_1, X_2, X_3)$. For the linear situation (labelled with L in Table I) (4.12) holds, so $\mathrm{LLR2}(X_4, X_5, X_6 | X_1, X_2, X_3) = \mathrm{LLR2}(X_4, X_5, X_6)$ and a linear relationship for LLR follows. For the non-linear situation (labelled with N in Table I) (4.13) holds, so LLR2 depends on the binary variables.

4.5 Distance between the two populations

A final characteristic of the data from the location model is the distance between the two populations Π_1 and Π_2. We will use the general distance measure for mixed distributions as developed by Krzanowski (1983).

Consider the general location model for two populations as described in Section 4.1. We have three binary variables X_1, X_2, X_3 which define the 8 possible states in each population. Again, p_{im} is the probability of observing state m in population Π_i ($i = 1, 2$; $m = 1, 2, \ldots, 8$). Within each state, the three continuous variables X_4, X_5 and X_6 are normally distributed, denoted by $N(\boldsymbol{\mu}_m^{(i)}, \Sigma_m^{(i)})$, with

$\Sigma_m^{(i)} = I$. The distance measure of choice is the measure based on affinity ρ, defined by Matusita (1956):

$$d = (2 - 2\rho)^{1/2}. \tag{4.14}$$

The affinity is given by

$$\rho = \int \{p_1(x) p_2(x)\}^{1/2} dx \tag{4.15}$$

with $p_i(x)$ the probability density of $X = (X_1, X_2, X_3, X_4, X_5, X_6)$ in population Π_i. Krzanowski (1983) derived that with $\Sigma_m^{(i)} = I$ for (4.15) can be written:

$$\rho = \sum_{m=1}^{8} \{(p_{1m} p_{2m})^{1/2} \cdot \exp(-\tfrac{1}{8}\Delta_m^2)\} \tag{4.16}$$

with

$$\Delta_m^2 = (\mu_m^{(1)} - \mu_m^{(2)})^T (\mu_m^{(1)} - \mu_m^{(2)}). \tag{4.17}$$

Using assumption (4.8), formula (4.17) becomes:

$$\Delta_m^2 = 3(\mu_m^{(1)} - \mu_m^{(2)})^2. \tag{4.18}$$

The state-probabilities p_{im} follow from the loglinear model (4.1) in each population. Then, after specifying the loglinear parameters α_i, α_{ij}, α_{123} in each population, and the parameters $\mu_m^{(i)}$ $(i = 1, 2)$, the distance d can be calculated by using (4.1), (4.18), (4.16) and (4.14).

For all situations studied, the distance d was calculated and is shown in the last column of Table I.

5. RESULTS

The five discrimination methods have been compared for a number of situations. A situation is defined by a combination of sample size, interaction structure of the binary variables and correlation structure between the continuous and the binary variables.

The following situations are distinguished:

a) Situations to study the performance of the discriminant analysis methods for *independently* distributed variables. Especially for these

situations the effect of *distance* between the two populations is studied (Table I, simulations 1–16).

b) Situations to study the effect of *equal interaction structures* between the binary variables in both populations, in combination with *linearity of the log likelihood* with respect to the "continuous" and "binary" part of this likelihood. The effect of sample size is studied in this group of situations, together with the group of situations in c) (Table I, simulations 17–28).

c) Situations to study the effect of *unequal interaction structures* between the binary variables in both populations, in combination with the *non-linearity of the log likelihood* with respect to its "continuous" and "binary" parts (Table I, simulations 29–44).

d) Situations in which some *specific interaction structures* were studied, such as asymmetrical loglinear parameters, and the combinations "equal interaction structures/non-linearity" and "unequal interaction structures/linearity" (Table I, simulations 45–59).

Moreover, some comparisons are made to study the effect of the value of the kernel smoothness parameters.

In each run of each situation (the number of runs depends on the sample sizes) the five discriminant analysis methods were applied to the training samples. The performance was estimated by calculating the posterior probabilities for the elements of the test samples, independently generated in each run. Means of the performance measures are calculated from the runs for each situation.

5.1 Independently distributed variables

Table I, simulations 1–16, presents the situations which we considered for studying the relative performance of the discriminant analysis methods for independently distributed variables. Also, distance between the populations is varied. The simulation results for these 16 situations are presented in Appendix I. As expected, with increasing distance d, performance improves for all methods. The relative performance of the five methods is about equal for these 16 situations. This justifies a summary taking the mean value of ERR and of MLS for the 16 situations: see Table II, first row (ERR1 and MLS1). The differences between the five methods are very small.

TABLE II

Summary of the simulation results[a]

	Error rate						Modified logarithmic score				
	LDA	LOG	QDA	IND	KER		LDA	LOG	QDA	IND	KER
$\overline{\text{ERR}}1$	0.20	0.20	0.21	0.21	0.23	$\overline{\text{MLS}}1$	0.09	0.09	0.10	0.09	0.10
$\overline{\text{ERR}}2$	0.20	0.20	0.23	0.26	0.24	$\overline{\text{MLS}}2$	0.10	0.10	0.13	0.12	0.12
$\overline{\text{ERR}}3$	0.32	0.32	0.16	0.30	0.20	$\overline{\text{MLS}}3$	0.13	0.14	0.09	0.13	0.10
ERRDIF	0.04	0.04	0.07	0.06	0.07	MLSDIF	0.02	0.02	0.06	0.02	0.03
ERROPT	0.26	0.26	0.08	0.28	0.11	MLSOPT	0.09	0.08	0.08	0.08	0.04

[a] $\overline{\text{ERR}}1$ and $\overline{\text{MLS}}1$ are the mean values of ERR and MLS for sumulations 1–16 (independently distributed variables). $\overline{\text{ERR}}2$ and $\overline{\text{MLS}}2$ are the mean values of ERR and MLS for simulations 17–28 (linear likelihood ratio). $\overline{\text{ERR}}3$ and $\overline{\text{MLS}}3$ are the mean values of ERR and MLS for simulations 29–44 (non-linear likelihood ratio). ERRDIF and MLSDIF are the differences of ERR and MLS between sample size situations (25, 25) and (500, 500) after averaging over simulations 17–44. ERROPT and MLSOPT are the smallest deviations of ERR and MLS from the optimum method for simulations 17–59.

ERR1 for the kernel method KER is 2–3% higher than for the other methods; the differences for MLS1 are even smaller.

5.2 Equal interaction structures and linear log likelihood ratios

Situations with equal interaction structure and linear log likelihood ratios with respect to their continuous and binary parts are summarized in Table I, simulations 17–28. Several sample sizes are considered. See Appendix I for the simulation results for these 12 situations; the mean scores for ERR and MLS are presented in Table II, second row (ERR2 and MLS2). The differences between the methods are somewhat larger than for independent variables. The linear methods LDA and LOG perform practically equal and better than the other methods.

5.3 Unequal interaction structures and non-linear log likelihood ratios

Four situations of unequal interaction structures and non-linear log likelihood ratios (with respect to the continuous and binary parts) times four sample sizes result in the 16 situations in Table I, simulations 29–44. The results are given again in Appendix I, and the overall performance, obtained by averaging over all the 16 situations, in Table II, third row (ERR3 and MLS3). It can be seen that QDA is the best method for these situations, followed by KER. The linear methods LDA, LOG and IND run far behind.

5.4 The effect of training sample sizes

In order to study the effect of training sample sizes on discriminatory ability, the results from simulations 17–44 are used. Because almost always the results for sizes 500 are the best and for sizes 25 the worst, the differences in results for sizes 500 and 25 have been calculated, and averaged over the seven situations (see ERRDIF and MLSDIF in Table II). The difference ERRDIF is largest for QDA and KER. The difference MLSDIF is largest for QDA, followed by KER.

5.5 Some specific interaction structures

Next, some situations with specific interaction structures and with other combinations "interaction structure/linearity of log likelihood ratio" are studied.

a) The sign of the loglinear parameters $\alpha_1=\alpha_2=\alpha_3$, $\alpha_{12}=\alpha_{13}=\alpha_{23}$ and α_{123} is always the same in each of the populations Π_1 and Π_2. Some other cases with mixed signs are considered in situations 45–48 (Table I).

b) Always, the first-order and second-order interaction terms in the loglinear model were chosen equal. Some "unequal situations" are covered by situations 49 to 52 (Table I).

c) When the interaction structures in both populations are equal, we always chose $\mu_m^{(2)}=\mu_m^{(1)}+1$ $(m=1,2,\ldots,8)$. Some more "non-linearities" are covered by situations 53–55 (Table I), where for equal interaction structures (4.13) was taken for the μ-parameters.

d) When the interaction structures in both populations are unequal, we will choose $\mu_m^{(2)}$ as in (4.12), so the log likelihood ratio is linear with respect to its continuous and binary parts: see situations 56–59 in Table I.

The results for these situations are also shown in Appendix I. Situations 45–52 concern some specific values for the loglinear parameters. In situations 45, 47, 49 and 50 the interaction structures are equal and "non-linearities" are not present. Thus, these results are comparable with situations 18 and 19. Indeed, LDA and LOG perform approximately equal and better than QDA and KER.

Likewise, situations 46, 48, 51 and 52 cover unequal interaction structures and presence of "non-linearities", so these results are comparable with those in situations 30 and 32: QDA and KER possess a much better discriminatory ability than LDA and LOG. The performance of the linear methods LDA, LOG and IND for situations 46 and 48 may even be called dramatically.

Situations 53–55 correspond with situations 17, 18 and 19. However, now "non-linearities" have been introduced. Whereas without "non-linearities" LDA and LOG perform the best, when "non-linearities" are added and the interaction structures remain the

same, KER, but especially QDA, has a superior performance compared with LDA and LOG.

The last situations 56–59 concern the opposite interaction structures as in situations 29, 30, 31 and 32, however without the "non-linearities" as in these last situations. No substantial differences can be detected here.

5.6 Smallest deviation from the optimum method

A question which can be answered utilizing the results from the foregoing sections is the question which method has the smallest deviation from the optimum method. Therefore, the measures ERROPT and MLSOPT are used. ERROPT is defined by:

$$\text{ERROPT} = \max_{\substack{\text{(over} \\ \text{situations)}}} \{\text{ERR} - \min_{\substack{\text{(over} \\ \text{methods,} \\ \text{for each} \\ \text{situation)}}} \text{ERR}\} \tag{5.1}$$

and analogously MLSOPT.

The situations considered for calculations of these measures are numbers 17–59. The values for ERROPT and MLSOPT, obtained in this manner, are shown in Table II.

From these values we may conclude that QDA (for ERR) and KER (both for MLS and for ERR) have the smallest deviation from the optimum method.

5.7 Combination of LDA and QDA

Because the combination of methods LDA and QDA appears to show a very good performance, it is interesting to compare the values of ERR (and MLS) for KER with the smallest value of ERR (and MLS) of the two ones for LDA and QDA. Say, $\text{ERRCOM} = \min\{\text{ERR}_{\text{LDA}}, \text{ERR}_{\text{QDA}}\}$ and $\text{MLSCOM} = \min\{\text{MLS}_{\text{LDA}}, \text{MLS}_{\text{QDA}}\}$. Then, for practically all situations 17–59 the values of ERRCOM (and MLSCOM) are substantially lower than ERR_{KER} (and MLS_{KER}). Averaged over these situations we obtained the values ERRCOM = 0.19, $\text{ERR}_{\text{KER}} = 0.23$, MLSCOM = 0.10 and $\text{MLS}_{\text{KER}} = 0.11$.

5.8 Kernel smoothness parameters

In Section 2 it was described in which way the smoothness parameters u_1 and u_2 are estimated in our kernel program. For seven situations (17–19 and 29–32) we applied the kernel model three times:

a) in a standard manner, so u_i is estimated as indicated in Section 2.

b) by doubling the value of u_i found in a).

c) by halving the value of u_i found in a).

In Appendix II the results are shown. No substantial differences occur in the performance of KER, when doubling or halving the smoothness parameters.

5.9 Summary of results

The results as described in this section may be summarized as follows:

1) If all the variables are independently distributed, the linear methods LDA, LOG and IND have approximately the same performance, somewhat better than QDA and KER.

2) For equal interaction structures and absence of non-linearities in the log likelihood ratio LDA and LOG have the best performance. For equal or unequal interaction structures and non-linearities in the log likelihood ratio with respect to its continuous and binary part, QDA generally is superior to the other methods, followed by KER. For unequal interaction structures and absence of these non-linearities, the five methods have approximately the same discriminatory ability. It may be concluded that the impact on performance of the interaction structure of the binary variables is less than that of the presence of non-linearity induced by the conditional distribution of the continuous variables.

3) The "effect" of sample size on the performance of the methods studied is largest for QDA and KER.

4) The smallest deviation from the optimum method have QDA and KER.

5) Choosing from LDA and QDA the best performing method results in superior results.

6) In the kernel method as used in the present study, the size of the smoothness parameters has relatively little influence on the performance.

6. DISCUSSION

A really "comprehensive" study of the performance of discriminant analysis methods for mixed data seems to be hardly possible. Many limitations are necessary. Nevertheless, in this study some important features for the mixed dataset (sample size, interaction structure, existence of non-linearities in the log likelihood ratio with respect to their continuous and binary parts) have been taken into account.

The variables have a structure which may be recognized in a particular dataset with not too many variables (especially with not too many binary variables). This may be reached by fitting an appropriate loglinear model to the binary variables (to detect the interaction structure) and next, by calculating the means of the continuous variables within the cells, which can be formed by the binary variables (to detect possible presence of non-linearities).

Our results are partly a confirmation, partly an addition to earlier results from mixed data studies (Titterington *et al.*, 1981; Knoke, 1982; Vlachonikolis and Marriott, 1982; Schmitz *et al.*, 1983a; Schmitz *et al.*, 1983b).

Titterington *et al.* (1981) discuss the application of several methods of discriminant analysis to a prognosis problem involving a series of 1000 patients with severe head injury. Among their conclusions are the robustness of LDA, the lack of success of KER (from their several KER-versions we used the best one), and the comparability of the results from LOG and LDA. QDA performed somewhere between LDA and KER. Indeed, for all situations studied by us, LOG and LDA show only marginal differences in discriminatory performance. We found that in situations where KER is better than LDA and LOG (opposite interaction structures and presence of non-linearities), QDA generally performs better than KER.

Knoke (1982) compares several approaches, among which LDA, QDA, LOG and an "augmented LDA-approach", i.e. a LDA augmented with appropriate higher-order terms (interaction terms; squared variables). The methods are applied among others on a

three-dimensional problem consisting of two mutually independent binary variables and one normally distributed continuous variable with different means within each cell of the two 2×2-tables defined by the binary variables. Important conclusions are: (a) If interactions are not present, LDA is a remarkably robust method with a good performance. (b) If first-order or second-order interactions are present, QDA is to be preferred. Especially if second-order interactions are present (comparable with the situations in our study, indicated as "presence of non-linearities"), QDA is better than LDA, but performs less than the "augmented LDA-approach". Therefore, Knoke does not recommend the use of QDA as an omnibus method. The augmented LDA-approach (possibly slightly improved by using the LOG-model corresponding with the selected augmented-LDA) is recommended by Knoke if interactions exist.

Vlachonikolis and Marriott (1982) apply LDA, LOG, some KER-methods and various modifications of LDA on two sets of data with binary and continuous variables. The modified LDA has the advantage of compensating for interactions due to the binary variables. They mention as the main conclusion that "LDA is flexible and efficient, and where interaction is present, the inclusion of appropriate terms can remove its most serious disadvantage". LOG performs marginally better than LDA. The mixed kernel methods were less effective than that parametric methods.

Schmitz *et al.* (1983a) study the performance of LDA, LOG, QDA and KER on a myocardial infarction dataset, consisting of three binary and nine continuous variables. LDA, LOG and KER performed nearly identically, and better than QDA. The authors ascribe the poor performance of QDA to the absence of any substantial interaction in the dataset.

Finally, Schmitz *et al.* (1983b) compare LDA, LOG, QDA and KER in a mixed data simulation study. The data were generated from two four-dimensional normal distributions. Three of the four variables were discretized. They found that for unequal covariance structures KER performs better than or equal to QDA, and QDA better than LDA, whereas for equal covariance structures LDA performs better than KER, and again KER is better than or equal to QDA. That QDA is never the best method in that study may be explained by the absence of substantial non-linearities in the generated data.

It is not very likely that there is an overall best method for discriminant analysis. Each method will be the best under certain circumstances, and the problem is to know what they are. In this study, using data with three continuous and three binary variables, it appeared that only in situations that these variables are independently distributed, IND does have a reasonable discriminatory ability. Therefore, this method will generally not be the method of choice for this type of data. Nevertheless, it should be remarked that for high-dimensional data the estimation of the parameters of models such as LDA and LOG may give trouble, so that IND will be preferred. Also, if considerable missing data are present, it has been shown that independence models are very robust (Titterington *et al.*, 1981).

It will depend on the interaction and correlation structure of the mixed dataset whether (LDA, LOG) or (QDA, KER) perform better. As mentioned by Schmitz *et al.* (1983b), if a choice has to be made between the equally performing methods LDA and LOG, use of LDA instead of LOG may have a certain preference, because LOG fails to produce a classification rule in some situations (indicated as "zero marginal proportions"). On the other hand, in a recent study LOG appeared to have a much better reliability (goodness-of-fit with respect to the posterior probabilities) than LDA (Schmitz *et al.*, 1985).

A choice from QDA and KER will also be difficult. The data used by Schmitz *et al.* (1983b) give indications in the direction of a preference for KER, while from the present study a preference for QDA may be deduced.

Remarkable good results are reached, in this study, by using the combination of LDA and QDA (that is: use both methods, and take the best performing one). Combined with the results from Knoke (1982) and Vlachonikolis and Marriott (1982), it may be concluded that the "augmented LDA-approach" generally will be a very good choice.

A final remark has to be made. Nowadays, due to the presence of highspeed computers, it is not necessary to make an *a priori* choice from discriminant analysis methods. If these methods have been brought together in a "master" programme, the user can apply all the five (or more) methods studied in this paper, and next a choice can be made. Provided an unbiased estimation of the performance

measures (see also Schmitz *et al.*, 1983a), this approach may be recommended.

Acknowledgements

The authors would like to thank Dr. W. J. Krzanowski, Dr. J. A. Anderson, Dr. J. Hilden and Dr. H. J. Trampisch for their valuable comments on a first proposal for the design of this simulation study. The detailed comment of the referee has greatly helped improve the presentation of the results.

References

Aitchison, J., Habbema, J. D. F. and Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Applied Statistics* **26**, 15–25.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Anderson, T. W. (1958). *Introduction to multivariate analysis.* John Wiley, New York.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1976). *Discrete multivariate analysis: Theory and practice.* MIT Press, Cambridge.

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

Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In: Compstat 1974, Proceedings in computational statistics. Physica Verlag, Wien, 101–110.

Habbema, J. D. F., Hermans, J. and Remme, J. (1978a). Variable kernel density estimation in discriminant analysis. Compstat Proceedings. Physica Verlag, Wien, 178–185.

Habbema, J. D. F. and Hilden, J. (1981). The measurement of performance in probabilistic diagnosis IV. Utility considerations in therapeutics and prognostics. *Methods of Information in Medicine* **20**, 80–96.

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. *Methods of Information in Medicine* **17**, 217–226.

Habbema, J. D. F., Hilden, J. and Bjerregaard, B. (1981). The measurement of performance in probabilistic diagnosis V. General Recommendations. *Methods of Information in Medicine* **20**, 97–100.

Hand, D. J. (1982). *Kernel discriminant analysis.* John Wiley, Chichester.

Hermans, J. and Habbema, J. D. F. (1975). Comparison of five methods to estimate posterior probabilities. *EDV Medizin Biol.* **6**, 14–19.

Hermans, J., Habbema, J. D. F., Kasanmoentalib, T. K. D. and Raatgever, J. W. (1983). Manual for the ALLOC80 discriminant analysis program. Technical Report, Dept. of Medical Statistics, University of Leiden, The Netherlands.

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. *Methods of Information in Medicine* **17**, 227–237.

Hilden, J., Habbema, J. D. F. and Bjerregaard, B. (1978b). The measurement of performance in probabilistic diagnosis III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 238–246.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics* **38**, 191–200.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association* **70**, 782–790.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics* **19**, 191–200.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36**, 493–499.

Krzanowski, W. J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika* **70**, 235–243.

Lee, E. T. (1974). A computer program for linear logistic regression analysis. *Computer Programs in Biomedicine* **4**, 80–92.

Matusita, K. (1956). Decision rule, based on the distance, for the classification problem. *Ann. Inst. Statist. Math.* **8**, 67–77.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika tables for statisticians. Vol. II.* Cambridge University Press, Cambridge.

Schmitz, P. I. M., Habbema, J. D. F. and Hermans, J. (1983a). The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. *Statistics in Medicine* **2**, 199–205.

Schmitz, P. I. M., Habbema, J. D. F., Hermans, J. and Raatgever, J. W. (1983b). Comparative performance of four discriminant analysis methods for mixtures of continuous and discrete variables. *Comm. in Stat.* **B12**, no. 6, 727–751.

Schmitz, P. I. M., van de Merwe, J. P. and Habbema, J. D. F. (1985). Construction and validation of a diagnostic support model for the diagnosis of Crohn's disease by agglutination tests. (submitted).

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F. and Gelpke, G. J. (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society Series A* **144**, 145–175.

Vlachonikolis, I. G. and Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Applied Statistics* **31**, 23–31.

Appendix I

SIMULATION RESULTS

	Error rate					Modified logarithmic score				
Simulation	LDA	LOG	QDA	IND	KER	LDA	LOG	QDA	IND	KER
1	0.48	0.47	0.46	0.46	0.44	0.16	0.16	0.18	0.16	0.17
2	0.33	0.32	0.34	0.34	0.38	0.14	0.14	0.15	0.14	0.15
3	0.19	0.20	0.21	0.21	0.22	0.09	0.09	0.10	0.10	0.11
4	0.10	0.10	0.11	0.11	0.12	0.05	0.05	0.06	0.06	0.06
5	0.31	0.31	0.30	0.30	0.31	0.13	0.12	0.13	0.12	0.13
6	0.23	0.23	0.27	0.23	0.28	0.10	0.10	0.12	0.10	0.12
7	0.14	0.13	0.13	0.14	0.15	0.07	0.07	0.07	0.07	0.08
8	0.07	0.07	0.07	0.08	0.08	0.04	0.04	0.04	0.04	0.05
9	0.18	0.18	0.18	0.18	0.19	0.10	0.10	0.10	0.09	0.10
10	0.18	0.18	0.17	0.18	0.18	0.10	0.09	0.10	0.09	0.10
11	0.11	0.11	0.12	0.12	0.13	0.06	0.06	0.07	0.06	0.07
12	0.07	0.08	0.08	0.07	0.11	0.04	0.05	0.05	0.04	0.06
13	0.35	0.35	0.39	0.37	0.42	0.15	0.15	0.16	0.15	0.16
14	0.28	0.28	0.30	0.30	0.32	0.13	0.13	0.13	0.13	0.14
15	0.18	0.18	0.19	0.20	0.20	0.09	0.09	0.09	0.09	0.10
16	0.08	0.08	0.09	0.10	0.10	0.04	0.05	0.05	0.06	0.06
17	0.18	0.18	0.19	0.20	0.20	0.09	0.09	0.09	0.09	0.10
18	0.20	0.21	0.26	0.28	0.26	0.10	0.10	0.14	0.13	0.13
19	0.20	0.21	0.21	0.28	0.26	0.10	0.10	0.10	0.12	0.13
20	0.21	0.22	0.24	0.22	0.25	0.11	0.12	0.14	0.11	0.11
21	0.24	0.23	0.31	0.33	0.29	0.12	0.12	0.20	0.14	0.14
22	0.22	0.22	0.28	0.32	0.31	0.11	0.11	0.17	0.15	0.14
23	0.16	0.16	0.17	0.19	0.18	0.08	0.08	0.09	0.09	0.09
24	0.18	0.18	0.17	0.24	0.19	0.08	0.08	0.09	0.11	0.10
25	0.20	0.20	0.20	0.29	0.20	0.09	0.09	0.09	0.12	0.10
26	0.19	0.19	0.23	0.22	0.22	0.09	0.09	0.11	0.10	0.11
27	0.22	0.22	0.26	0.28	0.25	0.10	0.11	0.14	0.12	0.13
28	0.22	0.22	0.28	0.31	0.28	0.10	0.11	0.15	0.14	0.14
29	0.32	0.32	0.16	0.33	0.21	0.14	0.14	0.08	0.13	0.10
30	0.22	0.23	0.15	0.23	0.16	0.12	0.12	0.10	0.11	0.09
31	0.31	0.31	0.15	0.33	0.22	0.13	0.13	0.08	0.14	0.11
32	0.32	0.32	0.13	0.28	0.20	0.13	0.13	0.07	0.12	0.10

Appendix I (continued)

	Error rate					Modified logarithmic score				
Simulation	LDA	LOG	QDA	IND	KER	LDA	LOG	QDA	IND	KER
33	0.35	0.35	0.22	0.34	0.26	0.15	0.15	0.12	0.15	0.12
34	0.29	0.30	0.17	0.28	0.18	0.13	0.13	0.14	0.12	0.10
35	0.38	0.38	0.22	0.36	0.26	0.15	0.15	0.12	0.16	0.13
36	0.33	0.33	0.17	0.33	0.23	0.14	0.14	0.10	0.14	0.11
37	0.28	0.28	0.14	0.30	0.15	0.13	0.13	0.07	0.13	0.08
38	0.30	0.30	0.17	0.19	0.16	0.11	0.12	0.09	0.10	0.07
39	0.32	0.32	0.13	0.29	0.18	0.13	0.13	0.07	0.12	0.09
40	0.30	0.30	0.11	0.29	0.22	0.13	0.13	0.06	0.13	0.10
41	0.36	0.37	0.20	0.35	0.21	0.14	0.14	0.11	0.14	0.11
42	0.28	0.28	0.16	0.23	0.15	0.12	0.12	0.12	0.11	0.08
43	0.35	0.35	0.18	0.32	0.21	0.14	0.14	0.10	0.13	0.10
44	0.35	0.35	0.14	0.33	0.20	0.16	0.15	0.07	0.14	0.10
45	0.19	0.20	0.21	0.27	0.24	0.10	0.10	0.10	0.12	0.11
46	0.50	0.50	0.36	0.56	0.35	0.20	0.19	0.23	0.22	0.19
47	0.22	0.22	0.22	0.32	0.27	0.10	0.10	0.11	0.15	0.12
48	0.48	0.48	0.24	0.43	0.31	0.17	0.17	0.12	0.17	0.14
49	0.20	0.20	0.22	0.24	0.22	0.09	0.10	0.10	0.11	0.11
50	0.19	0.20	0.25	0.26	0.25	0.09	0.10	0.13	0.12	0.13
51	0.07	0.09	0.06	0.12	0.07	0.06	0.07	0.04	0.08	0.04
52	0.39	0.39	0.25	0.40	0.27	0.16	0.16	0.16	0.17	0.19
53	0.34	0.34	0.18	0.37	0.22	0.14	0.14	0.09	0.15	0.10
54	0.37	0.37	0.24	0.44	0.28	0.17	0.17	0.13	0.18	0.14
55	0.47	0.47	0.21	0.49	0.27	0.17	0.17	0.10	0.18	0.12
56	0.18	0.19	0.18	0.18	0.21	0.09	0.09	0.09	0.09	0.10
57	0.16	0.17	0.14	0.11	0.15	0.08	0.08	0.09	0.06	0.09
58	0.16	0.17	0.17	0.16	0.21	0.08	0.08	0.08	0.08	0.10
59	0.16	0.16	0.15	0.16	0.22	0.08	0.08	0.08	0.08	0.10

Appendix II

SIMULATION RESULTS FOR THE KERNEL METHOD WITH DIFFERENT CHOICES FOR THE SMOOTHNESS PARAMETERS

	Error rate Estimation of u_i			Modified logarithmic score Estimation of u_i		
Simulation	standard	doubled	halved	standard	doubled	halved
17	0.20	0.23	0.18	0.10	0.12	0.12
18	0.26	0.26	0.26	0.13	0.14	0.12
19	0.21	0.21	0.21	0.10	0.10	0.10
29	0.16	0.16	0.15	0.09	0.10	0.07
30	0.26	0.26	0.24	0.13	0.14	0.12
31	0.22	0.24	0.19	0.11	0.12	0.10
32	0.20	0.20	0.21	0.10	0.11	0.09

CHAPTER 7

LOGISTIC DISCRIMINANT ANALYSIS FOR MODELLING QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS (QSAR)

Logistic Discriminant Analysis for Modelling Quantitative Structure-Activity Relationships. P. I. M. Schmitz and J. D. F. Habbema. Submitted for publication, February 1986.

LOGISTIC DISCRIMINANT ANALYSIS FOR MODELLING QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS (QSAR)

P.I.M.Schmitz
Institute of Biostatistics
Erasmus University
P.O.Box 1738
3000 DR Rotterdam
The Netherlands

J.D.F.Habbema
Department of Public Health
and Social Medicine
Erasmus University
Rotterdam

SUMMARY

In QSAR-studies with a discrete (often binary) biological activity variable, discriminant analysis (supervised pattern recognition) can be a valuable method. Logistic discriminant analysis is compared with Fisher's linear discriminant analysis, quadratic discriminant analysis, a kernel method and an independence model, both by reviewing some results and by application of the methods to a QSAR dataset from the pharmaceutical industry. The conditions for applicability of the standard logistic discrimination model and some of its generalizations are discussed.

1. INTRODUCTION

Quantitative structure-activity relationships (QSAR) relate one or more biological activity variables (y) in a series of chemical compounds to a set of characteristics (x) of the compounds, often including information about their structure. With QSAR models it may be possible to detect the characteristics of the chemical compounds which are prognostic for the biological activity. The biological activity generally depends on more than one characteristic (or structural indicator) of the compound. Therefore, multivariate methods are used in QSAR. For a review of these methods see Wold and Dunn (1). The classical approach is multiple regression analysis, which relates a number of variables $x_1, x_2, \ldots, x_p$ to a continuous, normally distributed, biological activity variable y. However, if activity is a qualitative (ordered or unordered) variable, multiple linear regression is less appropriate. For such situations, discriminant analysis (DA), one of the techniques falling under the umbrella term 'pattern recognition', may be applied. This paper presents an evaluation of some methods of discriminant analysis, in particular logistic discriminant analysis, by their application to results from chemical and biological investiagtions in a pharmaceutical firm.

First, logistic regression analysis, which may also be formulated as a DA method (logistic discriminant analysis), is discussed (section 2). Then, four other discriminant analysis methods used in our comparison are introduced (section 3). After defining the evaluation measures, and reviewing some earlier results, the five discriminant analysis methods (logistic discrimination, linear discriminant analysis, quadratic discriminant analysis, a kernel method and an independence model) are applied to a QSAR dataset (section 4). Some extensions of the standard model are considered in section 5, and the results of the comparisons are discussed in section 6.

2. LOGISTIC REGRESSION ANALYSIS

The classical statistical technique used in QSAR studies is multiple linear regression. The basic model is that the activity y_i of a compound is linearly related to a set of p physicochemical features x_{ij}:

$$y_i = \beta_o + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i , \qquad i = 1,2,\ldots,n \tag{2.1}$$

n is the number of compounds, ε_i is a random (or error) term, and β_j are coefficients which can be estimated from the data. Usual (but not necessary) assumptions are that the independent variables x_{ij} are measured without error, and that the error terms ε_i are independent and follow identical normal distributions with expectation zero.

Applications of multiple regression in QSAR studies include Hansch (2) and Free and Wilson (3). Increasingly, multiple regression is used for selecting of significant predictive variables. This has become a common practice after regression software packages such as BMDP, SPSS, etc. have become widely available.

It frequently occurs that the dependent variable in a QSAR-study, the biological activity y, is expressed in a dichotomous way: 'active' (y=1) or 'not active' (y=0). Use of standard multiple regression in those situations is possible, especially when only relative small changes in 'expected activity' (for binary data this is equivalent with the probability of activity P) are involved. For larger changes this approach may lead to interpretation problems because of the condition $0 \leq E(y)=P \leq 1$. A solution might be to consider estimators $\hat{\beta}_j$ of the multiple regression equation $E(y_i)= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ subject to the constraints $0 \leq E(y_i) \leq 1$. However, apart from computational difficulties, there are still interpretation problems for the parameters in this model.

The most appropriate alternative is the logistic regression model. This model assumes that the expectation of the 0-1 Bernoulli random variable y is a logistic function of a linear combination of the explanatory variables x_i:

$$P_i = E(y_i) = \frac{\exp \{\beta_o + \sum_{j=1}^{p} \beta_j x_{ij}\}}{1+\exp \{\beta_o + \sum_{j=1}^{p} \beta_j x_{ij}\}} \qquad (2.2)$$

or:

$$\log \left(\frac{P_i}{1-P_i}\right) = \beta_o + \sum_{j=1}^{p} \beta_j x_{ij} \qquad (2.3)$$

The unknown coefficients β_j are usually estimated by maximum likelihood (ML) (4). With ML estimates $\hat{\beta}_j$ inserted in expression (2.2), an estimate $\hat{P}_i$ for P_i is obtained. $\hat{P}_i$ is interpreted as the estimated probabiltiy that compound i with chemical characteristics $x_{i1}, x_{i2}, \ldots, x_{ip}$ will be biological active.

BMDP (5) and SPSS-X (6) contain subroutines for (stepwise) logistic regression analysis. Adding (or deleting) variables to (from) the regression equation in these subroutines is based on the increase in the likelihood. Since a good prediction of activity is of primary interest, we will prefer measures other than the likelihood criterion for the assessment of prediction. These measures are developed in the context of another statistical technique: discriminant analysis. Therefore, wel will discuss logistic regression analysis further in the following sections as a method for discrimination and classification.

3. DISCRIMINANT ANALYSIS

3.1 General formulation of discrimination problems.

The general discriminant analysis model in QSAR may be described for k (>2) populations (groups) of compounds with response y (activity). Here we will limit ourselves to two populations II_1 (indicated by y=1: the 'active' compounds) and II_2 (indicated by y=0: the 'non-active' compounds). Suppose that one is forced to allocate to one of these two populations. If none of the characteristics x_j are known, allocation can be done according to the so-called prior probabilities $P(II_1)$ and $P(II_2)=1-P(II_1)$. These prior probabilities can sometimes be estimated from the total sample of compounds: $\hat{P}(II_1)=n_1/(n_1+n_2)$, with n_1= number of active compounds, n_2= number of non-active compounds.

The feature vector $X'=(x_1,x_2,\ldots,x_p)$ with chemical characteristics of the compound gives additional probabilistic information about the activity. This information is expressed by the conditional or posterior probabilities $P(II_1|X)$ and $P(II_2|X)=1-P(II_1|X)$. Using Bayes' theorem the posterior probability of activity, $P(y=1|X) = P(II_1|X)$, can be written as:

$$P(y=1|X) = \frac{1}{1 + \frac{P(II_2)}{P(II_1)} \cdot \frac{P(X|II_2)}{P(X|II_1)}} \qquad (3.1.1)$$

$P(X|II_k)$, k=1,2 are the probability densities of vector X in populations II_k. These densities, or at least their ratio, have to be estimated from the data, consisting of two training samples, one from each population. After estimation of the prior probabilities and the densities (ratio), the posterior probabilities

are estimated with expression (3.1.1). Contrary to discriminant analysis methods, which are based on the estimation of the densities, logistic discriminant analysis directly estimates the posterior probabilities.

For measuring the performance of an allocation rule by the error rate, each element (compound) of the two training datasets can be allocated to the population with highest posterior probability. See section 4.1 for other measures and section 4.3 for a method for bias reduction.

3.2 The discriminant analysis methods used.

The expression for the posterior probabilities in logistic discrimination (LOG) is totally identical to the regression formulation (2.2). Again, the parameters β_j are estimated by maximum likelihood from two training sample datasets. This implicitly implies the assumption of proportionality of the prior probabilities to the sizes of the training samples (n_i from population II_i, i=1,2):

$$P(II_1) = n_1/(n_1 + n_2) \tag{3.2.1}$$

For other probabilities, say $P^*(II_i)$, the ML estimates for coefficients $\beta_1, \beta_2, \ldots, \beta_p$ remain unchanged, whereas the estimate $\hat{\beta}_0$ of the constant term β_0 has to be replaced by:

$$\hat{\beta}_o^* = \hat{\beta}_o + \ln \frac{P^*(II_1)}{P^*(II_2)} - \ln \frac{n_1}{n_2} \tag{3.2.2}$$

Besides logistic discriminant analysis, four other methods for discrimination will be considered in this paper. These four methods are characterised by different underlying densities $P(X|II_i)$, i=1,2. Linear and quadratic discriminant analysis (LDA and QDA respectively) assume multinormal densities with equal or unequal covariance matrices. With estimation of the mean vectors and the covariance matrices by their usual maximum likelihood estimates, the estimated density for QDA becomes:

$$P(X|II_i) = (2\pi)^{-\frac{1}{2}p} |S_i|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(X-\bar{x}_i)^T S_i^{-1} (X-\bar{x}_i)\} \tag{3.2.3}$$

with $\bar{x}_i$ the sample mean and S_i the sample covariance matrix (7). The density $P(X|II_i)$ for LDA is obtained by replacing S_i with the pooled sample covariance matrix S. The kernel method KER uses a kernel function K for mixed data,

described by Habbema et al.(8):

$$P(X|II_i) = \frac{1}{n_i} \sum_{k=1}^{n_i} \prod_{j=1}^{p} K_{type(j)}\,(X_j;X_{ikj}) \qquad (3.2.4)$$

Kernel components $K_{type(j)}$ are defined for each type of variable j; type may be continuous, ordinal, nominal or binary.

The last method is the independence model (IND). The computer program used calls for discretizing the continuous variables, and uses a 'zero-avoiding device':

$$P(X|II_i) = \prod_{j=1}^{p} \frac{n_i(x_j)+1/(2c_j)}{n_i+\frac{1}{2}} \qquad (3.2.5)$$

where

c_j = number of catgeories of variable x_j,

$n_i(x_j)$ = number of sample elements with score x_j on variable j, and

n_i = sample size.

4. COMPARISON OF FIVE DISCRIMINANT ANALYSIS METHODS

4.1 Evaluation measures.

After estimation of coefficients β_j in the logistic discriminant function (2.2) or of densities $P(X|II_k)$ in Bayes' formula (3.1.1) for other discriminant analysis methods, the posterior probabilities $P(II_1|X)$ can be calculated for new elements (compounds) from two test- or evaluation samples. In order to assess the prediction results (active or not active) of a discriminant analysis model, the so-called forced classification (or allocation) matrix can be constructed. By forced allocation is meant that a compound is allocated to the population with the highest posterior probability. Within each population the number of compounds classified as 'active' or 'not active' are entered in this matrix. For two populations a 2x2 table is obtained.

The error rate, defined as the expected fraction of misclassifications, can easily be calculated from the allocation matrix:

$$ERR = P(II_1).P(\rightarrow II_2 | II_1) + (1-P(II_1)).P(\rightarrow II_1 | II_2) \qquad (4.1.1)$$

where $P(\rightarrow II_j | II_i)$ is the probability of allocation to II_j for elements from II_i, the misclassification probabilities. Although the error rate is a very common performance measure, it also is a rather crude and thereby unstable one: probabilities higher (respectively lower) than 0.5 are not distinguished from each other. A first step towards more sophistication is the introduction of an allocation of doubt in case neither population has a very high (or low) probability (e.g. between 0.2 and 0.8). Other performance measures are based on so-called continuous scoring rules, see for example Hilden et al.(9). These rules measure the agreement between probabilistic predictions and actual outcomes, and satisfy the requirement that the larger the probability assigned to an element's (compound's) actual population (active or not active), the smaller the penalty score. One such scoring rule, the modified logarithmic scoring rule, to a close approximation equals to $-\ln(P(k)+c)$, with P(k) the probability assigned to the population of origin and c reflecting the seriousness of assigning a very low probability to the actual populations. By averaging the scores over all $n= n_1+n_2$ test sample elements (weigthed for the prior probabilities), we obtain the average modified logarithmic score MLS:

$$MLS = (1/n) \sum_{k=1}^{n} - \ln(P(k)+c) \qquad (4.1.2)$$

The value c = 0.02 will be used in the remainder of the present paper.

4.2 Review of earlier comparison results.

Because most QSAR datasets consist of mixtures of continuous and discrete variables, we will only consider the performance of DA methods for mixtures of variables. The performance of the methods described in section 3.2 (LOG, LDA, QDA, KER and IND) has been studied before, both by applying them to empirical datasets (Titterington et al.(10); Knoke (11); Vlachonikolis and Marriott (12); Schmitz et al. (13)) and by means of simulation (Schmitz et al. (14,15)). Titterington et al. (10) describe as main results the robustness of LDA, the lack of success of KER, the comparibility of the results from LOG and LDA, and the good results of IND.

Knoke (11) concludes that if interactions are not present, LDA is a remarkable robust method with a good performance. If first- or second order interactions are present QDA is better than LDA, but performs worse than the so-called 'augmented LDA-approach': i.e. LDA augmented with appropriate higher-order terms (interaction terms; squared variables).

Comparable results are presented by Vlachonikolis and Marriott (12). They also used several mixed kernel methods, which appear to be less effective than the parametric methods. Schmitz et al. (13) study the performance of LDA, LOG, QDA and KER on a mixed dataset. LDA, LOG and KER performed nearly identically and better than QDA. The authors ascribe the poor performance of QDA to a combination of the relatively small sample sizes and the absence of substantial interactions in the datasets.

In a mixed data simulation study, see Schmitz et al. (14), QDA was never better than KER. This can probably be explained by the absence of higher-order interactions in the generated data. In a second simulation study, Schmitz et al. (15) got sometimes better results for QDA than for KER. All studies which also included IND show that this method has only a reasonable discriminatory ability in situations with (approximately) independently distributed variables.

4.3 Application to a QSAR dataset.

The methods described in section 3.2 were applied to a dataset consisting of 400 compounds. For each compound the biological activity y on a particular species has been determined and classified as 'active' (y=1) or 'not active' (y=0). The training data consisted of 206 'active' and 194 'not active' compounds. Prior probabilities were chosen equal to 50%. Each compound was characterised by 6 variables (including structure components) which were selected from a larger set using one of the preprocessing methods described in section 5. The distributions of the variables for the active and not active compounds respectively are summarized in Table 1. The correlation matrices are not presented: most correlations are approximately zero, with exception of the correlation between x_2 and x_3 (r=0.16 for active and r=0.44 for not active compounds).

The classification performance, measured by the error rate ERR and the modified logarithmic score MLS, using the training data as test data

Table 1. Distributions of structure components in a QSAR dataset.

	active compounds (n=206)		not active compounds (n=194)	
continuous	$\bar{x}$	s.d.	$\bar{x}$	s.d.
x_1=B_1 ringsystem 3 para	1.68	0.34	1.49	0.32
x_2=B_4 ringsystem 3 para	7.68	5.86	6.67	9.76
x_3=(Pi ringsystem 1 + Pi ringsystem 2 + Pi ringsystem 3)^2	15.48	8.75	13.60	12.64
binary (0-1):	% present		% present	
x_4= ortho substitution ring 1	1.0		9.8	
x_5= meta substitution ring 1	1.5		9.3	
x_6= ortho substitution ring 3	0.5		9.3	

(resubstitution method), is shown in Table 2. The results suggest superior performance for the kernel method, approximately equal discriminatory ability for LOG, LDA and IND, and a worse performance for QDA. However, these results, especially those for the kernel method, are too optimistic because assessment of the performance measures was obtained by resubstitution (see also Schmitz et al (13)). Using the so-called 'split-sample method' results in less biased estimates of ERR and MLS. The two training groups were randomly split into two parts of equal size. One part was used as training sample from which the allocation rule was deduced, the other part was used as test sample to which the allocation rule is applied. The results for the 103 'active', 97 'not active' training sample elements and 103 'active', 97 'not active' test sample elements are shown in Table 3. The picture is quite different: for the error rate, LOG and LDA have the lowest values, while QDA, KER and IND perform worse. Using MLS as measure, KER and IND show a more favourable picture than with KERR.

Compared with earlier studies about the performance of the discriminant

Table 2. Performance of five discriminant analysis methods with the resubstitution method.(*) (206 active, 194 not active compounds) (low values indicate good performance)

	LOG	LDA	QDA	KER	IND
ERR	0.323	0.318	0.373	0.208	0.295
MLS	0.135	0.135	0.264	0.094	0.129

* Abbreviations:
ERR = Error Rate, MLS = Modified Logarithmic Score,
LOG = LOGistic discriminant analysis, LDA = Linear Discriminant Analysis,
QDA = Quadratic Discriminant Analysis, KER = KERnel method,
IND = INDependence model.

Table 3. Performance of five discriminant analysis methods with the split-sample method.(*) (103 active, 97 not active compounds) (low values indicate good performance)

	LOG	LDA	QDA	KER	IND
ERR	0.320	0.325	0.395	0.410	0.410
MLS	0.157	0.158	0.286	0.173	0.161

* Abbreviations: see Table 2.

analysis methods considered here, the strong detoriation in error rate of the independence model (from 0.295 to 0.410) is remarkable. This is undoubtly partly a chance phenomenon, but it also stresses the unstability of the error rate as a performance measure (compare the much more reasonable and consistent behaviour of the MLS).

5. EXTENSIONS OF STANDARD LOGISTIC DISCRIMINATION

Discriminant analysis (or supervised pattern recognition) is only one procedure in the general statistical analysis of chemical data, as can be performed with the main chemometric programs ARTHUR (16) and SIMCA (17). Before using a discrimination model, feature selection (preprocessing) will generally be necessary. A possible approach in QSAR problems consists of three steps (Kowalski and Wold (18)):

1. Perform a cluster analysis (unsupervised pattern analysis).
2. Select one variable (or aggregrate) from each variable cluster in the subsequent analysis (Massart, (19); Jain and Dubes, (20)).
3. Supposed the reduced number of variables is thereafter smaller than one third of the total of chemical compounds ($p<(1/3)n$), further selection can be based on discriminatory ability.

Selection of variables in discriminant analysis has been extensively studied, see for example Habbema and Hermans (21) and Rencher and Larson (22). Only few programs use slection criteria based on the performance measures described in section 4.1. The selection of the six variables in our QSAR dataset (section 4.3) was performed with the INDEP-SELECT program (Habbema and Gelpke, (23)) which is built around the independence model IND (section 3.2). The six variables were selected from 25 features using the error rate as selection criterion. Actually, the choice of a selection method (including a strategy of selection, a criterion of optimality for selecting the variables, a stopping rule in the selection procedure and the discriminant analysis model) can have a considerable effect on the evaluation of standard logistic discrimination.

Some useful extensions of the standard model for discriminant analysis have been described by Habbema (24). Two of these will be discussed here in the context of logistic discrimination: extension to more than two populations (or groups), and improving performance by adding appropriate higher-order terms.

Biological activity in QSAR was considered until sofar as continuous or binary. Intermediate situations, where the activity variable y is categorical with k (>2) categories, can be analysed by the so-called polychotomous logistic model (Anderson, (25)). This model is not widely used, however, because of the large number of parameters that has to be estimated. Recently, Begg and Gray (26) showed that use of so-called individualized logistic

discrimination, in which a series of separate simple logistic models are analysed is an efficient alternative for the polychotomous logistic model. Schmitz et al.(27) present an application with three categories. Explicit use of the ordinal character of the outcome categories (for example, biological activity y is 'not or weak active', 'intermediate active' or 'strongly active') is studied by Anderson and Philips (28).

Another extension of standard logistic discrimination is to non-linear models. In section 4.2 it was already mentioned that if first- or second-order interactions are present in the data, QDA is better than LDA, but performs worse than the 'augmented LDA-approach'. This does hold also for logistic discrimination. A good strategy is to start with the calculation and evaluation of linear logistic discrimination. Next, one should aim at improvement of the model by adding higher-order terms. Although quadratic terms and three- or higher-order interactions may contribute to this improvement, generally the most contribution will come from adding two-factor interaction terms $x_i x_j$ ($i \neq j$) to the linear function (2.3). Adding all these terms at once may result in too many parameters; they should be screened in advance by calculating the correlations between the interaction terms $x_i x_j$ with the residuals from model (2.3) ((0,1)-variable minus its predicted probability). These correlations are tested for essential non-zero values, and the corresponding interaction terms are added. Alternative approaches are described by Vlachonikolis and Marriott (12) and Knoke (11).

6. DISCUSSION

For modelling quantitative structure-activity relationships with a dichotomous biological activity variable, logistic regression analysis (formulated as discriminant analysis method) is an obvious and valuable approach, especially in the situation of a mixture of continuous and binary explanatory variables. One of the main results of a comparison of the logistic model with alternative discriminant analysis models is that Fisher's linear discriminant analysis (LDA) consistently shows approximately the same discriminatory performance as logistic discriminant analysis (LOG). Actually this reflects the robustness of LDA: if one of its statistical conditions, an underlying multivariate distribution, is not fulfilled it still can show a satisfactory discriminatory or predictive performance.

If the dataset has a structure with unequal covariance matrices or higher-order interactions, other discriminant analysis methods such as quadratic analysis or a kernel model may be preferred. However, in these situations the performance of LOG and LDA is generally increased by adding higher order terms to the linear discriminant function. If the alternative discriminant analysis methods are brought together in a 'master' program, the methods could be compared for a specific dataset. An a priori choice would be necessary in that case.

Acknowledgements.

The authors would like to thank Dr. A.Vardy and some of his colleagues from Duphar, who made the dataset available and commented a first draft of this paper.

REFERENCES

(1) S.Wold and W.J.Dunn III, J.Chem.Inf.Comput.Sci.,23(1983)6.
(2) E.J.Ariens, ed., "Drug Design 1", Medicinal Chemistry Series, Vol.11, Academic Press, New York, 1971.
(3) S.M.Free and J.W.Wilson, J.Med.Chem., 7(1964)395.
(4) D.R.Cox, "The analysis of binary data", Chapman and Hall,London,1970.
(5) W.J.Dixon,ed., "BMDP, Biomedical Computer Programs", University of California Press, Berkeley, 1975.
(6) "SPSS-X User's Guide, SPSS Inc.", McGraw-Hill, New York,1983.
(7) T.W.Anderson, "Introduction to Multivariate Analysis", J.Wiley, New York,1958.
(8) J.D.F.Habbema, J.Hermans and J.Remme, "Compstat Proceedings", Physica Verlag, Wien, 1978.
(9) J.Hilden, J.D.F.Habbema and B.Bjerregaard, Meth.Inform.Med., 17(1978)238.
(10) D.M.Titterington, G.D.Murray, L.S.Murray, D.J.Spiegelhalter, A.M.Skene, J.D.F.Habbema and G.J.Gelpke, J.Royal Stat.Soc. Series A, 144(1981)145.
(11) J.D.Knoke, Biometrics, 38(1982)191.
(12) I.G.Vlachonikolis and F.H.C.Marriott, Applied Statistics,31(1982)23.

(13) P.I.M.Schmitz, J.D.F.Habbema and J.Hermans, Stat.in Med.,2(1983)199.
(14) P.I.M.Schmitz, J.D.F.Habbema, J.Hermans and J.W.Raatgever, Comm.in Stat.B, 12(1983)727.
(15) P.I.M.Schmitz, J.D.F.Habbema and J.Hermans, J.Comput.and Simul., 23(1985)69.
(16) A.M.Harper, D.L.Duewer, B.R.Kowalski and J.L.Fasching, in B.R.Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society, Washington, 1977.
(17) S.Wold and M.J.Sjöström, in B.R.Kowalski (Ed.), Chemometrics: Theory and Application, American Chemical Society,Washington,1977.
(18) B.R.Kowalski and S.Wold, in P.R.Krishnaiah and L.N.Kanal, eds., Handbook of Statistics 2, North-Holland Publ. Comp., 1982.
(19) D.L.Massart, A.Kijkstra and L.Kaufman, "Evaluation and Optimization of Laboratory Methods and Analytical Procedures", Elsevier, Amsterdam,1978.
(20) A.K.Jain and R.Dubes, Pattern Recognition, 10(1978)85.
(21) J.D.F.Habbema and J.Hermans, Technometrics, 19(1977)487.
(22) A.C.Rencher and S.F.Larson, Technometrics, 22(1980)349.
(23) J.D.F.Habbema and G.J.Gelpke, Comp.Progr.in Biomed.,13(1981)251.
(24) J.D.F.Habbema, Analytica Chimica Acta, 150(1983)1.
(25) J.A.Anderson, Biometrika, 59(1972)19.
(26) C.B.Begg and R.Gray, Biometrika, 71(1984)11.
(27) P.I.M.Schmitz, J.P.van de Merwe and J.D.F.Habbema (1985). Submitted.
(28) J.A.Anderson and P.R.Philips, Applied Statistics, 30(1981)22.

CHAPTER 8

MULTIVARIATE LOGISTIC ANALYSIS OF RISK FACTORS FOR STROKE IN TILBURG, THE NETHERLANDS

Multivariate Logistic Analysis of Risk Factors for Stroke in Tilburg, The Netherlands. Amer. J. of Epid. 118 (1983) 514 - 525.

AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 118, No. 4

Printed in U.S.A.

MULTIVARIATE LOGISTIC ANALYSIS OF RISK FACTORS FOR STROKE IN TILBURG, THE NETHERLANDS

BERTRAM HERMAN,[1] PAUL I. M. SCHMITZ,[2] ANTON C. M. LEYTEN,[1] JACOB H. VAN LUIJK,[1] CORNELUS W. G. M. FRENKEN,[1] ADOLF A. W. OP DE COUL[1] AND BENTO P. M. SCHULTE[3]

Herman, B. (Tilburg Epidemiological Study of Stroke, Tilburg, The Netherlands), P. I. M. Schmitz, A. C. M. Leyten, J. H. van Luijk, C. W. G. M. Frenken, A. A. W. Op de Coul and B. P. M. Schulte. Multivariate logistic analysis of risk factors for stroke in Tilburg, The Netherlands. *Am J Epidemiol* 1983;118: 514-25.

By means of a case-control study conducted between October 1, 1978, and July 31, 1981, in Tilburg, The Netherlands, various characteristics and events, including personal data, health-related behavior, and medical history, were evaluated as risk factors for stroke. The study subjects included 132 stroke patients and 239 age- and sex-matched control patients interviewed at the two city hospitals. To assess joint effects and possible interactions, and to control for multiple confounding factors, a series of multivariate logistic models for matched data were studied. From this analysis, it appeared that hypertension, acute myocardial infarction, cardiac arrhythmias, transient cerebral ischemic attacks, obesity, physical activity during leisure time, education of head of household, and Rhesus factor were all significant stroke risk factors. These risk determinants demonstrated a multiplicative effect in general; however, the influence of some variables on stroke risk was not constant with age (hypertension, acute myocardial infarction, cardiac arrhythmias, obesity, and Rhesus factor) and sex (hypertension and education of head of household). The relationship of diabetes mellitus to stroke slightly decreased and became nonsignificant after adjustment for factors besides age and sex. Stroke risk was not associated with cigarette and alcohol use, family history of stroke and related disorders, marital status, and ABO blood typing.

cerebrovascular disorders; regression analysis

Stroke remains a major health problem in the world today. It is, moreover, a disorder requiring a multifactorial approach to the understanding of its etiology and, thus, its prevention. A number of factors (e.g., blood pressure and heart defects) are, at present, viewed as important in increasing one's stroke risk. Other factors remain more questionable in terms of their etiologic significance; in these cases,

Received for publication July 26, 1982, and in final form March 14, 1983.

[1] The Tilburg Epidemiological Study of Stroke and the Department of Neurology, St. Elisabeth Hospital and Maria Hospital, Tilburg, The Netherlands.
[2] Institute of Biostatistics, Erasmus University, Rotterdam, The Netherlands.
[3] Institute of Neurology, Catholic University, Nijmegen, The Netherlands.
Reprint requests to Dr. Herman, Project Director, Tilburg Epidemiological Study of Stroke, Klein Brabant 54, 5262 RM Vught, The Netherlands.

This investigation was supported by The Netherlands Fund for Preventive Medicine Grant No. 28-641 and The Netherlands Heart Foundation Grant No. 76.005.

The authors thank the medical staff and administration of the two Tilburg hospitals for their contribution to this investigation and H. Nunn-Smit and M. Akkermans-Takkenberg for their valuable assistance.

either there is less consensus regarding results noted from study to study or little attention is devoted to their investigation (1, 2). To help promote a better understanding of stroke risk determinants, a neuro-epidemiologic study group in Tilburg attempted to examine these factors by means of a case-control investigation.

In comparison with classical methods of analyzing matched case-control data (3–5), recent techniques have proven more flexible regarding the 1) types of data for study, 2) assessment of interactions, 3) control of multiple confounders, and 4) study of the joint influence of two or more factors. The present paper incorporates such more sophisticated approaches in examining the findings regarding risk factors for stroke in the city of Tilburg, The Netherlands. In doing so, it considers not only known major risk determinants, but other possible associations as well.

Materials and methods

A population-based stroke incidence register has been operating in the city of Tilburg since October 1, 1978 (6). A stroke is here defined as rapidly developed clinical signs of focal (or global: applied only to patients with subarachnoid hemorrhage and some patients in deep coma) disturbance of cerebral function, lasting more than 24 hours or leading to death, with no apparent cause other than a vascular origin. The diagnosis of stroke type is based on criteria set by both the World Health Organization (7) and the ad hoc Committee on Cerebrovascular Diseases of the National Institute of Neurological Diseases and Blindness (later, the National Institute of Neurological and Communicative Disorders and Stroke) (8, 9).

Case selection. Prospectively, from October 1, 1978, to July 31, 1981, 132 stroke cases (40–74 years of age, residing in Tilburg, of the Dutch nationality, having their first attack, and being treated at one of the two city hospitals) who were identified via the above-mentioned register underwent a specific study interview and examination, once their clinical condition allowed it (on the average, about two weeks after onset). These stroke patients were in most instances diagnosed as thrombotic infarction (86 per cent). Not all eligible cases could participate in the study. For example, of the 199 eligible thrombotic infarction patients, 42 (21 per cent) died and 44 (22 per cent) did not participate for various reasons (e.g., refusal, too ill).

Control selection. An attempt was made to obtain, for each case studied, two age- and sex-matched control patients admitted to the same hospital. Like the cases, controls were Tilburg residents who were 40–74 years of age, of the Dutch nationality, and without prior history of stroke. Age matching occurred on the basis of five-year age groups. The illnesses of these controls were not to have any known relationship to the suspect risk determinants and were, for the most part, of an acute surgical or nonsurgical nature. Persons with less acute sicknesses were likewise selected when such disorders were experienced for the first time, required hospital care, and began a short time before such hospital admission. The specific individuals chosen as candidates for study (i.e., undergoing the same questioning and examination as the cases) were the next patients admitted to a hospital after the case was studied who satisfied the above selection criteria. None of these patients died prior to the study procedures. However, 20 per cent of them did not participate because of refusal, severe illness, and so on. The final 239 controls studied included patients with the following conditions: 63 accidents (fractures and other injuries), 48 diseases of the urogenital tract (benign prostatic hypertrophy, uterine prolapse, etc.), 55 diseases of the digestive tract (inguinal hernia, appendicitis, etc.), 34 diseases of muscle,

bone, and connective tissue (meniscus, intervertebral disc hernia, etc.), and 39 other miscellaneous disorders.

Study variables. The analysis here relates primarily to the interview portion of the procedures that both cases and controls underwent within a similar study period in designated rooms of the neurologic departments of the two Tilburg hospitals. The interview involved the use of a structured questionnaire covering the following characteristics and experiences of the patients before their present illness: personal data (age, sex, marital status, and education of the head of the household); physical activity during leisure time (greatest portion of one's lifetime); cigarette consumption (amount and duration); alcohol intake (number of glasses of various alcoholic beverages per day during the last year and the greatest portion of one's lifetime); ever having clinically diagnosed acute myocardial infarction, cardiac arrhythmias, other heart conditions, hypertension, diabetes mellitus, and overweight (medical records for all subjects were further checked regarding the presence of each of these disorders); presence ever of possible transient cerebral ischemic attacks (i.e., acute focal deficits of the brain or of the retina, of vascular origin, which last no longer than 24 hours); and a history of stroke, high blood pressure, diabetes mellitus, and acute myocardial infarction among first-degree relatives (parents and siblings). All interviews were conducted by the same trained, nonmedical interviewer. She was able to obtain information personally from almost all subjects (questioning of family members was necessary for seven cases with extreme speech difficulties). The association of blood type with stroke risk was also studied. The necessary information was obtained through direct reports from the immunologic laboratory of the hospitals.

Statistical analysis. Two general methods of analysis were employed. First, for each factor under study, the method of Mantel-Haenszel (4) and its extension (10) was applied, for which eight strata were constructed by classification according to sex and three 10- and one five-year age intervals. The 95 per cent confidence intervals for the common relative risks were based on a procedure of Miettinen (11). If the categories of a polytomous variable had a natural order regarding level of exposure, a chi-square test for trend was performed (12, formula 4.43). When such sequence was not the case, the global null hypothesis that there is no effect of exposure on stroke was tested by means of the chi-square statistic described by Breslow and Day (12, formula 4.41).

To account for possible confounding factors (other than age and sex) simultaneously and to incorporate significant interaction terms, a more refined analysis was carried out using a conditional logistic regression model for matched data (12). These calculations were made by applying the program MATCH (12), which allows for matched sets consisting of a single case and a variable number of controls.

RESULTS

The distribution of the 132 cases and 239 controls by age and sex is shown in table 1. Since matching was done for age and sex, the age and sex distributions for cases and controls were very similar. Complete sets of two matched controls were obtained for 117 of the 132 cases. Five cases had only one control each. The 10 cases without controls were used only for the Mantel-Haenszel analysis. Table 2 summarizes the univariate findings regarding study factors.

The conditional logistic model is presented in table 3. This was arrived at after fitting a series of possible models, with good fit, significant relative risks (significance level of 0.05), and important confounding influences (not necessarily significant factors (13)) as criteria for the

TABLE 1

Age and sex distribution of stroke cases and controls, Tilburg, The Netherlands, October 1978–July 1981

Age (years)	Cases		Controls	
	Females	Males	Females	Males
	No. (%)	No. (%)	No. (%)	No. (%)
40–49	4 (3)	8 (6)	8 (3)	14 (6)
50–59	14 (11)	26 (20)	25 (11)	52 (22)
60–69	15 (11)	30 (23)	28 (12)	49 (21)
70–74	16 (12)	19 (14)	26 (11)	37 (16)
Total, all ages	49 (37)	83 (63)	87 (36)	152 (64)
Total, all ages and both sexes	132 (100)		239 (100)	

most appropriate model. Sex is coded as 0 (female) or 1 (male), and age is coded as 0 (40–49 years), 1 (50–59 years), 2 (60–69 years), or 3 (70–74 years). The codes for the other variables in the model are presented in the Appendix. None of the interactions between two risk factors reached statistical significance, so these terms were not incorporated into the final multivariate model. However, some of the interactions between a risk factor and the matching variables age and sex were found to be significant. Two main terms in the regression model, code 1 for hypertension and code 1 for obesity, were approximately zero (and nonsignificant) in the presence of their age and sex interaction terms. In the final analysis, these main terms were therefore omitted. In table 4, the conditional relative risk estimates (and 95 per cent confidence limits) for risk factors for stroke are summarized. These risk determinations were calculated from the model in table 3. By using the covariance matrix for the coefficients in this logistic model, the 95 per cent confidence limits could be obtained.

Personal data. No obvious relationship of marital status to stroke risk was noted ($\chi^2_2 = 3.42$, $p = 0.18$). However, the amount of schooling received by the head of the household had a significant overall association with stroke occurrence ($\chi^2_1 = 4.98$, $p = 0.03$). Individuals from households where wage earners had completed eight to nine years of schooling had 0.67 times the risk of those from homes where the head had only six to seven years of education. If the head of household had still further education, the relative risk dropped even lower (0.58). In addition, it can be seen from the logistic model (table 3) that education of the head of household has an interaction with sex. Apparently such education was only a risk factor for women (relative risks for intermediate and higher educational levels for females were 0.55 and 0.30, respectively; for males, these risks were 1.15 and 1.32) (table 4).

Health-related behavior. Physical activity during leisure time was found to have a significant association with stroke ($\chi^2_1 = 6.99$, $p = 0.01$). The multivariate regression analysis led to a relative risk for regular light activity of 0.49 and for regular heavy activity of 0.24. Several characteristics concerning smoking behavior and alcohol use were studied (table 2), but none demonstrated a significant relationship to stroke.

Medical history. A significantly higher stroke risk was observed among individuals who had experienced an acute myocardial infarction than those who had not. The Mantel-Haenszel estimate of the relative risk was 2.73. When the logistic regression analysis was utilized, an interaction of acute myocardial infarction with

TABLE 2

Distributions of cases and controls by study factors, relative risk (RR) and 95 per cent confidence interval (CI) estimates, and chi-square tests for polytomous factors, Tilburg, The Netherlands, October 1978–July 1981

Study factor	Cases No. (%)	Controls No. (%)	RR	95% CI	χ^2 tests*
			Personal data		
Marital status					
Married	90 (68)	182 (76)	1.00		
Single	14 (11)	15 (6)	1.83	0.87–3.85	
Divorced, separated or widowed	28 (21)	42 (18)	1.27	0.70–2.30	$\chi^2_2 = 3.42$ ($p = 0.18$)
Education of head of household (years)					
6–7	56 (42)	76 (32)	1.00		
8–9	43 (33)	82 (34)	0.67	0.40–1.13	
10+	33 (25)	81 (34)	0.58	0.35–0.97	$\chi^2_1 = 4.98$ ($p = 0.03$)
			Health-related behavior		
Physical activity during leisure time†					
Little	21 (16)	24 (10)	1.00		
Regular light	76 (58)	124 (52)	0.72	0.37–1.42	
Regular heavy	35 (27)	91 (38)	0.41	0.21–0.84	$\chi^2_1 = 6.99$ ($p = 0.01$)
No. of cigarettes smoked per day					
0	31 (23)	63 (26)	1.00		
1–10	34 (26)	47 (20)	1.51	0.76–3.02	
11–20	35 (27)	73 (31)	0.76	0.32–1.80	
>20	32 (24)	56 (23)	1.54	0.68–3.49	$\chi^2_1 = 0.01$ ($p = 0.93$)
Smoking exposure‡					
<25	34 (26)	67 (28)	1.00		
25–399	35 (27)	50 (21)	1.26	0.63–2.52	
400–699	25 (19)	45 (19)	1.43	0.57–3.57	
≥700	38 (29)	77 (32)	1.47	0.67–3.21	$\chi^2_1 = 0.16$ ($p = 0.69$)
Alcohol use§ (last 12 months)					
0	59 (45)	106 (44)	1.00		
1–29	36 (27)	73 (31)	0.89	0.52–1.53	
≥30	37 (28)	60 (25)	1.30	0.71–2.38	$\chi^2_1 = 0.15$ ($p = 0.69$)
Alcohol use§ (most of life)					
0	75 (57)	129 (54)	1.00		
1–9	18 (14)	26 (11)	1.19	0.61–2.31	
≥10	39 (30)	84 (35)	0.77	0.46–1.29	$\chi^2_1 = 0.85$ ($p = 0.36$)
			Medical history		
Acute myocardial infarction					
No	114 (86)	226 (95)	1.00		
Yes	18 (14)	13 (5)	2.73	1.32–5.64	
Cardiac arrhythmias					
No	112 (85)	228 (95)	1.00		
Yes	20 (15)	11 (5)	3.97	1.85–8.52	
Other heart condition					
No	105 (80)	215 (90)	1.00		
Yes	27 (20)	24 (10)	2.28	1.26–4.15	

TABLE 2 (*Continued*)

Distributions of cases and controls by study factors, relative risk (RR) and 95 per cent confidence interval (CI) estimates, and chi-square tests for polytomous factors, Tilburg, The Netherlands, October 1978–July 1981

Study factor	Cases No. (%)	Controls No. (%)	RR	95% CI	χ^2 tests*
			Personal data		
High blood pressure					
No	62 (47)	150 (63)	1.00		
Yes	70 (53)	89 (37)	1.88	1.23–2.87	
Diabetes mellitus					
No	111 (84)	222 (93)	1.00		
Yes	21 (16)	17 (7)	2.46	1.25–4.83	
Obesity					
No	75 (57)	169 (71)	1.00		
Yes	57 (43)	70 (29)	1.84	1.18–2.88	
Possible transient cerebral ischemic attack symptoms‖					
No	89 (67)	218 (91)	1.00		
Yes	43 (33)	21 (9)	5.22	2.98–9.13	
			Medical history of first-degree relatives		
Stroke					
No	103 (78)	181 (76)	1.00		
Yes	29 (22)	58 (24)	0.86	0.52–1.43	
Acute myocardial infarction					
No	75 (57)	154 (64)	1.00		
Yes	57 (43)	85 (36)	1.36	0.89–2.09	
Arterial hypertension					
No	100 (76)	174 (73)	1.00		
Yes	32 (24)	65 (27)	0.87	0.52–1.45	
Diabetes mellitus					
No	104 (79)	188 (79)	1.00		
Yes	28 (21)	51 (21)	0.99	0.59–1.66	
			Blood type		
Rhesus factor					
−	19 (14)	53 (22)	1.00		
+	113 (86)	186 (78)	1.66	0.95–2.93	
ABO grouping					
A	72 (55)	125 (52)	1.00		
B	8 (6)	15 (6)	1.03	0.40–2.65	
O	49 (37)	93 (39)	0.92	0.59–1.44	
AB	3 (2)	6 (3)	0.80	0.19–3.38	$\chi^2_3 = 0.12$ ($p = 0.98$)

* Chi-square tests were performed only for polytomous factors. If the categories of these factors had a natural order, a chi-square test for trend was carried out.

† Little = mainly sitting; regular light = regularly each week mainly light physical activity such as walking and cycling (recreational); regular heavy = regularly each week more intense physical activity such as heavy gardening and heavy sporting events.

‡ Number of cigarettes smoked per day × number of years smoking cigarettes.

§ Cubic centimeters of alcohol per day. The conversion of number of glasses of alcoholic beverages per day to cubic centimeters of alcohol consumed per day is based on data provided by the Dutch Traffic Safety Organization (14).

‖ A history ever of one or more of the following (all of acute onset and subsiding within 24 hours): 1) loss of vision in one eye; 2) diplopia; 3) motor weakness, numbness, or heaviness in one limb or both on the same side of the body; 4) difficulty with speech; and 5) sensory alteration involving one or both limbs or face on one side of the body.

TABLE 3

Conditional logistic regression model for data from Tilburg, The Netherlands, October 1978–July 1981

	Logistic coefficient	Standard error	Z-value*
Education	−0.604	0.321	−1.882
(Education) × sex	0.741	0.393	1.885
Physical activity, leisure time	−0.716	0.230	−3.113
Acute myocardial infarction	2.999	1.374	2.183
(Acute myocardial infarction) × age	−1.320	0.664	−1.988
Cardiac arrhythmias	5.856	3.103	1.887
(Cardiac arrhythmias) × age	−1.926	1.165	−1.653
(High blood pressure) × sex	1.507	0.546	2.760
(High blood pressure) × age	−0.408	0.222	−1.838
Diabetes mellitus	0.796	0.497	1.602
(Obesity) × age	0.394	0.159	2.478
Transient cerebral ischemic attack	1.991	0.454	4.385
Rhesus factor	−1.598	0.762	−2.097
(Rhesus factor) × age	1.385	0.401	3.454

* Z is the standard normal distributed standardized regression coefficient.

age was detected for both men and women. The relative risk was therefore higher among younger persons than older persons. For example, 50–59-year-olds with a history of acute myocardial infarction had, compared with those of the same age but lacking such disorder, a relative risk of 5.36; for individuals 60–69 years of age, this relative risk was 1.43. Comparable results were obtained regarding prior occurrence of cardiac arrhythmias. The relative risk estimate (Mantel-Haenszel) here was 3.97 (the risk being significantly different from 1). An interaction with age was likewise found with this variable: younger individuals with cardiac arrhythmias had a higher relative risk than older ones (for 50–59-year-olds, relative risk = 50.91; for 60–69-year-olds, relative risk = 7.42). A history of other heart conditions (for the most part, angina pectoris) presented a significant positive relative risk (2.28). However, after adjustment for other factors in the logistic model, this association became negligible, so the factor was omitted from the final regression.

A statistically significant overall relationship of stroke to high blood pressure was observed (95 per cent confidence limits for the relative risk being 1.23 and 2.87). A strong interaction with sex and a weaker one with age was further noted in the final multivariate model. Moreover, when similar logistic regressions were performed for each sex separately, high blood pressure appeared as a significant risk factor especially for men, more specifically for younger men. When the model in table 3 was used, hypertensive males 50–59 years of age had a relative risk of 3.00, while those 70–74 years had a relative risk of 1.33. A consequence of the absence of the main term for hypertension in the logistic regression was that for females (sex = 0) 40–49 years (age = 0), the relative risk for stroke, regarding high blood pressure, was exactly 1.

A medical history of diabetes mellitus also resulted in a greater relative stroke risk (2.46), but this risk slightly decreased and became nonsignificant after controlling for other factors in the multivariate analysis.

In contrast, a diagnosis of obesity remained as a significant risk factor for stroke, confirmed by both Mantel-Haenszel and logistic regression analysis. A significant interaction of overweight with age was detected. For example, the rela-

TABLE 4

Conditional logistic regression estimates of the relative risks (RR) and 95 per cent confidence limits for stroke risk factors, Tilburg, The Netherlands, October 1978–July 1981

Study factor	Category*	Stratum	RR	95% confidence limits
Education of head of household	8–9 years	Female	0.55	(0.29, 1.03)
		Male	1.15	(0.73, 1.80)
	10+ years	Female	0.30	(0.08, 1.05)
		Male	1.32	(0.60, 2.87)
Physical activity	Regular light	†	0.49	(0.31, 0.77)
	Regular heavy	†	0.24	(0.10, 0.59)
Acute myocardial infarction	Yes	40–49 years	20.07	(1.36, 296.74)
		50–59 years	5.36	(1.10, 26.16)
		60–69 years	1.43	(0.49, 2.58)
		70–74 years	0.38	(0.06, 1.95)
Cardiac arrhythmias	Yes	40–49 years	349.32	(0.80, >500)
		50–59 years	50.91	(1.07, >500)
		60–69 years	7.42	(1.21, 45.49)
		70–74 years	1.08	(0.26, 4.57)
High blood pressure	Yes	Female		
		40–49 years	1.00	‡
		50–59 years	0.66	(0.43, 2.32)
		60–69 years	0.44	(0.19, 1.05)
		70–74 years	0.29	(0.24, 1.08)
		Male		
		40–49 years	4.51	(1.55, 13.17)
		50–59 years	3.00	(1.32, 6.82)
		60–69 years	2.00	(0.93, 4.28)
		70–74 years	1.33	(0.52, 3.36)
Diabetes mellitus	Yes	†	2.22	(0.84, 5.87)
Obesity	Yes	40–49 years	1.00	‡
		50–59 years	1.48	(1.08, 2.03)
		60–69 years	2.20	(1.18, 4.11)
		70–74 years	3.26	(1.28, 8.33)
Transient cerebral ischemic attack	Yes	†	7.32	(3.01, 17.83)
Rhesus factor	Yes	40–49 years	0.20	(0.05, 0.90)
		50–59 years	0.81	(0.32, 2.07)
		60–69 years	3.23	(1.35, 7.76)
		70–74 years	12.91	(3.27, 50.98)

* All relative risks are calculated with respect to the baseline category 0 (see also Appendix).
† Relative risks are calculated for all persons regardless of age or sex.
‡ Confidence limits cannot be calculated because the standard error is zero.

tive risk for obese individuals 70–74 years of age was 3.26; the risk for those 50–59 years was 1.48.

Persons with a background of possible transient cerebral ischemic attacks presented a significant excess risk relative to those without such symptoms (relative risk of 5.22 if adjusted for age and sex only; 7.32 if corrected for all factors present in the final logistic model).

Medical history of first-degree relatives. The occurrence of either stroke, acute myocardial infarction, arterial hypertension, or diabetes mellitus among one's parents or siblings did not appear to influence stroke risk (table 2).

Blood type. Rhesus-positive individuals had a 1.66 times higher risk than those who were Rhesus-negative. A significant interaction with age could be determined,

with older Rhesus-positive persons having a higher relative stroke risk than those who were younger. On the other hand, ABO blood grouping demonstrated no association with stroke ($\chi^2_3 = 0.12$, $p = 0.98$).

Discussion

Conflicting results have been reported regarding the association of socioeconomic level with stroke morbidity (15–19). In Tilburg, an inverse relationship, specific to females, was noted between stroke risk and social status (using education of head of household as the indicator). Marital status, however, whether for males or females, did not differentiate those at greater risk for stroke.

The present findings indicate that leisure time physical activity is related to a substantial reduction in risk of developing a stroke. Very few findings on the association of habitual physical activity with stroke onset have been published. A recent longitudinal study of stroke morbidity in eastern Finland (20) found that low physical activity at work, but not in leisure time, was connected with an increased risk of cerebral stroke even after adjustment for confounding factors. As is the case in most other centers (1, 2, 21–23), cigarette smoking had little tie-in with one's risk of stroke in Tilburg. Alcohol use also appeared to have no bearing on stroke risk in the present research. Although some studies have noted an association of alcohol consumption with hemorrhagic stroke (24-27), its relationship to stroke in general and cerebral infarction in particular has not as yet been demonstrated.

High blood pressure was here, as in other investigations (1, 2, 21, 28, 29), an important component of the high stroke risk profile. A significant overall association between hypertension and stroke could be ascertained. After regarding the relationships within age and sex strata, it can be concluded that high blood pressure appeared to be a stroke risk factor primarily for younger males. Although the literature about the hypertension-stroke association within age and sex strata is rather sparse, our results appear to be consistent with reports from other studies. Salonen et al. (30) maintain that the association of stroke incidence with hypertension has not been found to be as strong in females as in males. Peacock et al. (24) found that males were more affected than females in terms of the relationship of thrombotic stroke and a history of high blood pressure. Librach et al. (31) state that hypertension appears to play a less important role regarding stroke risk among older persons compared with younger persons. Cardiac disorders are also consistently identified as very important risk factors for stroke from study to study (2, 21, 32–34). The findings in Tilburg substantiate these associations and point to acute myocardial infarction and, even more so, to dysrhythmias of the heart (particularly atrial fibrillation) as highly related. The extent of their connection with stroke, moreover, varies with age.

The literature regarding the role of obesity in stroke is more or less evenly divided; some studies note a positive association (15, 16, 23, 35, 36) while others do not (19, 24, 27, 31, 37). It appears from the research here that overweight is a definite risk factor for stroke in its own right (the connection being stronger with age) and not merely a consequence of its relationship to high blood pressure, diabetes mellitus, and various heart disorders. The precise mode of action of obesity has yet to be ascertained, although it has been suggested that an extra workload on the heart and increased intravascular volume might mediate its effect (38). It is unlikely that serum cholesterol, associated with overweight but not considered in the present investigation, confounded the relationship demonstrated. It has

proven a weak correlate of stroke risk, and when found related is specific to persons under 50 years of age (2).

Diabetes mellitus was not found to be directly associated with stroke after adjustment for other study factors. Obesity, especially, appeared to have a considerable confounding influence on the previously demonstrated relationship between diabetes and stroke when accounting for age and sex alone. Kuller (1) mentions that although clinical diabetes and stroke are frequently shown to be related in various studies, such an association may not be causal due to the former's connection with other stroke risk factors. Transient cerebral ischemic attacks are considered early clinical indications of a vascular pathology basic to stroke onset and, as such, important forewarnings of impending strokes (39). This study showed a significantly greater stroke risk among those experiencing such symptoms.

No convincing evidence exists for the role of family history of stroke and other related disorders in stroke risk. In Tilburg, the occurrence of stroke and other conditions considered stroke precursors (i.e., diabetes mellitus, high blood pressure, and acute myocardial infarction) among first-degree blood relatives (parents and siblings) did not increase one's chances of developing stroke. There is, thus, nothing from these data to suggest that something common to such close relatives (whether heredity, life-style, environment, or other factors) plays a part in stroke occurrence.

As in other studies (16), the present investigation's categorization of subjects by ABO blood grouping did not aid in identifying those with a higher stroke risk. What was unexpected, however, was the finding of a significant relation of Rhesus factor to stroke occurrence (more so for older persons). Confirmation of such results remains to be obtained.

The possibility that bias affected the study outcome should be considered. It is unlikely that the associations noted are spurious as a result of an improper selection of the study sample. Utilizing stroke registry data, it was possible to see that almost all (about 95 per cent) first-ever stroke cases 40–74 years of age in Tilburg during the study period were admitted to the hospital. Not all stroke patients under hospital care were interviewed, in large part because of death. This might lead to a bias in study outcome, if the factors being considered influence one's chances of dying from stroke a short time after the initiation of such care. It is very possible that if such bias were present, its effect would be an underestimation of associations noted because the study variables are more than likely to reduce one's survival chances, if anything. One of the most important requisites for a case-control study is the selection of controls from an appropriate population, in order to obtain unbiased relative risk estimates. Here, hospital controls with various acute diseases (as for stroke, necessitating hospitalization) considered unrelated to the suspect risk factors were chosen. Although, as is often the case, substantial evidence on which to base this latter conclusion is absent, it is most likely that, if an association of the risk variables with the diseases among control persons should exist, it would lead to a relative risk estimate closer to the value one than that of the unbiased determination.

Regarding possible bias in obtaining information, although the one interviewer (nonmedical) knew who cases and controls were, she did not appear to question one group more intensively than the other. The mean duration of an interview was 20.11 minutes for cases (standard error = 0.51) and 19.77 minutes for controls (standard error = 0.56) ($p > 0.60$). Not all study variables were found to be associated with stroke, making interviewer bias less suspect. Recall bias may have

contributed somewhat to the association of transient cerebral ischemic attack symptoms with stroke due to the more frequent prior questioning of stroke cases than controls concerning such complaints. However, regarding the remaining factors investigated, such bias may have had less influence because these experiences are not uncommon among people of the ages included in the present study and because both cases and controls underwent interviews within the same clinical setting. Furthermore, an examination of medical reports was of assistance in demonstrating good agreement between answers given by patients and recorded information.

References

1. Kuller LH. Epidemiology of stroke. In: Schoenberg BS, ed. Advances in neurology, Vol 19. Neurological epidemiology: principles and clinical applications. New York: Raven Press, 1978;281-310.
2. Ostfeld AM. A review of stroke epidemiology. Epidemiol Rev 1980;2:136-52.
3. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153-7.
4. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. JNCI 1959;22:719-48.
5. Miettinen OS. Estimation of relative risk from individually matched series. Biometrics 1970;26:75-86.
6. Herman B, Schulte BPM, van Luijk JH, et al. Epidemiology of stroke in Tilburg, The Netherlands: the population-based stroke incidence register. I. Introduction and preliminary results. Stroke 1980;11:162-5.
7. World Health Organization. Cerebrovascular diseases—prevention, treatment and rehabilitation. Technical Report Series No. 469. Geneva: World Health Organization, 1971.
8. Millikan CH, Adams RD, Fang H, et al. A classification and outline of cerebrovascular diseases. Neurology 1958;8:396-434.
9. Millikan CH, Bauer RB, Goldschmidt J, et al. A classification and outline of cerebrovascular diseases II. Stroke 1975;6:564-616.
10. Mantel N. Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. J Am Stat Assoc 1963;58:690-700.
11. Miettinen OS. Estimability and estimation in case-referent studies. Am J Epidemiol 1976;103:226-35.
12. Breslow NE, Day NE. Statistical methods in cancer research, Vol 1: The analysis of case-control studies. IARC Scientific Publications No. 32. Lyon: International Agency for Research on Cancer, 1980.
13. Dales LG, Ury HK. An improper use of statistical significance testing in studying covariables. Int J Epidemiol 1978;7:373-7.
14. Editorial: Alcoholgebruik stijgt. Tijdschrift voor Bejaarden- Kraam- en Ziekenverzorging 1975; 8:311.
15. Chapman JM, Reeder LG, Borun ER, et al. Epidemiology of vascular lesions affecting the central nervous system: the occurrence of strokes in a sample population under observation for cardiovascular disease. Am J Public Health 1966;56:191-201.
16. Johnson KG, Yano K, Kato H. Cerebral vascular disease in Hiroshima, Japan. J Chronic Dis 1967;20:545-59.
17. Pell S, D'Alonzo CA. Chronic disease morbidity and income level in an employed population. Am J Public Health 1970;60:116-29.
18. Abu-Zeid HAH, Choi NW, Maini KK, et al. Incidence and epidemiologic features of cerebrovascular disease (stroke) in Manitoba, Canada. Study of population 20-64 years of age. Prev Med 1975;4:567-78.
19. Abu-Zeid HAH, Choi NW, Maini KK, et al. Relative role of factors associated with cerebral infarction and cerebral hemorrhage. Stroke 1977;8:106-12.
20. Salonen JT, Puska P, Tuomilehto J. Physical activity and risk of myocardial infarction, cerebral stroke and death: a longitudinal study in Eastern Finland. Am J Epidemiol 1982;115: 526-37.
21. Stallones RA, Dyken ML, Fang HCH, et al. Epidemiology for stroke facilities planning. Stroke 1972;3:360-71.
22. Nomura A, Comstock GW, Kuller L, et al. Cigarette smoking and strokes. Stroke 1974;5:483-6.
23. Wolf PA, Kannel WB, Dawber TR. Prospective investigations: the Framingham study and the epidemiology of stroke. In: Schoenberg BS, ed. Advances in Neurology, Vol 19: Neurological epidemiology: principles and clinical applications. New York: Raven Press, 1978;107-20.
24. Peacock PB, Riley CP, Lampton TD, et al. The Birmingham stroke, epidemiology and rehabilitation study. In: Stewart GT, ed. Trends in epidemiology: application to health service research and training. Springfield, ILL: Charles C Thomas, 1972;231-345.
25. Kannel WB, Wolf PA, Dawber TR. An evaluation of the epidemiology of atherothrombotic brain infarction. Milbank Mem Fund Q 1975;53:405-48.
26. Kagan A, Popper JS, Rhoads GG. Factors related to stroke incidence in Hawaii Japanese men: the Honolulu Heart Study. Stroke 1980;11:14-21.
27. Tanaka H, Ueda Y, Hayashi M, et al. Risk factors for cerebral hemorrhage and cerebral infarction in a Japanese rural community. Stroke 1982;13:62-73.
28. Kannel WB, Wolf PA, Verter J, et al. Epidemiologic assessment of the role of blood pressure

in stroke: the Framingham study. JAMA 1970;214:301-10.
29. Marquardsen J. The epidemiology of cerebrovascular disease. Acta Neurol Scand 1978; 57[Suppl 67]:57-75.
30. Salonen JT, Puska P, Tuomilehto J, et al. Relation of blood pressure, serum lipids, and smoking to the risk of cerebral stroke: a longitudinal study in Eastern Finland. Stroke 1982;13:327-33.
31. Librach G, Schadel M, Seltzer M, et al. Stroke: incidence and risk factors. Geriatrics 1977;32: 85-96.
32. Friedman GD, Loveland DB, Ehrlich SP Jr. Relationship of stroke to other cardiovascular disease. Circulation 1968;38:533-41.
33. Kannel WB, Wolf P, Dawber TR. Hypertension and cardiac impairments increase stroke risk. Geriatrics 1978;33:71-83.
34. Wolf PA, Dawber TR, Thomas HE Jr. et al. Epidemiologic assessment of chronic atrial fibrillation and risk of stroke: the Framingham study. Neurology 1978;28:973-7.
35. Heyden S, Hames CG, Bartel A, et al. Weight and weight history in relation to cerebrovascular and ischemic heart disease. Arch Intern Med 1971;128:956-60.
36. Heyman A, Karp HR, Heyden S, et al. Cerebrovascular disease in the biracial population of Evans County, Georgia. Arch Intern Med 1971;128:949-55.
37. Okada H, Horibe H, Ohno Y, et al. A prospective study of cerebrovascular disease in Japanese rural communities, Akabane and Asahi. Part 1: Evaluation of risk factors in the occurrence of cerebral hemorrhage and thrombosis. Stroke 1976;7:599-607.
38. Gordon T, Kannel WB. The effects of overweight on cardiovascular diseases. Geriatrics 1973; 28:80-8.
39. Ramirez-Lassepas M. TIAs: the forewarning of stroke. Geriatrics 1980;35:73-83.

APPENDIX

Code for use of the logistic model

Sex	0 = female, 1 = male
Age	0 = 40–49 years, 1 = 50–59 years, 2 = 60–69 years, 3 = 70–74 years
Education	0 = 6–7 years, 1 = 8–9 years, 2 = 10+ years
Physical activity, leisure time	0 = little, 1 = regular light, 2 = regular heavy
Acute myocardial infarction	0 = no, 1 = yes
Cardiac arrhythmias	0 = no, 1 = yes
High blood pressure	0 = no, 1 = yes
Diabetes mellitus	0 = no, 1 = yes
Obesity	0 = no, 1 = yes
Transient cerebral ischemic attack	0 = no, 1 = yes
Rhesus factor	0 = negative, 1 = positive

CHAPTER 9

UNIVARIATE DOSE-RESPONSE MODELS IN CASE-CONTROL STUDIES

Univariate Dose-Response Models in Case-Control Studies. P. I. M. Schmitz. Biometrical Journal, 28 (1986). In press.

UNIVARIATE DOSE-RESPONSE MODELS IN CASE-CONTROL STUDIES

P.I.M. Schmitz

ABSTRACT

For modelling dose-response relationships in case-control studies the multiplicative logistic regression model, assuming the relative risk to be an exponential function of the dose, is widely known. If the relative risk is assumed to be a linear function of the dose, several authors (see e.g. Berry (1980)) have proposed an additive (linear) model. This model has a better fit with the data if such a linear relation holds. Confidence limits for the relative risk derived from the information matrix, however, appeared to be rather inaccurate. Therefore, use of the 'standard' logistic model in two different ways was studied: extension with a quadratic term or a logarithmic transformation of the dose. By applying the methods both to an empirical data set and in a simulation experiment, it is shown that appropriate transformation (often logarithmic) of the dosage and then applying the 'standard' logistic model is an useful approach if a linear dose-response relationship holds.

Key words: Case-control studies, dose-response, logistic model, relative risk.

1. INTRODUCTION

In modelling the dependence of the probability of a binary response D (disease incidence, death, etc.) on a set of p exposure variables, the use of multiple logistic regression is widely known (Breslow and Day (1980)). This logistic model implies that the odds ratio (which can be considered as an approximation of the relative risk if the disease incidence is low) is expressed as the exponent of a linear function of the explanatory variables. However, many disease processes do not follow a multiplicative model such as logistic regression. Sometimes, modelling of the odds ratio with an additive function, or a function between the multiplicative and the additive model, is more adequate (Walter and Holford (1978)). This may be the case for both forms of dose-response

relationships with a single exposure variable, and with forms of interactions between multiple exposure variables (Thomas (1981)). Therefore, the logistic model has been extended to general relative risk models, which express the odds ratio as an arbitrary function of the exposure variables. Thomas (1981) describes this approach for matched case-control data, with an arbitrary (variable) number of controls matched with each case. Storer et al.(1983) extend this approach to stratified case-control data with a variable number of cases and controls in each stratum. For non-stratified case-control studies Thompson and Baker (1981) describe the use of GLIM (Baker and Nelder (1978)) for fitting other than logistic models. For non-stratified case-control studies, Greenland (1983) has investigated the power of some tests for interaction (for the case with two binary risk factors) in additive and in multiplicative models. In this paper, a main objective is a comparison of some univariate dose-response models for non-stratified case-control studies when the data suggest a linear relation. This comparison is performed both with respect to bias of the risk estimates and with respect to the accuracy of the confidence intervals for the relative risk.

In the next section two main classes of relative risk models are described: additive and multiplicative models. Methods for discriminating between additive and multiplicative models are considered in section 3. Some of these approaches are illustrated by applying them to a non-stratified case-control data set in section 4. In section 5 two versions of the logistic model and the linear model are compared, when a linear relationship holds, by means of a simulation experiment. Some practical implications are discussed in section 6.

2. ADDITIVE AND MULTIPLICATIVE MODELS FOR THE RELATIVE RISK.

In linear logistic regression, the binary outcome variable D (disease incidence, death, etc.) is related to p explanatory variables x_i by the logistic function:

$$\Pr\{D=1 \mid x;\beta_0,\beta\} = \frac{\exp\{\beta_0 + \sum_{i=1}^{p} \beta_i x_i\}}{1+\exp\{\beta_0 + \sum_{i=1}^{p} \beta_i x_i\}} \qquad (2.1)$$

with $x = (x_1, x_2, \ldots, x_p)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$.

For this model the odds ratio with baseline value x=0 is:

$$RR(x;\beta) = \frac{Pr\{D=1|x\} \,/\, Pr\{D=0|x\}}{Pr\{D=1|0\} \,/\, Pr\{D=0|0\}} = \exp\{\sum_{i=1}^{p} \beta_i x_i\} \qquad (2.2)$$

Estimation of the parameters and testing of hypotheses concerning these parameters in unstratified case-control studies can be based on the unconditional likelihood function

$$L = \prod_{j=1}^{n_1} Pr_j\{D=1|x;\beta_0,\beta\} . \prod_{j=n_1+1}^{n_1+n_0} Pr_j\{D=0|x;\ \beta_0,\beta\} \qquad (2.3)$$

with n_1 = number of cases, and n_0 = number of controls.
This approach is widely known, see e.g. Breslow and Day (1980).
However, in many situations modelling of the odds ratio as an exponential function as in (2.2) is not appropriate.
Other functions, for example the additive model

$$RR(x;\beta) = 1 + \sum_{i=1}^{p} \beta_i x_i \qquad (2.4)$$

or intermediate models (between additive and multiplicative), could then better be fitted. Then, estimation of β_i and hypotheses testing may be based again on the likelihood (2.3). If the additive model (2.4) is taken, the conditional probability on the disease outcome is:

$$Pr\{D=1|x;\beta_0,\beta^*\} = \frac{\beta_0 + \sum_{i=1}^{p} \beta_i^* x_i}{1 + \beta_0 + \sum_{i=1}^{p} \beta_i^* x_i} \qquad (2.5)$$

with $\beta_i^* = \beta_0 \beta_i$, i= 1,2,...,p.

Both models (2.1) and (2.5) may be fitted using GLIM (Thompson and Baker (1981)). Guerrero and Johnson (1982) proposed a general relative risk model which may be viewed as an extension of the logistic model with a parameter λ, so that models (2.2) and (2.4) are obtained for specific values of λ.
We have written a special purpose computer programme for fitting multiplicative models (2.2) and additive models (2.4). Maximum likelihood estimates for the

parameters are calculated, using the Newton-Raphson algorithm. For the logistic model this is a standard method. For additive models there are some differences. In the Newton-Raphson approach for iterative solution of the likelihood equations, the matrix of negative second partials ('observed' information matrix) must be calculated. This matrix may be replaced by its expectation (Fisher information matrix; method of scoring; Rao (1965)). For logistic models these two matrices are equal, but for the other non-standard models they are different. Following the recommendations of Storer et al. (1983), the expected matrix is used. This (Thomas (1981)) also has the advantage of quicker convergence and has the certainty of a positive-definite matrix for inversion. Another difference is the existence of the ML solution. While for logistic models for most situations (some extreme ones excluded) the ML estimate exists, this is not the case for additive models. Besides this, if a ML solution for the additive model exists, finding of the numerical solution is rather sensitive for the initial values of the parameters. An appropriate scaling of the risk factors, combined with initial values not too far from the optimal ones, will, using the Newton-Raphson algorithm, generally lead to a convergent maximum likelihood estimate.

3. DISCRIMINATING BETWEEN ADDITIVE AND MULTIPLICATIVE MODELS

In this section an approach for discriminating between additive and multiplicative models will be outlined.
The additive model

$$RR(x_1;\beta_1) = 1 + \beta_1 x_1 \tag{3.1}$$

might be tested against the extended model

$$RR(x_1;\beta_1,\gamma_1) = 1 + \beta_1 x_1 + \gamma_1 x_1^2 \tag{3.2}$$

Similarly, the multiplicative model

$$RR(x_1;\beta_1) = \exp(\beta_1 x_1) \tag{3.3}$$

might be tested against the extended model

$$RR(x_1;\beta_1,\gamma_1) = \exp(\beta_1 x_1 + \gamma_1 x_1^2) \tag{3.4}$$

The additional parameter γ_1 may be viewed as a measure of departure from the

non-extended model. For testing this parameter three asymptotically equivalent tests are available: the likelihood ratio (LR) statistic G^2, the score statistic S and the Wald statistic W (see e.g. Breslow and Day (1980)).
After performing one or more of these tests, the results may be compared in order to make a decision on which model has the best fit: the additive or the multiplicative model. For example, when using the LR-test for H_0: $\gamma_1 = 0$ in (3.2) we obtain LR_I. Similarly, when using the LR-test for H_0: $\gamma_1 = 0$ in (3.4) we obtain LR_{II}. We might choose the model with the smallest LR. However, the two LR-tests do not test against the same alternative, so are, strictly speaking, not mutually comparable. Therefore, when studying univariate dose-resonse relationships, Thomas (1981) proposes to test against a mixture model, using the LR test, score test or Wald test. Using LR tests for model selection, this gives the same model choices as a direct comparison of the likelihoods: choosing the model with the larger likelihood (Walker and Rothman (1982); Greenland (1983)). Greenland (1983) compared the LR-, score- and Wald-tests, as 'tests for interaction' for two binary risk factors. The main results are (i) the low power of these tests (even for sample sizes until 500, cases plus controls), and (ii) likelihood comparison provides a useful model selection criterion when these tests fail to reject either model.
In this paper we will concentrate further on univariate dose-response models, while for the data a linear dose-response relationship holds. So it should be expected that the null hypothesis $\gamma_1 = 0$ in (3.2), using one of the three test statistics mentioned, is not rejected.

4. APPLICATION TO A NON-STRATIFIED CASE-CONTROL DATA SET

Data from a study among Quebec asbestos miners and millers, as presented in Table 1 of Thompson and Baker (1981), are given in Table 1. Dust exposure is related with lung cancer cases. We will treat these data as non-stratified case-control data, although they were originally matched. This theoretically may introduce a bias, but we only use the data for illustrative reasons. In the models we consider, dust exposure x is the lower limit of the concerning category.
We start with 2x2-table analyses (Table 2), and it is clear from that, that a linear model will show a good fit. Because the data are grouped , a goodness-of-fit test is possible now. An appropriate version of the loglikelihood ratio

Table 1. Distribution of dust exposure for cases and controls
(Thompson and Baker, 1981)

Dust exposure x (in units of (million particles per cubic foot). (years), up to 7 years before death of the case).

x	0	6	10	30	100	300	600	1000	1500	2000	Totals
Cases	43	10	24	37	31	27	18	10	6	9	215
Controls	285	62	166	211	168	95	50	19	8	11	1075
Totals	328	72	190	248	199	122	68	29	14	20	1290

statistic is:

$$G = 2 \sum_j O_j \log \frac{O_j}{E_j} \qquad (4.1)$$

with summation j over cells, O_j = number of observed subjects in cell j, E_j = number of expected subjects in cell j. For cells in category x $E = P_x n_x$ for cases and $E = (1-P_x)n_x$ for controls (n_x is the total number of cases and controls in category x; P_x is the probability on the disease for category x). The degrees of freedom df are calculated with df = (number of categories)-(number of parameters). In Table 3 the goodness-of-fit tests and tests for additional parameters for the multiplicative and additive models (3.1) to (3.4) are summarized. Both multiplicative models show a good fit, but because the quadratic term appears to be significant on a 10% level (keeping in mind the low power of these tests), the extended multiplicative model will be preferred. Also, both additive models fit well, but here the quadratic term obviously can be neglected. A further analysis will be performed for the linear model and the extended multiplicative model. In Table 2 the estimates and 95% confidence limits for the relative risk calculated with these models are presented. Although both models have a good fit, the confidence limits for the extended logistic model are wider than those for the linear model, especially the upper limit is higher for high dust exposure values. This is illustrated in Figure 1. The smaller confidence intervals for the linear model may have the consequence that this

Table 2. Analyses of the relative risks for the dust exposure data in Table 1.

dust exposure x	0	6	10	30	100	300	600	1000	1500	2000
2x2-table analysis										
odds ratio	1	1.07	0.96	1.16	1.22	1.88	2.39	3.49	4.97	5.42
95% conf.limits	-	(0.51-2.24)	(0.56-1.64)	(0.72-1.87)	(0.74-2.02)	(1.11-3.20)	(1.29-4.41)	(1.58-7.69)	(1.81-13.67)	(2.31-12.74)
width conf.int.	-	1.73	1.08	1.15	1.28	2.09	3.12	6.11	11.86	10.43
extended logistic model RR = exp $(0.001784x - 0.0000004859x^2)$ (G_7=0.70, p=1)										
odds ratio	1	1.01	1.02	1.05	1.19	1.63	2.45	3.66	4.87	5.07
95% conf.limits	-	(1.00-1.02)	(1.01-1.03)	(1.02-1.09)	(1.08-1.31)	(1.27-2.10)	(1.62-3.69)	(2.23-6.03)	(2.87-8.26)	(2.23-11.56)
width conf.int.	-	0.01	0.02	0.06	0.23	0.83	2.07	3.80	5.39	9.33
linear model RR = 1 + 0.002434x (G_8=0.50, p=1)										
odds ratio	1	1.01	1.02	1.07	1.24	1.73	2.46	3.43	4.65	5.87
95% conf.limits	-	(1.01-1.02)	(1.01-1.04)	(1.03-1.12)	(1.10-1.39)	(1.30-2.16)	(1.59-3.33)	(1.99-4.88)	(2.49-6.82)	(2.98-8.75)
width conf.int.	-	0.02	0.03	0.09	0.29	0.87	1.73	2.89	4.33	5.77
logistic model RR = exp (0.5178.log(x+100)) (G_8=1.07, p=0.99)										
odds ratio	1	1.03	1.05	1.15	1.43	2.05	2.74	3.46	4.20	4.84
95% conf.limits	-	(1.02-1.04)	(1.03-1.07)	(1.09-1.20)	(1.27-1.62)	(1.61-2.61)	(1.95-3.85)	(2.28-5.26)	(2.59-6.82)	(2.84-8.23)
width conf.int.	-	0.02	0.03	0.10	0.35	1.00	1.90	2.98	4.23	5.38

Table 3. Test statistics for some multiplicative and additive models for the dust exposure data in Table 1.

	Model	Goodness-of-fit test	Null-hypothesis	Test-statistics for the null hypothesis		
				LR-test	Wald-test	Score test
(1)	'No effect' RR=1	$G_9 = 32.72$ ($p<0.001$)	-	-	-	-
(2)	'Multiplicative model' $RR= \exp(\beta_1 x)$	$G_8 = 3.35$ ($p=0.91$)	$\beta_1 = 0$	29.34 ($p<0.001$)	31.44 ($p<0.001$)	36.94 ($p<0.001$)
(3)	'Extended multiplicative model' $RR=\exp(\beta_1 x+\beta_2 x^2)$	$G_7 = 0.70$ ($p=1$)	$\beta_2 = 0$	2.61 ($p=0.11$)	2.68 ($p=0.10$)	2.70 ($p=0.10$)
(4)	'Additive model' $RR=1 + \frac{\beta_1}{\beta_0} x$	$G_8 = 0.50$ ($p=1$)	$\beta_1 = 0$	32.14 ($p<0.001$)	15.78 ($p<0.001$)	36.94 ($p<0.001$)
(5)	'Extended additive model' $RR=1+\frac{\beta_1}{\beta_0} x + \frac{\beta_2}{\beta_0} x^2$	$G_7 = 0.45$ ($p=1$)	$\beta_2 = 0$	0 ($p=1$)	0.05 ($p=0.82$)	0.05 ($p=0.83$)

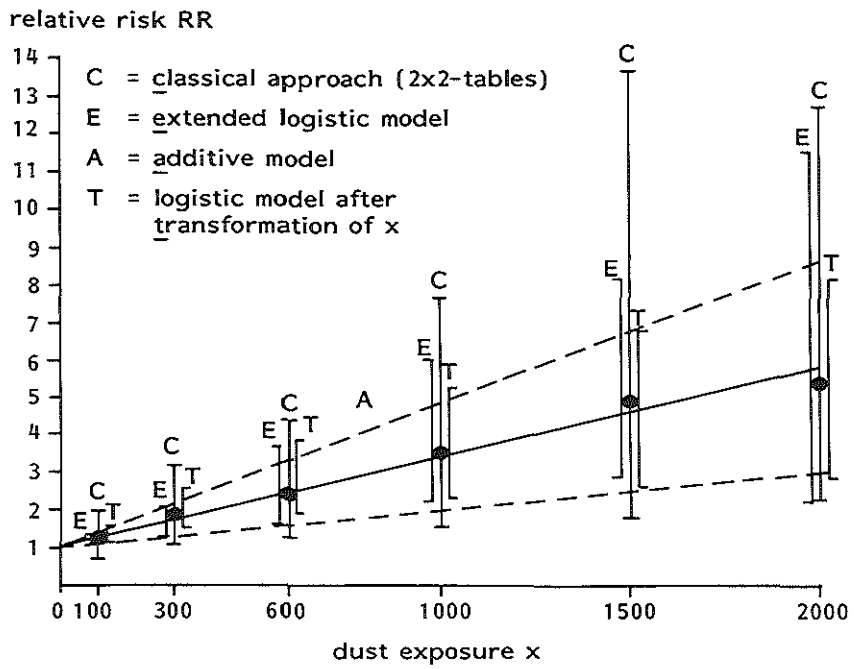


Figure 1. Point estimates and 95% confidence intervals of the relative risk obtained with four different approaches to the dust exposure data in Table 1.

model is preferred. However, it is uncertain, until so far, whether both models meet the conditions on which the calculations of the confidence intervals are based: the (asymptotically) normality of the parameters. This may be studied by using a measure of non-normality for parameters, as described by Sprott (1973), but we will use the more practical criterion as suggested by Thomas (1981). If there is a discrepancy between the Wald statistic and either the LR statistic or the score statistic, this is often an indication that the asymptotic normality does not hold.

For the dust exposure data, we see that these three statistics are in reasonable agreement for the extended multiplicative model, but that the Wald statistic is approximately the half of the LR and score statistic in the additive model (Table 3). This probably points out to the non-normality of the parameters in the additive model, so the relative 'favoured' confidence intervals for the

relative risk, based on this model, are biased. This does not mean that we have to choose the extended logistic model for these data. Other methods for constructing a confidence interval could be used, such as a direct search method based on the loglikelihood ratio, a method based on higher order derivatives of the loglikelihood, or using an appropriate reparameterization of the model parameter (Thomas (1981)). The methods based on the loglikelihood may fail to produce upper limits (Thomas (1981)). Especially when using unconditional likelihood, it may be difficult to find an appropriate reparameterization. Added to further disadvantages of additive model-fitting (see section 6), an appropriate adaptation of the logistic model may be preferred.

Still another possibility when using the logistic model will be studied: transformation of the dose x, actually a classical approach for dose-response relationships. To find a suitable constant c in the transformation log(x+c), a programme for non-linear modelfitting (based on weighted least squares) may be used, but when fitting 'by trial and error', c = 100 appeared to be a good choice for our data. The results for fitting the model

$$RR(x) = \exp\{\beta_1 \log(x+100)\} \qquad (4.2)$$

are summarized in Table 2. The model has a good fit (G_8=1.07; p=0.99), while the Wald statistic is in reasonable agreemeent with both LR- and score statistic for testing if the parameter β_1 is zero (Wald-χ_1^2=33.84, LR-χ_1^2=31.58, score-χ_1^2=35.69). So, the confidence intervals (Table 2) based on the asymptotic normality of $\hat{\beta}_1$ will be justified. Although the differences with the results of the additive model are not impressing it is shown in this manner that the additive model certainly is not the best choice for these dose-response data. In order to study whether the results from these analysis are consistent and may be generalized to other comparable data situations, a small simulation experiment is performed and described in the next section.

.5. COMPARISON OF THREE MODELLING APPROACHES FOR THE RELATIVE RISK

In order to compare the extended multiplicative model (3.4) (say model I), the additive model (3.1) (say model II) and the multiplicative model for transformed dose x (4.2) (say model III), a small simulation experiment was performed. Consider five possible dose-outcomes x, coded as 0,1,2,3 and 4 respectively (x=0 is the baseline category). Model III is based on the transformation

Table 4. Underlying risk factor distributions for cases and controls in a simulation experiment

risk factor x	0	1	2	3	4
cases	p_0	p_1	p_2	p_3	p_4
controls	q_0	q_1	q_2	q_3	q_4

$x' = \log(x+1)$. All situations have a common underlying risk factor distribution (see also Table 4). Among controls this distribution is multinomial with probabilities $q_0=q_1=q_2=q_3=q_4= 0.2$.
The distribution $\{p_0,p_1,p_2,p_3,p_4\}$ of the risk factor among cases is chosen in such a way that the odds ratios are linearly related with dose x:

$$RR(x) = 1 + x,\ x= 0,1,2,3,4. \qquad (5.1)$$

From (5.1) it follows that the probabilities of the multinomial risk factor distribution among the cases are:
$p_0 = 1/15=0.07$, $p_1= 2/15=0.13$, $p_2= 3/15=0.20$, $p_3=4/15=0.27$ and $p_4=5/15=0.33$.
If n_1 is the total number of cases and n_0 the total number of controls, the disease probabilities P_x are calculated with:

$$P_x = \frac{p_x n_1}{p_x n_1 + q_x n_0},\quad x= 0,1,2,3,4. \qquad (5.2)$$

In the simulation experiment, the number of controls was taken equal to the number of cases ($n_1=n_0=n$). For size n, two values were taken: 'moderate' sizes (n=300) and 'large' sizes (n=600).
From (5.2) and the distribution parameters p_i and q_i it is easily seen that the data follow the additive model

$$P_x = \frac{\beta_0+\beta_1 x}{1+ \beta_0+ \beta_1 x},\quad x = 0,1,2,3,4, \qquad (5.3)$$

with $\beta_0 = \beta_1 = 1/3 = 0.33$.
For generating the data, the random generator for a multinomial distribution GGMIN from the IMSL-library (see references) was used; the sequence of numbers for the controls was generated independently of the sequence of

Table 5. Simulation results for three models
(mean over 500 trials)

	Model I extended multiplicative $RR=\exp(\beta_1 x+\beta_2 x^2)$		Model II additive $RR=1+\frac{\beta_1}{\beta_0}$		Model III multiplicative for transformed x $RR=\exp(\beta_1 \log(x+1))$	
	$\lvert\hat{RR}-RR\rvert$	cov.rate†	$\lvert\hat{RR}-RR\rvert$	cov.rate	$\lvert\hat{RR}-RR\rvert$	cov.rate
sizes 300/300						
x=1, RR=2	0.32	91.6%	0.31	95.6%	0.19	95.8%
x=2, RR=3	0.71	95.0%	0.62	95.6%	0.47	95.8%
x=3, RR=4	0.99	95.4%	0.93	95.6%	0.80	95.8%
x=4, RR=5	1.08	96.0%	1.24	95.6%	1.18	95.8%
sizes 600/600						
x=1, RR=2	0.25	86.6%	0.20	92.4%	0.13	94.6%
x=2, RR=3	0.46	94.4%	0.40	92.4%	0.31	94.6%
x=3, RR=4	0.63	95.8%	0.60	92.4%	0.53	94.6%
x=4, RR=5	0.75	92.6%	0.80	92.4%	0.64	94.6%

† The coverage rate is the percentage of trials for which the 'true' relative risk is covered by the 95% confidence interval.

numbers for the cases. For each of the two sample size situations, 500 trials were produced. For each of the three models I, II and III, the absolute differences of the true relative risk and the estimated relative risk, $\lvert\hat{RR}-RR\rvert$, are calculated for categories x = 1,2,3 and 4. Further, for each model the coverage rate is calculated. This is the percentage of trials for which the 'true' relative risk is covered by the 95% confidence interval. For the fitted additive model this rate has the same value for all categories. However, for the two logistic models, especially the extended logistic model, this rate may differ over categories. The results (means over 500 trials) are summarized in Table 5.

From Table 5 it is seen that there are no substantial differences in bias $|\hat{RR}-RR|$ between the three models. This is consistent with the findings in section 4: these models were selected with respect to a good fit. The bias in model III is slightly less than in models I and II. The coverage rate, however, shows more serious differences between the models. For 'moderate' sizes models II and III reach the optimal 95%-value very close. For 'large' sizes model II has a coverage rate which is 2.6% below the optimal value, and model III only 0.4%. Both for 'moderate' and for 'large' sizes, the coverage rate for model I is too low in category x=1; for 'large' sizes this rate is too low also in category x=4.

6. DISCUSSION

A common practice for modelling univariate dose-response relationships for non-stratified) case-control data is fitting the logistic model. This model is based, however, on an exponential relationship between dose x and response RR(x). Situations occur, that the data are fitted better by a linear model. The fit of the multiplicative linear logistic model may be improved by extending the model with a quadratic term. On theoretical considerations, it may be expected that the confidence intervals for the additive model are smaller than those for the extended logistic model, supposing an underlying linear relation and a reasonable fit for both models. This is because the additive model reaches the same fit with one parameter less.
Nevertheless, additive model-fitting has several disadvantages in comparison with standard logistic fitting. First, there are more situations where the maximum likelihood solution does not exist. Second, the numerical solution (usually found by Newton-Raphson iterative fitting) appears to be rather sensitive for the initial values chosen. The third, and most serious, disadvantage is that the parameters in the additive model have a distribution which may show departures from the normal distribution. Therefore, confidence intervals based on the asymptotic normality of these parameters may be inaccurate. Several alternatives may be studied. Two of them, investigated in this paper, are small adaptations of the logistic model for one parameter:
(a) extension with a quadratic term
(b) transformation of the dose.
Both adaptations generally fit reasonably well, as confirmed in this study both by application to an empirical data set and by a simulation experiment. The

confidence intervals for the relative risk may be inaccurate when using an additive model. This accuracy was determined as the coverage rate in a simulation experiment. Especially for large sample sizes the coverage rate derived from the additive model may be lower than the optimal value of 95%: the confidence intervals are too narrow. From the two adaptations of the logistic model, only the model based on a transformation of the dose has a coverage rate close to the optimal value.

Not only because the confidence intervals are possibly inaccurate, but also because finding the maximum likelihood solution gives more numerical problems when using an additive model, it is suggested in situations where a linear dose-response relationship appears to hold, to use an appropriate transformation (often logarithmic) of the dosage and then applying the 'standard' logistic model.

REFERENCES

Baker,R.J. and Nelder,J.A. 'The GLIM System, Release 3', Numerical Algorithms Group, Oxford, 1978.

Berry,G. 'Dose-response in case-control studies', J.of Epid.and Comm.Health, 34, 217-222(1980).

Breslow,N.E. and Day,N.E. 'Statistical methods in cancer research. Volume 1 - The analysis of case-control studies', IARC Scientific Publications No.32, IARC, Lyon, 1980.

Greenland,S. 'Tests for interaction in epidemiologic studies: a review and a study of power', Statistics in Medicine, 2, 243-251(1983).

Guerrero,V.M. and Johnson,R.A. 'Use of the Box-Cox transformation with binary response models', Biometrika, 69, 309-314(1982).

IMSL Library, Edition 8, IMSL, Houston, 1980.

Rao,C.R. 'Linear statistical inference and its applications', Wiley, New York, 1965.

Sprott,D.A. 'Normal likelihoods and their relation to large sample theory of estimation', Biometrika, 60, 457-465(1973).

Storer,B.E., Wacholder,S. and Breslow,N.E. 'Maximum likelihood fitting of general risk models to stratified data', Applied Statistics, 32, 172-181(1983).

Thomas,D.C. 'General relative risk models for survival time and matched case-control analysis', Biometrics, 37, 673-686(1981).

Thompson,R. and Baker,R.J. 'Composite link functions in generalized linear models', Applied Statistics, 30, 125-131(1981).

Walker,A.M. and Rothman,K.J. 'Models of varying parametric form in case-referent studies', American Journal of Epidemiology, 115, 129-137(1982).

Walter,S.D. and Holford,T.R. 'Additive, multiplicative, and other models for disease risks', American Journal of Epidemiology, 108, 341-346(1978).

CHAPTER 10

SELECTION OF VARIABLES IN ETIOLOGIC-EPIDEMIOLOGICAL STUDIES

Selection of Variables in Etiologic-Epidemiological Studies. P. I. M. Schmitz. T. v. Soc. Gez. 60 (1982) 851 - 854. In Dutch.

SELECTION OF VARIABLES IN ETIOLOGIC-EPIDEMIOLOGICAL STUDIES.

P.I.M. Schmitz

SUMMARY

Since high speed computers are available, it is common that in etiologic-epidemiological studies (cohort and case-control studies) many variables (100-200) are considered. In order to obtain estimates of risks which are as unbiased as possible, a multivariate analysis with subsets of these factors must be performed. In situations with so many variables an automatic stepwise selection procedure incurs the risk of serious biases and is therefore not be recommended.

A simple selection strategy for extensive etiologic-epidemiological investigations is introduced. This strategy is based on a division of the study variables in groups of similar variables, so-called batteries. First, variables are selected within each battery, next these selected variables are combined to form one set, and from this set the final selection is performed.

1. INTRODUCTION

Computerized stepwise procedures for the selection of variables in etiologic-epidemiological studies may hide important characteristics of the study data. These methods are directed to the selection of a subset of variables with good discriminatory power, much more than to an adequate assessment of the effect of a particular risk factor. Especially if the effects of potential risk factors are the subject of study, these selection procedures are not to be recommended. An alternative approach for the selection of variables is described in this paper.

1.1 Etiologic-epidemiological studies.

The two designs most frequently applied in observational etiologic-epidemiological studies, in order to investigate possible causal relationships between risk factors and disease occurence, are cohort and case-control studies (MacMahon and Pugh, 1970). In cohort studies disease incidence is compared between individuals with different exposure outcomes. In case-control studies a sample is taken from individuals which have the disease (the cases) and a sample from individuals which do not have the disease (the controls) The values of the potential risk factors are then obtained, in general by means of interviews. The considerations in this paper concern both designs, and we will use therefore the term 'observational study' for them shortly.

1.2 Developments in observational studies.

The decision to initiate an observational study may have different reasons. For example, an unexpected high incidence of certain cases of cancer in a hospital may lead to searching for factors possibly associated with the disease cases. This search may end up in a systematically designed case-control study. In such an investigation, beside potential risk factors, some possibly confounding variables (confounders) such as age and sex will also be recorded. Such an analytical approach is generally concerned with the question: does a causal relationship between risk factor F and disease D exist? With help of both simple univariate statistical techniques and more complex multivariate models relevant hypotheses can be tested. Besides, the relative risk for a certain exposure value of the risk factor can be estimated.

Methods for the design and analysis of such analytical observational studies have been extensively described in the literature. Especially about case-control studies some excellent monographs have been published recently (Breslow and Day, 1980; Schlesselman, 1982). In addition to analytical studies, more and more explorative observational studies are performed which is particularly stimulated by the availability of fast and inexpensive computers. The explorative study ('fishing expedition') is characterized by the lack of pre-specified hypotheses, especially if there is little known about causes of the disease concerned. The emphasis is on generating hypotheses rather than

on testing. The explorative approach may be followed also with the intention to gather additional information where some causal relationships are known. Most diseases have multifactorial causes, and in this way it is possible to detect etiologic factors which are as yet unknown. The lacking of pre-specified hypotheses at the start of an explorative study implies that this approach requires many variables (potential risk factors and confounders) to be recorded. Especially in studies that involve psycho-social factors (and consequently numerous items), the number of variables will readily increase to 100 or 200. The problem of variables selection arises therefore particularly in explorative observational studies. It is almost self-evident that the distinction between an analytical and an explorative study cannot always be made in practice. Many analytical studies have an explorative aspect. However, it is a prerequisite at the start of any study to distinguish between (causal) relations that have been established earlier and those that are not. That is because this prior information influences the decisions to be made in our selection strategy.

1.3 The problem of selection.

Consider an explorative observational study in which data have been collected involving a large number of variables (at least 100). The relative risks for the disease associated with several potential risk factors are assessed by means of statistical methods for estimation and testing. The usual procedure is first to give univariate descriptions of the variables recorded, including raw relative risk estimates. Next, adjusted estimates of relative risks, based on stratification techniques, are calculated. For that purpose the association between an exposure factor and presence of disease is analyzed within a limited number of strata formed by the most important confounding factors (often age and sex). After estimation of the raw relative risk in the first analysis phase and the adjusted relative risk in the second phase, a multivariate analysis of the data is performed in a third phase. In this analysis adjustment for several confounders is possible in a flexible way, and hypotheses concerning the simultaneous effect of two or more risk factors and their interactions may be studied (Breslow and Day, 1980; Schlesselman, 1982).

Apart from statistical intricacies concerning the interpretation of p-values in explorative investigations (see the discussion), the first two analysis phases

mentioned above are rather straightforward. It is in the multivariate analysis phase, however, that the investigator is confronted with the problems of selection. The number of variables that can be studied simultaneously in a multivariate model is limited for statistical and technical reasons. Dependent on sample size and computer facilities this model will usually consist of not more than 10 to 20 variables (risk factors, confounders, interactions, etc.). In the statistical literature methods have been described for model selection with a limited number of terms, but guidelines for a selection strategy starting from much larger numbers of variables have not been found. Except that, stepwise computerized automatic selection methods based on significancies are well known. This latter approach is not to be recommended, however, for the studies under consideration (see the next section). Yet, the investigator has to find the best variables to be included in a definite model. These are the problems of choice which are discussed in this paper.

2. STEPWISE SELECTION METHODS.

A well known method for the selection of variables in a multivariate model (such as in multiple linear or logistic regression) is a stepwise selection procedure, based on significance testing or a comparable evaluation. Especially after the implementation of such procedures in computer packages these methods have received some popularity. Forward selection is usually applied in situations with very many variables. This means that the factor with the most significant effect will be selected first. In the following step the most significant term of the remaining variables will be added to the model. This process is repeated until a certain criterion threshold is crossed. If the number of variables is not too large, a backward procedure may be preferable (Mantel, 1970).

Anyhow, these stepwise methods cannot be recommended for observational studies for a number of reasons.

(a) The procedure treats each variable in an equivalent manner. It will lead to a set of variables with possibly good discriminatory power, rather than to an adequate assessment of the effects of individual risk factors (adjusted for confounding). In some fields, for example in medical decision making, where a correctly predicted outcome is the only aim, such an approach may be very useful. It does not matter much if a set of variables with good prediction is

replaced by another set with equally good prediction. Stated differently: only the success rate of the final prediction is of interest, rather than the influence of each factor separately. In observational epidemiological studies, however, quantitative assessment of the causal association between each risk factor separately and the disease is the very first objective.

(b) Computerized automatic stepwise selection procedures do not use subject matter information on the medical significance of a factor or association. Usually at most a certain priority for inclusion in the model can be assigned to the variables.

(c) Stepwise selection procedures, based on statistical tests, do not distinguish risk factors from confounders. Selection of confounders may lead to incorrect conclusions (Dales and Ury, 1978).

Because of these objections it is clear that the use of an automatic stepwise selection method in observational studies cannot be recommended generally. An alternative approach will be proposed in the next section.

3. A SIMPLE SELECTION STRATEGY.

The selection strategy described below may be divided in three phases.

Phase 1. Grouping into batteries.

Group the variables into batteries of similar variables. For example one battery with 'personal data' (such as civil status, age, sex and education), a second battery with 'health - related behaviour' (such as smoking habits and physical activity), a third battery with 'psycho-social factors' etc. Precise guidelines for the number of batteries and the number of variables in each battery cannot and need not be given. It is dependent on the type of study and besides it is somewhat subjective. A feasable grouping of a total number of 100 variables, might be into 10 batteries with on average 10 variables in each battery. It is strongly recommended to take into account such groupings already in the design phase of the study (Cox and Snell, 1974).

Phase 2. Selection within a battery.

Within each battery (consisting of say 10 to 20 variables) the significant risk factors and the possibly confounding factors are to be selected. This phase can be conceived as divisible into two stages.

(a) A first step consists of screening the variables on their mutual associations. It is desirable that from two highly correlated factors only one is taken into the multivariate model in order to prevent problems of collinearity. (Collinearity is a phenomenon in regression models that occurs when two highly correlated factors are both in the model. The coefficient estimates of these factors are then very unstable: another sample may yield quite different results). A related situation occurs if two factors A and B are in the same causal chain. For example, A causes B, B causes the disease. Then, only one of the factors A and B should be selected. Considerable reduction of variables may also be possible in this phase by the use of sumscores for similar variables, especially for psycho-social factors.

(b) For those variables that remain within a battery the best fitting model has to be searched. Although in this stage the potential number of factors in the model may be quite manageable, an automatic stepwise procedure is still not feasible (Henderson and Velleman, 1981). The best 'feeling with the data' is obtained through the analysis of a great number of possible models. Among them the following are almost obligatory:

- all factors simultaneously in one model
- each factor separately in one model
- every obvious model between these extremes
- some reasonably good fitting models with interaction terms added.

From the models thus obtained the 'best fitting model' must be selected. Selection criteria are:

- statistical significance of a risk factor
- consistency of a risk factor viewed over several models (effect remaining similar in all models)
- substantial influence of a confounder (see also the discussion).

The selection strategy in this phase using logistic models in case-control studies was excellently demonstrated by Breslow and Day (1980, chapters 6 and 7).

Phase 3. Merging of batteries.

In the final phase the variables selected within each battery are joined together. From this set the final best fitting model is selected using the methods as described for phase 2.

4. DISCUSSION.

Some comments concerning the selection strategy proposed in section 3 are now in order.

4.1 Selection of confounders.

The final objective of the selection procedure is a multivariate model that describes the data satisfactory with a minimum number of parameters. The selection of the risk factors in this model is based mainly on p-values concerning significance of regression coefficients in the multivariate model. As was noted before, parallel with risk factors also confounders are selected. For the latter a significance criterion is not suitable. A frequently applied approach for detecting confounders with substantial effects is the following. Assume that the relative risk ('effect') of a risk factor E is RR1 when the confounder C is incorporated in the model and RR0 when the confounder is not. If the difference between RR1 and RR0 is not negligable, then C is to be considered a confounder for the association between E and the disease. Consequently, a (subjective) threshold for this difference has to be chosen. Rigid application of this approach may raise problems in the selection procedure. A variable from battery 1 which is only a confounder for a risk factor in battery 2 will not be included upon strict application of the strategy. Therefore it is desirable that variables for which a substantial confounding effect may be expected, will not be deleted during phase 2 (selection within a battery). Those variables should be selected in phase 3 (merging of variables).

4.2 Interpretation of p-values.

The interpretation of p-values in observational studies, and certainly in explorative observational studies, is questionable. The estimation of an effect (risk) is generally more important (in observational studies) than the testing of hypotheses for this effect. It is acceptable to view p-values as a guideline for the interpretation of the data (Cox, 1977). Rigid application of fixed significance levels (e.g. 0.05 or 0.01) has to be dissuaded. Important factors may be 'missed' in this way, especially with moderate sample sizes. In the selection procedure proposed in this paper p-values are used as an aid for

selection. For inclusion in the model a reasonable threshold is $p = 0.05$ or $p = 0.10$, but this rule has to be handled flexibly, dependent on prior information. For example, a risk factor, known to be important from the literature, with a p-value of 0.07 in a certain phase of the selection procedure, should not be deleted even if a threshold 0.05 is used. Decisions on how far we may go in such situations have a subjective character without any doubt. Maintaining a rigid threshold, however, does not improve the quality of the selection procedure. Above this: in situations where the study clearly has an explorative nature, the significant effects have to be evaluated always in a subsequent analytical study.

4.3 Studies with relatively few variables.

In studies with fewer variables (say less than 20) the selection procedure proposed in this paper may seem unnecessary. It is not sufficient, however, to use a once-only analysis then with all variables together in one multivariate model. Unfortunately that is what has been done in numerous studies. This is also stimulated by the generally available computer program packages. Without detailed investigation of possible models, supported by univariate stratification techniques, the result of a multivariate analysis may be quite unreliable.

4.4 Concluding remarks.

The selection procedure for observational epidemiological studies, as introduced in this paper, must be viewed as a possible approach, that certainly is not to be used too strictly. There is not a 'best selection method' and also there is not 'a best set of variables' or 'a best model'. The procedure described here is one of the possible alternatives, necessary when other known approaches (automatic stepwise methods) may well lead to incorrect results. It is like Schlesselman (1982, p.62) remarks: 'In any analysis the choice of variables is largely a subjective matter. It should depend on one's hypothesized biological model for the disease process, and on one's assumptions regarding potential sources of bias that may distort the magnitude of the estimated association between exposure and disease. Beyond this, one can give little advice that will be incontrovertible.'

REFERENCES

Breslow,N.E. and Day,N.E.: Statistical methods in cancer research, vol.I. The analysis of case-control studies. IARC Scientific Publications No.32. International Agency for Research on Cancer, Lyon, 1980.

Cox,D.R.: The role of significance tests (with discussion). Scand.J.Stat. 4(1977)49-70.

Cox,D.R. and Snell, E.J.: The choice of variables in observational studies. Applied Statistics 23(1974)51-59.

Dales,L.G. and Ury,H.K.: An improper use of statistical significance testing in studying covariables. Int.J.Epid. 7(1978)373-375.

Henderson,H.V. and Velleman,P.F.: Building multiple regression models interactively. Biometrics 37(1981)391-411.

MacMahon,B. and Pugh,T.F.: Epidemiology: Principles and methods. Little, Brown and Co., Boston, 1970.

Mantel,N.: Why stepdown procedures in variable selection. Technometrics 12 (1970)621-625.

Schlesselman,J.J.: Case-control studies. Design, conduct, analysis. Monographs in epidemiology and biostatistics. Oxford University Press. New York, 1982.

SUMMARY

Samenvatting

SUMMARY

The past 10 to 15 years have witnessed important developments in the methodological approach of medical research, especially in the fields of medical decision making and of etiological epidemiology. These developments have been accelerated by the growing pressure on clinicians for more cost-effective use of medical resources. Simultaneously, computer systems have become more sophisticated, accessible and relatively inexpensive. Therefore the more widespread use of modern quantitative methods from biostatistics has become feasible. This thesis is concerned with one such method which is based on the logistic regression model. The logistic model is reviewed as to its developments, compared with other models and shown to be applicable in a variety of situations occurring in practice.

Chapter 1 reviews important theoretical developments of the multiple logistic regression model from 1970 onward. Introduction of the conditional likelihood approach made unbiased estimation in matched case-control studies possible. Several useful procedures for testing goodness of fit have been introduced and sophisticated regression diagnostics for detecting model departures have been described. Generalizations to outcome variables with more than two response classes are summarized as are models of which the logistic is a special case.

Chapter 2 deals with an analysis of the merits of four diagnostic tests (agglutination tests) for the differential diagnosis of patients with Crohn's disease, patients with ulcerative colitis and control subjects. From the results of a polychotomous logistic discriminant analysis it is concluded that differentiation of patients with ulcerative colitis either from patients with Crohn's disease or from healthy subjects is hardly possible on the basis of these agglutination tests. A dichotomous logistic model for differentiation between Crohn's and non-Crohn's was extensively validated, both by comparison with alternative discrimination methods and by evaluation of the model performance on different datasets: Dutch validation cases and international cases. From the three discriminant analysis models compared (Fisher's linear discriminant analysis, an independence model and the logistic model) the logistic model shows better reliability ("goodness of fit"), while discriminatory performance is apparently comparable. Evaluation on different datasets was considered for the logistic model only. A relatively strong deterioration of the performance of the logistic model in both the Dutch validation cases and the international cases is apparent. This is presumably due to lack of precise definitions for the several diagnostic categories in these latter datasets. Chapter 3 is concerned with the first application of logistic discriminant analysis to agglutination tests for distinguishing patients with Crohn's disease and healthy subjects. It introduces a more differentiated interpretation of the estimated probability of Crohn's disease.

Chapters 4, 5 and 6 are concerned with the comparison of logistic discrimination with some other discriminant analysis methods. In chapter 4 we consider the IMIR (Imminent Myocardial Infarction study Rotterdam) dataset, consisting of continuous and binary variables. Four discrimination methods are applied to these data: LOG (logistic model), LDA (Fisher's linear discriminant analysis), QDA (quadratic discriminant analysis) and KER (a kernel approach). It is concluded that the resubstitution method for performance evaluations leads to an overly optimistic picture, especially for the kernel approach. Application to independent test samples showed that LOG, LDA and KER are nearly identical with respect to discriminatory performance measures, and are preferable to QDA. A simulation study to compare LOG, LDA, QDA and KER for a wide range of mixed continuous and discrete data structures is described in chapter 5. The mixed data are generated from the fourdimensional normal distribution, with three of the variables being discretized. The results show an almost identical performance of Fisher's linear discrimination and logistic discrimination. A choice from LDA and KER seems to be appropriate as far as discriminatory ability is concerned. It is suspected that the results of this study are dependent on the underlying normal distribution, however. Because of this and because the data structures studied may not be detectable in actual applications, a second simulation study was performed. Chapter 6 describes this second simulation study, where the same four discrimination techniques as in chapter 5 are compared, together with a fifth method, an independence model (IND). Much attention was paid to the design of this simulation study, where data were generated according to a particular choice of a location model. The results of this study indicate that the interaction and correlation structure of the mixed dataset determine whether LDA and LOG at one hand, or QDA and KER at the other hand perform better. Remarkably good discriminatory performance is reached by using the combination of LDA and QDA, which means that the "augmented LDA-approach" generally will be a very good choice. If these results are integrated with those from chapter 2, it is clear that if besides discriminatory ability, also a good reliability of the discrimination rule is desired, the "augmented LOG-approach" will give satisfactory results. Our experience with these simulation studies resulted in an application of logistic discrimination and the four discriminant analysis methods mentioned above (LDA, QDA, KER, IND) to an actual dataset. The data are obtained from a pharmaceutical industry, and concern the relationships of biological activity variables to a set of characteristics of chemical compounds, including information about their molecular structure: see chapter 7. In this QSAR-study (quantitative structure-activity relationship) discriminant analysis is used to identify characteristics which significantly affect biological activity. In this field logistic discriminant analysis is yet relatively unknown.

An application of the multiple logistic model for the analysis of matched case-control studies is presented in chapter 8. The study subjects include stroke patients and age- and sex-matched control subjects. A series of multiple logistic models were studied in order to assess joint effects and possible interactions and to control for possible confounding factors. Several significant stroke risk factors, among which hypertension and acute myocardial infarction, were assessed. Because the influence of some variables on stroke risk was not constant for age and sex differences the relevant interaction terms were also incorporated into the conditional logistic regression model. As a consequence estimates of the relative risks of some factors have to be presented within age and sex strata.

Chapter 9 deals with modelling dose-response relationships in case-control studies. The standard logistic model, extended with a quadratic term, and a linear model were applied to both an empirical dataset and simulated data. It appears that the standard logistic model gives useful results if a linear dose-response relationship holds.

In chapter 10 a selection strategy for extensive etiologic-epidemiological studies is introduced. This strategy is based on a division of the study variables into groups of similar variables (batteries). As a start, variables are selected within each battery, next these selected variables are combined to form one set, and from this set the final selection is performed. This approach is recommended for use with the logistic model.

Throughout this thesis the underlying theme is the assesment of when and how to apply the multiple logistic regression model to problems arising in the extensive fields of medical decision making, which concern diagnostic, therapeutic and prognostic assessments, and of epidemiology, especially case-control and cohort studies. Implementation of this model in widely available computer packages like BMDP, shows that it has proven its important role among modern advanced statistical techniques. Meanwhile, the model is developing still further. To conclude, it is worth stressing that adequate application requires insight both in the clinical subject matter and in the statistical theory of model fitting. Close collaboration between subject matter experts and biostatisticians is therefore needed for the benefit of medical research and practice.

SAMENVATTING

In de afgelopen 10 tot 15 jaar hebben belangrijke ontwikkelingen plaats gevonden in de methodologie van medisch-wetenschappelijk onderzoek, vooral op het gebied van de medische besliskunde en van de etiologische epidemiologie. Deze ontwikkelingen zijn versneld door de toenemende druk op clinici om een kosten-effectief gebruik van het medische potentieel te maken. Tegelijkertijd zijn computers geavanceerder, toegankelijker en relatief goedkoper geworden. Hierdoor konden moderne kwantitatieve methoden uit de biostatistiek in ruimer verband worden geïmplementeerd. Dit proefschrift behandelt één zo'n methode, gebaseerd op het logistische regressiemodel. De ontwikkelingen van het logistische model worden beschreven, het model wordt vergeleken met andere modellen, en de toepasbaarheid ervan in tal van uiteenlopende praktijksituaties wordt getoond.

Hoofdstuk 1 geeft een overzicht van belangrijke theoretische ontwikkelingen van het multipele logistische regressiemodel vanaf 1970. Zuivere schatting in gematcht case-control onderzoek werd mogelijk door invoering van de conditionele likelihood methode. Technieken voor het toetsen van de mate van aanpassing (goodness of fit) werden ontwikkeld en geavanceerde regressie-diagnostiek voor het opsporen van modelafwijkingen werd geïntroduceerd. Voorts wordt aandacht geschonken aan generalisaties naar respons variabelen met meer dan twee categorieën, en klassen van modellen waarvan het logistische een bijzonder geval is.

Hoofdstuk 2 behandelt een analyse van de merites van vier diagnostische testen, (agglutinatie testen) voor de differentiaaldiagnose van patiënten met de ziekte van Crohn, patiënten met colitis ulcerosa en controle personen. De resultaten van een polychotome logistische discriminantanalyse leiden tot de conclusie dat differentiatie van patiënten met colitis ulcerosa van òf patiënten met de ziekte van Crohn òf van gezonde personen nauwelijks mogelijk is op basis van deze agglutinatietesten. Een dichotoom logistisch model voor de differentiatie tussen patiënten met en zonder de ziekte van Crohn werd uitgebreid gevalideerd, zowel door een vergelijking met andere discriminantanalyse methoden als door een evaluatie van de werking van het model met verschillende gegevensverzamelingen: de Nederlandse validatie cases en internationale cases. Van de drie discriminantanalyse modellen die werden vergeleken (Fisher's lineaire discriminantanalyse, een onafhankelijkheidsmodel en het logistische model) laat het logistische model een betere aanpassing zien, terwijl het discriminerend vermogen van de modellen weinig verschilt. Validatie met verschillende gegevensverzamelingen werd alleen voor het logistische model beschouwd. Er blijkt een relatief sterke afname van de werking van het logistische model bij zowel de Nederlandse validatie cases als de internationale cases op te treden. Dit is waarschijnlijk toe te schrijven aan het ontbreken van preciese definities van de verschillende diagnostische categorieën bij

deze laatste cases. Hoofdstuk 3 behandelt de eerste toepassing van logistische discriminantanalyse voor agglutinatie testen voor het onderscheiden van patiënten met de ziekte van Crohn en gezonde personen. In dit hoofdstuk wordt een genuanceerde interpretatie ingevoerd van de geschatte kans op de ziekte van Crohn.

In hoofdstuk 4, 5, en 6 wordt de vergelijking van logistische discriminantanalyse met andere discriminantanalyse methoden behandeld. In hoofdstuk 4 worden de IMIR-gegevens (Imminent Myocardial Infarction study Rotterdam) beschouwd, bestaande uit continue en binaire variabelen. Vier discriminantanalyse methoden worden op deze gegevens toegepast: LOG (logistische model), LDA (Fisher's lineaire discriminantanalyse), QDA (kwadratische discriminantanalyse) en KER (een kernel methode). De conclusie is dat de hersubstitutiemethode voor de evaluatie van de werking van het model een te optimistisch beeld geeft, vooral voor de kernel methode. Toepassing op onafhankelijke test steekproeven laat zien dat LOG, LDA en KER een bijna gelijk discriminerend vermogen hebben, en te prefereren zijn boven QDA. Een simulatiestudie waarin LOG, LDA, QDA en KER worden vergeleken voor een brede verzameling van gemengd continue en discrete gegevensstructuren wordt beschreven in hoofdstuk 5. De gemengde gegevens worden gegenereerd uit de vierdimensionale normale verdeling, waarbij drie van de variabelen gediscretiseerd werden. De resultaten laten een bijna identieke werking van Fisher's lineaire discriminantanalyse en logistische discriminantanalyse zien. Een keuze uit LDA en KER lijkt aangewezen voor wat betreft het discriminerend vermogen. Het is echter enigzins te verwachten dat de resultaten van deze studie afhangen van de onderliggende normale verdeling. Hierom, en ook omdat de bestudeerde gegevensstructuren mogelijk niet herkenbaar zouden zijn in feitelijke toepassingen, werd een tweede simulatiestudie verricht. In hoofdstuk 6 wordt deze tweede simulatiestudie beschreven, waarin dezelfde vier discriminanttechnieken als in hoofdstuk 5 worden vergeleken, samen met een vijfde methode, een onafhankelijkheidsmodel (IND). Er werd veel aandacht geschonken aan de opzet van deze simulatiestudie, waarin gegevens werden gegenereerd volgens een bijzonder geval van het locatiemodel. De resultaten van deze studie geven aan dat de interactie- en correlatiestructuren van de gemengde gegevens bepalen of LDA en LOG aan de ene kant of QDA en KER aan de andere kant beter werken. Een opmerkelijk goede discriminerende werking wordt bereikt door de combinatie van LDA en QDA te gebruiken, hetgeen betekent dat de "uitgebreide LDA-methode" in het algemeen een heel goede keus zal zijn. Als deze resultaten worden gecombineerd met die uit hoofdstuk 2 is het duidelijk dat als naast het discriminerend vermogen ook een goede aanpassing wenselijk is, de "uitgebreide LOG-methode" bevredigende resultaten zal geven. Onze ervaring met deze simulatiestudies resulteerde in een toepassing van logistische discriminantanalyse en de hiervoor genoemde methoden LDA, QDA, KER en IND op een gegevensverzameling uit de praktijk. De gegevens zijn afkomstig uit de farmaceutische industrie en betreffen de relaties tussen biologische activiteit en een aantal kenmerken van chemische verbindingen, o.a. betreffende hun moleculaire structuur: zie hoofdstuk 7. In deze QSAR studie (quantitative structure-activity relationship) wordt discriminantanalyse gebruikt om kenmerken op te sporen welke een significante invloed op de biologische activiteit bezitten. Op dit terrein is logistische discriminantanalyse nog relatief onbekend.

Een toepassing van het multipele logistische model voor de analyse van gematcht case-control onderzoek wordt in hoofdstuk 8 gegeven. De onderzoekpersonen betreffen patiënten met een beroerte en controlepersonen die op leeftijd en geslacht werden gematcht. Een serie multipele logistische modellen werd bestudeerd teneinde risicofactoren en mogelijke interacties op te sporen en tevens te corrigeren voor mogelijke confounding factoren. Verschillende significante risicofactoren voor een beroerte werden gevonden, waaronder hypertensie en een acuut myocard infarct. Omdat de invloed van enkele variabelen op het risico op een beroerte niet constant is voor verschillende leeftijden en voor geslacht, werden ook de betreffende interactietermen opgenomen in het conditionele logistische regressiemodel. Bijgevolg moeten schattingen van het relatieve risico van sommige factoren worden gepresenteerd binnen leeftijd- en geslacht-stata.

In hoofdstuk 9 wordt het modelleren van dosis-respons relaties in case-control onderzoek behandeld. Het standaard logistische model, uitgebreid met een kwadratische term, en een lineair model werd toegepast op zowel een empirische gegevensverzameling als op gesimuleerde gegevens. Het blijkt dat het standaard logistische model (met kwadratische term) zinvolle resultaten geeft in geval er sprake is van een lineaire dosis-respons relatie.

In hoofdstuk 10 wordt een selektiestrategie voor uitgebreide etiologisch-epidemiologische studies geïntroduceerd. Deze strategie is gebaseerd op een verdeling van de studievariabelen in groepen gelijksoortige variabelen (batterijen). Om te beginnen worden variabelen binnen iedere batterij geselecteerd, vervolgens worden deze variabelen gecombineerd tot één verzameling, en uit deze verzameling wordt de uiteindelijke selectie uitgevoerd. Deze aanpak wordt aanbevolen voor gebruik met het logistische model.

Het hoofdthema in dit proefschrift is het vaststellen van wanneer en hoe het multipele logistische regressie model toe te passen op problemen uit het uitgebreide terrein van de medische besliskunde, welk diagnostische, therapeutische en prognostische problemen bestrijkt, alsook op problemen uit de epidemiologie, vooral case-control en cohort studies. Implementatie van dit model in ruim beschikbare computerpakketten zoals BMDP laat zien dat het model zijn belangrijke rol onder de hedendaagse geavanceerde statistische technieken heeft bewezen. Ondertussen is het model nog steeds verder in ontwikkeling. Tenslotte dient te worden benadrukt dat adequate toepassing van het model inzicht vereist in zowel de klinische materie als in de statistische theorie van het aanpassen van modellen. Nauwe samenwerking tussen medici en biostatistici is daarom een vereiste voor goed wetenschappelijk onderzoek op eerder genoemde gebieden.

CURRICULUM VITAE

P. I. M. Schmitz werd geboren op 8 april 1946 te 's Gravenhage. Hij legde het eindexamen HBS-B af in 1963 aan het Thomas More college te 's Gravenhage. Van 1963 tot 1973 studeerde hij aan de Technische Hogeschool te Delft. Aanvankelijk werd de opleiding tot elektrotechnisch ingenieur gevolgd. De studie werd besloten met het wiskundig ingenieurs-examen, met als specialisatie mathematische statistiek bij prof. ir. J. W. Sieben. Vanaf december 1973 is hij werkzaam binnen het Instituut voor Biostatistica van de Erasmus Universiteit Rotterdam (hoofd: Prof. R. van Strik).