STATISTICS IN MEDICINE, VOL. 5, 503–515 (1986)

SELECTION OF VARIABLES USING 'INDEPENDENCE BAYES' IN COMPUTER-AIDED DIAGNOSIS OF UPPER GASTROINTESTINAL BLEEDING

C. OHMANN, M. KÜNNEKE, R. ZACZYK, K. THON AND W. LORENZ

Institute of Theoretical Surgery, Centre of Operative Medicine I, University of Marburg, Baldingerstraße, D-3550 Marburg, W. Germany

SUMMARY

In this paper two problems of computer-aided diagnosis with 'independence Bayes' were investigated: selection of variables and monotonicity in performance as the number of measurements is increased. Using prospective data from patients with upper gastrointestinal bleeding, the stepwise forward selection approach maximizing the apparent diagnostic accuracy was analysed with respect to different kinds of bias in estimation of the true diagnostic accuracy and to the stability of the number and type of variables selected. The results of this study suggest first that the selection of variables should be evaluated against the estimated true diagnostic accuracy obtained using all variables, and secondly that the results of a single selected sequence may be severely biased.

KEY WORDS Computer-aided diagnosis 'Independence Bayes' model Selection of variables Upper gastrointestinal haemorrhage Diagnostic accuracy Monotonicity in performance

INTRODUCTION

In the most popular statistical methods used in computer-aided diagnosis, such as Fisher's linear discriminant, logistic regression and 'independence Bayes', diagnostic probabilities are assigned to a patient on the basis of observed variables (for example symptoms, signs, test results).[1] Usually the allocation of a patient to a diagnostic category is based on these posterior probabilities. A natural criterion of a model's performance is the rate of correct classification, which is expected to be maximal when the Bayes criterion of allocation to the disease class with the highest *a posteriori* probability is used.[2]

An important problem in the use of such statistical methods is the selection of a subset of variables from a large set of possible informative variables. There are many practical and theoretical reasons for not using all variables.[3] One of the most irritating aspects of the problem is that in practice quite often the performance of a computer-aided system improves up to a point and then deteriorates as further measurements are added.[4] This indicates that increasing the dimensionality of a measurement vector may not only be useless, for example when the added measurements do not contribute at all to the classification, but may also be harmful under certain circumstances. Variable selection procedures based on 'diagnostic accuracy'[5] should be evaluated against several criteria including monotonicity or peaking in the 'diagnostic accuracy' as the number of measurements is increased.

Another striking problem in the evaluation of variable selection procedures is the estimation of the true diagnostic accuracy, especially when it is used as a criterion for selection. The 'optimistic

0277–6715/86/050503–13$06.50

Received September 1985

bias', which often occurs when the reclassification method is used, may be controlled by cross-validation or by dividing the data into a separate training and test set.[6] However, caution has to be exercised in the application of variable selection procedures, as there will be a 'selection bias' associated with the choice of the optimum from a number of possible combinations.[7]

This paper discusses selection of variables in computer-aided diagnosis as applied to prospective data from 612 patients with upper gastrointestinal bleeding using the 'independence Bayes' model with 46 variables and 4 diagnostic categories. The stepwise forward selection procedure in INDEP-SELECT[5] using the 'diagnostic accuracy' (respectively the 'error rate') as selection criterion is investigated in relation to the monotonicity or peaking problem and to different models of controlling 'optimistic' and 'selection bias'.

PATIENTS AND METHODS

The database consisted of a consecutive series of 612 upper gastrointestinal tract emergencies who were admitted to the Surgical Clinic, Marburg between January 1978 and January 1985. For every patient, a detailed history was taken and a careful clinical examination was performed. Forty-six variables from the prospective documentation were used for computer-aided diagnosis (Table I). The final diagnosis of the bleeding source was based on the findings at emergency endoscopy which was performed in every patient.[8] Four disease categories were formed: gastric ulcer, duodenal ulcer, oesophageal varices and a group containing all other possible bleeding sources.

Computer-aided diagnosis was performed with the 'independence Bayes' model, which assumes the conditional independence of the symptoms within every disease category, and uses Bayes' theorem to calculate the posterior probabilities. An *a priori* probability, $P(D)$ of 0·25 for every disease category D was chosen. The conditional probabilities $P(S|D)$ were estimated as follows:

$$P(S|D) = \frac{(\text{No. of patients with disease } D \text{ and variable } S) + 1/c}{(\text{No. of patients with disease } D) + 1},$$

where c is the number of categories of symptom S.[9] For each patient the disease D with the highest posterior probability was taken as the computer prediction.

'Diagnostic accuracy' was used to assess the performance of the computer-aided model. The term 'diagnostic accuracy' is confusing; therefore a distinction between the various measures of diagnostic accuracy used in this paper is made.[6] Ideally, one would like to know the diagnostic accuracy for future patient samples when the computer-aided model is trained (that is its parameters are estimated) on the given data set. This type of diagnostic accuracy is referred to as the *true diagnostic accuracy* of the discriminant function on new patients. However, true diagnostic accuracy cannot be obtained exactly, and thus methods of *estimating the true diagnostic accuracy* have to be used.

The use of diagnostic accuracy in connection with the development of a computer-aided model is quite different. If a given data set is used *both* to build up the model (estimate the parameters) and to assess its performance in terms of diagnostic accuracy, this accuracy is referred to as *apparent diagnostic accuracy*.[6] If selection of variables procedures are applied in computer-aided diagnosis they are part of the training of the model. Selection strategies based on the 'diagnostic accuracy' need a measure of accuracy which has to be calculated from a given set of patients. If the performance of the sequence of variables selected by the procedure is assessed with the same data, this measure is called *apparent diagnostic accuracy for variable selection.*

Table I. Forty-six variables used for computer-aided diagnosis

Anamnesis		Examination
admission from	appetite	weight
age	change of weight	height
sex	bowel habit	blood group
place of origin	bowel movements per day	—
occupation	—	general appearance
marital status	ulcer disease	mental state
season of admission	ulcer complications	skin
—	ulcer operation	abdominal examination
haematemesis,	heart disease	rectal examination
since when	liver disease	—
retching	—	pulse rate
melaena,	anticoagulants	systolic blood pressure
since when	antiphlogistics	diastolic blood pressure
nausea	analgesics	arrythmias
vomiting	date of drugs used	
regurgitation	smoking	
heartburn	alcohol	
dysphagia		
pain		
duration of symptoms		

Monotonicity

For the investigation of *monotonicity*[4] in 'diagnostic accuracy' (that is no decrease in the 'diagnostic accuracy' if the dimensionality of the measurement vector is increased), 1000 stepwise forward sequences were formed, where at each step a randomly chosen variable was included. Since investigation of all 46 variables was not feasible, this analysis was restricted to the 17 variables with the highest single apparent diagnostic accuracy calculated by the reclassification method.[6] To examine the monotonicity problem in relation to the method of measuring 'diagnostic accuracy', three different models applied to the same 1000 random sequences were investigated.[6]

Model 1 (*reclassification*)

All patients ($N = 362$, 1978–1981) were used for the estimation of the probabilities. The computer-aided model using these estimates was tested on the same patients. The apparent accuracy was used as a (bad) estimate of the true diagnostic accuracy.

Model 2 (*cross-validation* (*leaving-one-out*))

All patients except one ($N = 361$) were used for the estimation of the probabilities. The computer-aided model using these estimates was tested on the one patient left out. This was done repeatedly for all 362 patients. The percentage of patients correctly diagnosed by the computer was used as an estimate of the true diagnostic accuracy.

Model 3 (*consecutive test set*)

The data were split into two sets (consecutive). The first set ($N = 212$, 1978–1979) was used to estimate the probabilities. The computer-aided model using these estimates was tested on the

Table II. Models used for investigation of the variable selection problem. See 'Patients and methods' for description of the models.

Model		Method of measuring apparent accuracy: Reclassification	Method of measuring apparent accuracy: Cross-validation
Method of estimating true accuracy	Reclassification	1a	1b
	Consecutive test set	2a	2b
	Random test set	3a	3b

Apparent accuracy: accuracy used in developing the model, that is in the variable selection procedure.
True accuracy: accuracy on future performance of the sequence of variables selected.

second set ($N = 150$, 1980–1981). The diagnostic accuracies obtained from the test set were used as estimates of the true diagnostic accuracy.

Monotonicity of every random sequence was investigated with the Spearman rank correlation coefficient.[10] For summary descriptive statistics the median–quartile system was used.

Selection of variables

The selection of variables was performed with a stepwise forward strategy of adding a new variable to the set of already selected variables in each selection step. As the criterion for selection the apparent diagnostic accuracy (respectively error rate) was used: in the first step the variable with the maximum apparent diagnostic accuracy is selected; in the second step another variable is selected, such that this new variable together with the first one has a maximum apparent diagnostic accuracy. This process is continued until all 46 variables are included (INDEP-SELECT[5]).

Different methods of measuring the apparent accuracy used in the selection procedure and of estimating the true accuracy of the selected sequence on future performance were investigated (see Table II).

Model 1a (reclassification, no separate testing)

All patients ($N = 612$) were used for the estimation of the probabilities. The apparent diagnostic accuracy needed for the variable selection was calculated by applying the model to the same set of patients (reclassification). No separate test group was used for estimation of the true accuracy of the selected sequence.

Model 1b (cross-validation, no separate testing)

All patients except one ($N = 611$) were used for the estimation of the probabilities. The performance of a combination of variables was assessed by testing the model with this combination on the one patient left out. This was repeated for all 612 patients (cross-validation). The apparent diagnostic accuracy of the combination of variables under study was then calculated as the percentage of patients correctly diagnosed by the computer. This measure was used in the stepwise

forward selection process. Again no separate test group was used for estimation of the true diagnostic accuracy of the selected sequence.

Model 2a (reclassification with separate testing using consecutive splitting)

Patients from 1978–1981 ($N = 362$) were used for estimation of the probabilities. The apparent diagnostic accuracy needed for the variable selection was calculated by applying the model to the same set of patients (reclassification). The true accuracy of the selected sequence was then estimated by separate testing on the patients from January 1982–January 1985 ($N = 250$).

Model 2b (cross-validation with separate testing using consecutive splitting)

All of the patients from 1978–1981 ($N = 362$) except one were used for the estimation of the probabilities. The performance of a combination of variables was assessed by testing the model on the one patient left out. This was done repeatedly for all 362 patients (cross-validation) and the apparent diagnostic accuracy of the combination of variables under study was calculated as the percentage of patients correctly diagnosed by the computer. This measure was then used in the stepwise forward selection process. After the sequence of variables had been selected, *all* 362 patients were used for re-estimation of the probabilities. The true accuracy of the selected sequence was then estimated by separate testing on the patients from January 1982–January 1985 ($N = 250$).

Model 3a (reclassification with separate testing using random splitting)

The data were split randomly into two sets ($N = 612$). Similarly to model 2a, one half ($N = 306$) was used to estimate the probabilities and to perform the selection of variables with reclassification. The true accuracy of the selected sequence was then estimated by separate testing on the other half ($N = 306$).

Model 3b (cross-validation with separate testing using random splitting)

The data were split randomly into two sets ($N = 612$). Similarly to model 2b, one half ($N = 306$) was used to estimate the probabilities and to perform the selection of variables with cross-validation. The true accuracy of the selected sequence was again estimated by separate testing on the other half ($N = 306$).

All calculations were performed with computer programs written by the authors in BASIC and run on an IBM PC-AT.

RESULTS

Monotonicity

To investigate whether the 'diagnostic accuracy' of the computer-aided system generally increases or at least stays the same if one more variable is added into the model (monotonicity[4]), 1000 stepwise forward sequences were created randomly. The three models used to explore monotonicity with these sequences produced different results dependent on the method used for estimating the true diagnostic accuracy.

With model 1 (reclassification, estimated true diagnostic accuracy = apparent diagnostic accuracy) in less than 7 per cent of all steps, where a randomly chosen variable was included, a decrease in the estimated true diagnostic accuracy of more than 1 per cent was observed, compared

Table III. Correlation between the number of randomly chosen variables and the estimated true diagnostic accuracy. Distribution of the Spearman rank correlation coefficient using three different models applied to the same 1000 random stepwise forward sequences. See 'Patients and methods' for description of the models

Spearman rank correlation coefficient	Model 1 Reclassification	Model 2 Cross-validation	Model 3 Consecutive test set
$\geqslant 0{\cdot}95$	943	494	278
0·90–0·94	53	322	318
0·85–0·89	3	110	209
0·80–0·84	0	43	89
0·75–0·79	1	16	36
0·70–0·74	0	5	27
0·65–0·70	0	6	15
0·60–0·64	0	2	7
0·55–0·59	0	1	7
0·50–0·54	0	0	6
$< 0{\cdot}50$	0	1	8
Total	1000	1000	1000

to 15 per cent with model 2 (cross-validation) and 19 per cent with model 3 (consecutive test set). In 278 of the 1000 sequences no such decrease occurred with model 1 at any of the selection steps (model 2: 37 sequences, model 3: 9 sequences). A decrease of more than 2 per cent occured only in 1 per cent of all selection steps with model 1, in 5 per cent of the steps with model 2 and in 13 per cent of the steps with model 3. Two hundred and three sequences (model 1), 621 sequences (model 2) and 930 sequences (model 3) were affected by a decrease of diagnostic accuracy of more than 2 per cent in at least one of the selection steps. With models 1 and 3 the number of decreases in the estimated true diagnostic accuracy when a variable was added (for example at least one per cent), was uniformly distributed with respect to the selection steps (except in steps 1 and 16, where the 2nd and 17th variables were added, respectively). In the cross-validation approach (model 2) the number of such decreases reached its maximum in step 2 (3rd variable added) and declined monotonically with the number of variables added.

A non-parametric technique, Spearman's rank correlation coefficient, was used to measure for each of the stepwise forward sequences whether a monotonic relationship between the number of variables and the estimated true diagnostic accuracy exists.[10] Table III shows the distribution of the 1000 coefficients using the three different models. With model 1 (reclassification), the great majority of the coefficients are centred near the value 1, which indicates a perfect monotonic relationship. This could not be shown so convincingly with model 3 (consecutive test set), where about 10 per cent of the sequences had coefficients less than 0·80. Nevertheless, the majority of all Spearman rank correlation coefficients (60 per cent) exceeded the value of 0·90 with model 3. The distribution of the coefficients with model 2 (cross-validation) lay between the distributions of models 1 and 3. In Figure 1, the summary descriptive statistics of all 1000 sequences, which were calculated separately for each selection step and each model, are given. For all three models the median-curve was monotonic and flattening towards the maximum value obtained from using all variables. The area bounded by the 25th and 75th percentile curves, which were also monotonic, decreased continuously from the second to all variables in the three models. However, the variation in the estimated true diagnostic accuracy is clearly greater with models 2 and 3 compared to model 1.

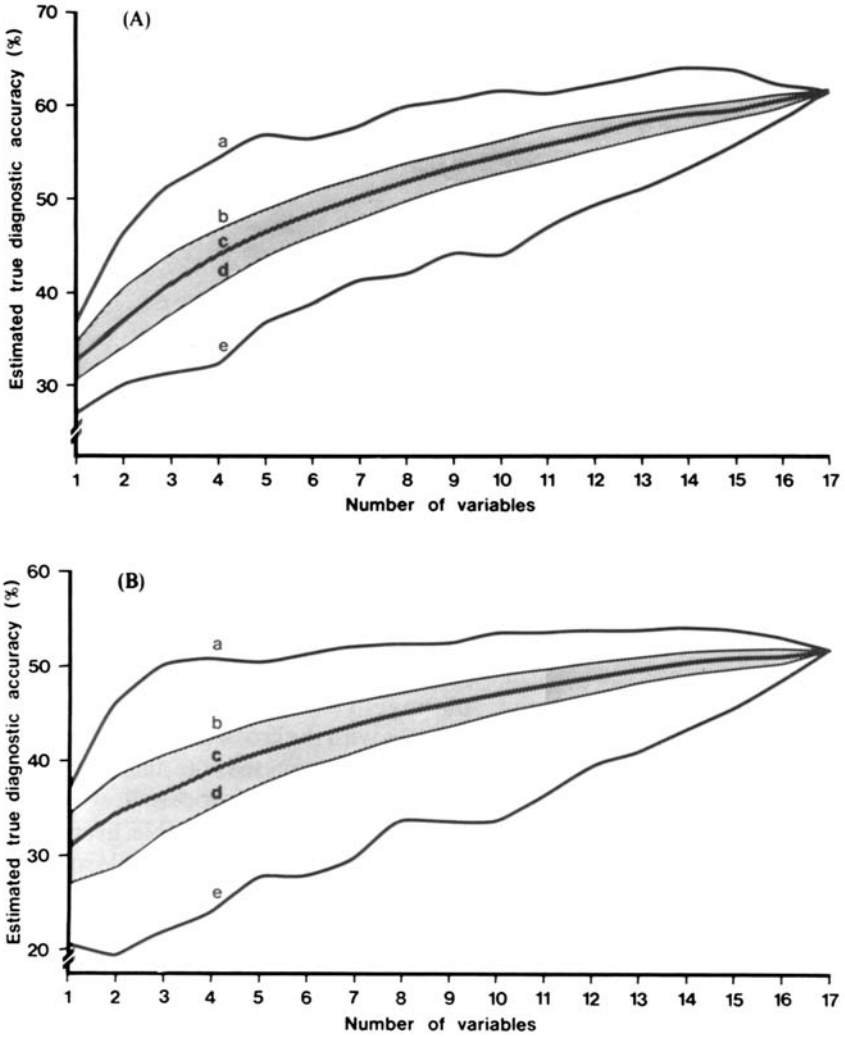

Figure 1, continued on next page

Selection of variables

Using 6 different models and all 46 variables (see 'Patients and methods'), the stepwise forward selection approach maximizing the apparent diagnostic accuracy was performed.[5] Using the criterion 'stop the selection if no increase of the apparent diagnostic accuracy occurs', the number and type of the selected variables and the final 'accuracy' were investigated. From Figure 2 it can be seen that the number of variables selected by model 1a (reclassification, $N = 612$) was double the number of variables selected by model 1b (cross-validation, $N = 612$) with a final (apparent) diagnostic accuracy of 63 per cent with reclassification and 57 per cent with cross-validation. The selected sequences agreed up to the second step (1st variable: skin, 2nd: liver disease). In addition the variables pain, melaena and bowel habit were selected by both models at different steps (cross-validation: the 3rd, 4th and 8th variables selected; reclassification: 7th, 5th and 8th).

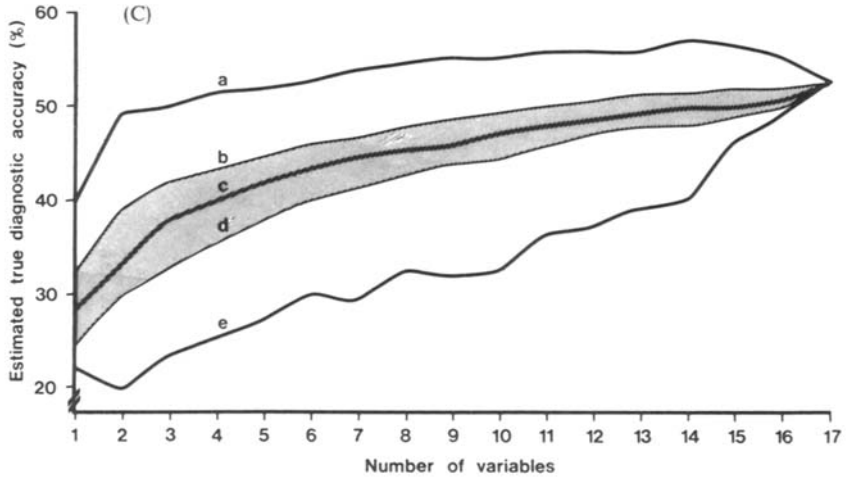

Figure 1. Summary descriptive statistics for 1000 random stepwise forward sequences. For each model (see 'Patients and methods') the same random sequences were used. For each number of variables (1, 2 . . ., 17) selected, the following descriptive statistics are presented: (a) maximum; (b) 75th percentile; (c) median; (d) 25th percentile and (e) minimum: (A) model 1 (reclassification); (B) model 2 (cross-validation); (C) model 3 (consecutive test set)

A similar pattern was observed when using models 2a and 2b (reclassification or cross-validation with separate testing using consecutive splitting). Twenty variables were selected with the reclassification method (see Figure 2) and 8 variables with the cross-validation method, resulting in apparent diagnostic accuracies of 67 and 53 per cent and estimated true diagnostic accuracies of 46 and 45 per cent, respectively. Four variables were selected by both models: liver disease, age, height and smoking. Again the sequences agreed up to the second step (1st variable: liver disease, 2nd: age). The variable height was included in the 6th step (both models) and the variable smoking in the 12th and 7th steps (reclassification, cross-validation).

Models 3a and 3b (reclassification or cross-validation with separate testing using random splitting) were applied repeatedly using different randomly split data sets for training and testing (also different between the models). Table IV shows the results of 10 selected sequences for each model. With selection based on reclassification (model 3a) the number of variables selected was again higher than with selection based on cross-validation (model 3b). Whereas the apparent diagnostic accuracy differed between the models (on the average 7 per cent), no difference could be observed in the estimated true diagnostic accuracy. Thirty-five out of 46 variables were included in at least 1 of 10 selected sequences with model 3a and 31 out of 46 variables with model 3b. Only variable 36 (liver disease) was selected in all repeated applications of both models. One variable (pain) was included in 6 sequences with model 3a and 7 sequences with model 3b. Another 6 variables were included in half of the sequences (model 3a: blood group, age, haematemesis, dysphagia and anticoagulants; both models: time of melaena). With model 3a about 50 per cent and with model 3b about 68 per cent of the selected variables were observed only in one or at most two sequences.

Stepwise forward selections maximizing the apparent diagnostic accuracy and without employing a stopping criterion are shown in Figure 2 for five different models. Except for the first selections, where model 2a (reclassification: $N = 362$, no separate testing), model 1a (reclassification: $N = 612$) and model 1b (cross-validation: $N = 612$) agreed in the apparent diagnostic accuracy, these three curves were well separated. Whereas model 1a and model 2a (no separate

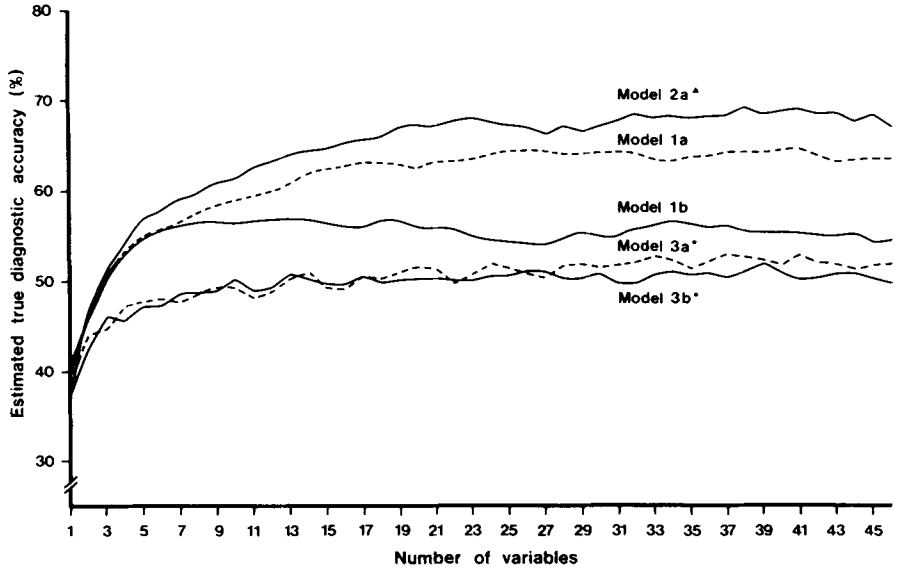

Figure 2. Selection of variables using different models for estimation of the 'diagnostic accuracy'. See 'Patients and methods' for detailed description of the models and the variable selection procedure. For models with no separate testing, the apparent diagnostic accuracy was taken as an estimate of the true diagnostic accuracy: ▲ training set $N = 362$ (1978–1981), no separate testing; * Median of 10 selected sequences based on different random splits

Table IV. Selection of variables applied to the models based on random splitting. See 'Patients and methods' for description of the models 3a and 3b

	Number of variables selected†		Apparent diagnostic accuracy (%)		Estimated true diagnostic accuracy (%)	
Sequence*	Model 3a	Model 3b	Model 3a	Model 3b	Model 3a	Model 3b
1	16	12	69	60	47	49
2	14	10	68	55	51	47
3	13	9	65	57	43	48
4	11	8	62	57	47	51
5	11	7	64	57	50	48
6	10	7	64	58	49	53
7	8	6	60	54	49	51
8	8	5	57	55	50	47
9	7	4	61	53	46	48
10	7	3	58	51	54	47
median	11	7	63	56	49	48

* 10 repeated applications of each model using different randomly split data sets.
† Stopping criterion: no increase in the apparent diagnostic accuracy.

testing) produced smooth and approximately monotonic curves, there was a slight peaking in the curve produced by model 1b. This latter model produced results approaching the models with no separate test group in the beginning of the selection process and the models with a separate test group in the middle and end of the selection process. Model 3a (reclassification with separate

testing using random splitting) and model 3b (cross-validation with separate testing using random splitting) produced rather similar results. However, these models resulted in unsmooth curves with a clear worsening of the estimated true diagnostic accuracy compared to the other models without separate testing.

DISCUSSION

The problem of monotonicity and of selection of variables has been investigated in this study using the very simple 'independence Bayes' model. This model has been extensively criticized, especially for the simplifying assumption of independence.[11] Nevertheless it is one of the most popular statistical methods in computer-aided diagnosis[1] and has often been described as a good discriminator.[2, 9] For our problem with missing data, four disease categories and a great number of probabilities to be estimated, such a simple model is of great advantage. It must be stressed that the strategy of selection and the criterion of optimality used in this paper are based on posterior probabilities, which can be obtained with many models.[5]

Concerning the criterion of optimality for selection of variables, it has been shown that, for example, the F-criterion used in the SPSS and BMDP packages is not appropriate, and can even be inversely related to diagnostic performance.[12] The criterion of 'diagnostic accuracy' used in this paper belongs to one of several criteria,[5] of real interest to the clinician, when using a computer-aided system.

There are many factors influencing the kind of problems investigated in this paper. Even when restricted to the 'independence Bayes' model and to 'diagnostic accuracy' as a measure of performance, many other factors, such as different methods for estimating probability densities, interactions between the variables and the ratio of dimensionality to the sample size,[4] influence the monotonicity and the variable selection problem. In this paper a clinical data set ($N = 612$), typical for such diagnostic problems, was used as an example to describe some of the statistical problems connected with selection and monotonicity which seem to be independent of interactions between the variables.

Stepwise forward selection maximizing the apparent diagnostic accuracy on a given data set leads to a single subset of variables. One of the main points of interest is the diagnostic accuracy of the selected subset on future samples.[6] Two kinds of bias in estimating the true diagnostic accuracy have to be considered. The first is the selection bias associated with choosing the optimal one from a large number of possible subsets.[7, 13, 14] Murray[7] described a model in which data were simulated from two populations with independent normal random variables with unit variance and different means ($+1/4$, $-1/4$) and the likelihood ratio was used as the (optimal) discriminant. This gives a monotonically increasing true diagnostic accuracy as the number of variables is increased. Different procedures for the selection of variables used in that study resulted in very optimistic estimates of the true diagnostic accuracy and showed a clearly peaking behaviour of the apparent diagnostic accuracy.

The second bias in estimating the true diagnostic accuracy is connected with the problem of having only a finite number of samples for training and testing of the computer-aided system. The dilemma is that if one is interested in having full efficiency of the model, all available data must be used for training and one cannot check whether the model is correct. If one is interested in whether the model is correct, not all available data can be used for training and thus one cannot have full efficiency.[14] The reclassification method with training and testing on the same data usually gives optimistically biased results ('optimistic bias'), whereas dividing the sample into training and test sets may result in pessimistically biased estimates of the true error rate ('pessimistic bias').[6]

Figure 2 clearly demonstrates the influence of both kinds of bias on the estimation of the true

diagnostic accuracy. In the reclassification method without separate testing, the results are mostly optimistic, owing to the additive effects of selection and optimistic bias. With increasing sample size ($N = 612$ instead of $N = 362$) the bias is reduced, but the results are still overoptimistic. The cross-validation method (leaving-one-out[5, 6]) without separate testing does not suffer from the optimistic bias, thus resulting in better estimates (Figure 2). However, the selection bias is still present, as can be seen from the nearly identical curves of both the reclassification and cross-validation methods ($N = 612$) in the first seven selection steps. The peaking behaviour of the cross-validation curve ($N = 612$) underlines the strong effect of selection bias. In stepwise forward selection procedures which result in only a few selected variables, neither the reclassification nor the cross-validation method without separate testing give accurate estimates of the diagnostic accuracy on future performance.

In the different approaches with separate testing, the selection bias is eliminated, as can be seen from Figure 2. The difference in the estimated true diagnostic accuracy between cross-validation with separate testing and reclassification with separate testing is surprisingly small, indicating that in our data set selection of variables using better estimates of the 'diagnostic accuracy' (cross-validation omitting the optimistic bias) do not lead to selected sequences with a better performance. These two models seem to approximate the true diagnostic accuracy from below with only small differences to the cross-validation model without separate testing ($N = 612$) in the second half of the selection process. To estimate the diagnostic accuracies on future performance, both estimates from below (models with separate testing) and from above (cross-validation without separate testing) should be calculated and compared.

Another important question also related to the estimation of the diagnostic accuracy is, whether a selected sequence of variables is the best on future performance.[15] Although the selected sequence is optimal with regard to the training set, it need not be optimal on new data. The reclassification method and the cross-validation method with separate testing produced unsmooth curves, indicating that a non-optimal set of variables had been selected with respect to the new data (Figure 2). One approach to solving this problem is the production not of a single subset of variables, but of different 'optimal' sets of selected variables to be tested on separate data and compared against each other. Unfortunately this introduces another kind of selection bias which has to be controlled.

The choice of the model in selection of variables is not only important for estimation of the diagnostic accuracy on future performance, but also for the number and type of selected variables. Marked differences occurred when the stepwise forward selection approach maximizing the apparent diagnostic accuracy[5] was applied using the reclassification or the cross-validation methods.[11] Repeated application of the variable selection approach using random splitting (models 3a and 3b) also produced unstable results with respect to the number, the type and the order of variables included and the estimated diagnostic accuracies (Table IV). The fact that this selection approach is very sensitive to changes in the model and changes in the stopping criterion, indicates that a single selected sequence of variables cannot be trusted. One approach to investigating this problem could be repeated application of variable selection using random splitting with an analysis of the distribution of the selected variables.

The last problem discussed in this paper is the often observed, but undesirable feature of premature termination in stepwise forward selection approaches.[16] The results of our study show that tests of monotonicity (estimated true diagnostic accuracy improves monotonically as the number of measurements is increased) depend on the model used for estimating the true accuracy (Figure 1). Monotonicity in our data could be demonstrated clearly when using reclassification without separate testing and the apparent diagnostic accuracy (Table III, Figure 1(A)). In the case of reclassification and separate testing, using the estimated true diagnostic accuracy in the test set for

the analysis of monotonicity, the results were not so impressive (Table III, Figure 1(C)). However, the median performance and the performance of the majority of the 1000 individual random sequences could be demonstrated to be (nearly) monotonic. Model 2 (cross-validation) with a more efficient use of the data, produced, as expected, results between the two extremes. Overall, the results show that in our data the estimated true diagnostic accuracy generally improves monotonically as the number of measurements is increased.

Monotonicity has been proved in the case of completely known class-conditional densities (infinite training set)[4] and in the Bayesian approach with known prior densities.[17] For completely unknown class-conditional densities, as assumed in this paper, Hughes[18] showed that in the estimative approach (maximum-likelihood estimates of the conditional probabilities $P(S|D)$) peaking occurs when an average is taken over all possible sets of random samples of fixed size and over all possible problems generated by the prior densities. If the ratio of the sample size and the number of variables is large enough this so called 'optimal measurement complexity' is not reached.[4, 18] Despite the violation of the model's underlying assumptions in our study (conditional independence), no such optimal measurement complexity could be observed with a ratio of about 20 (362 patients, 17 variables). From Figure 2 it can be assumed that even with a ratio of 14 (612 patients, 46 variables) no optimal measurement complexity exists. Therefore for similar data, a stepwise forward selected sequence[5] should be evaluated against the estimated true diagnostic accuracy obtained from all variables. Copas[14] makes similar points from a theoretical perspective. He argues that in regression analysis applied to prediction, empirical selection of variables mostly gives worse results than fitting the whole regression. In addition, the amount by which validation fit on new data falls short of retrospective fit (shrinkage) can be particularly marked when stepwise fitting is used.[14]

In this paper, a particular clinical data set was used to investigate the problems of monotonicity and of selection of variables. However, other studies and our own results in other fields of computer-aided diagnosis[19] suggest that most of the results can be extended to other sets of data. From the results of this study it is concluded that the full model using all variables should be used as a reference point for selection procedures in computer-aided diagnosis with 'independence Bayes'.[14] Furthermore, stepwise forward selection of variables maximizing the apparent diagnostic accuracy has intractable sampling properties, and thus must be treated with extreme caution, even if the selection bias and the optimistic bias are controlled by separate testing.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Madeleine Ennis for helping with the English and careful reading of the manuscript, Marlene Verfürth for typing the manuscript and Doris Weber for preparing the drawings. This work was supported by grant of Deutsche Forschungsgemeinschaft (DFG) Oh 39/2-1.

REFERENCES

1. Spiegelhalter, D. J. and Knill-Jones, R. P. 'Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology', *Journal of the Royal Statistical Society, Series A*, **147**, 35–77 (1984).
2. Croft, D. J. 'Mathematical methods in medical diagnosis', *Annals of Biomedical Engineering*, **2**, 69–89 (1974).
3. Schaafsma, W. 'Selecting variables in discriminant analysis for improving upon classical procedures', in Krishnaiah, P. R. and Kanal, L. N. (eds), *Handbook of Statistics, Vol.* 2, North-Holland Publishing Company, Amsterdam, 1982, pp. 857–881.

4. Jain, A. K. and Chandrasekaran, B. 'Dimensionality and sample size considerations in pattern recognition practice', in Krishnaiah, P. R. and Kanal, L. N. (eds), *Handbook of Statistics, Vol.* 2, North-Holland Publishing Company, Amsterdam, 1982, pp. 835–855.
5. Habbema, J. D. F. and Gelpke, G. J. 'A computer program for selection of variables in diagnostic and prognostic problems', *Computer Programs in Biomedicine*, **13**, 251–270 (1981).
6. Toussaint, G. T. and Sharpe, P. M. 'An efficient method for estimating the probability of misclassification applied to a problem in medical diagnosis', *Computers in Biology and Medicine*, **4**, 269–278 (1975).
7. Murray, G. D. 'A cautionary note on selection of variables in discriminant analysis', *Applied Statistics*, **26**, 246–250 (1977).
8 Ohmann, C., Thon, K., Stöltzing, H., Yang Qin and Lorenz, W. 'Upper gastro-intestinal bleeding: assessing the diagnostic contribution of anamnestic and clinical findings', *Medical Decision Making* (in press).
9. 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 data set of head injured patients', *Journal of the Royal Statistical Society, Series A*, **144**, 145–175 (1981).
10. Lienert, G. A. *Verteilungsfreie Methoden in der Biostatistik*, Verlag Anton Hain, Meisenheim am Glan, 1973.
11. Spiegelhalter, D. J. 'Statistical aids in clinical decision making', *The Statistician*, **31**, 19–36 (1982).
12. Habbema, J. D. F. and Hermans, J. 'Selection of variables in discriminant analysis by F-statistic and error rate', *Technometrics*, **19**, 487–493 (1977).
13. Hecker, R. and Wegener, H. 'The valuation of classification rates in stepwise discriminant analysis', *Biometrical Journal*, **20**, 713–727 (1979).
14. Copas, J. B. 'Regression, prediction and shrinkage', *Journal of the Royal Statistical Society, Series B*, **45**, 311–354 (1983).
15. Morris, J. O. 'On selecting the best set of regression predictors', *Journal of Experimental Education*, **48**, 100–103 (1979).
16. Kuk, A. Y. C. 'All subsets regression in a proportional hazards model', *Biometrika*, **71**, 587–592 (1984).
17. Menzefricke, U. 'A decision-theoretic approach to variable selection in discriminant analysis', *Communications in Statistics: Theory and Methods, Series A*, **10**, 669–686 (1981).
18. Hughes, G. F. 'On the mean accuracy of statistical pattern recognizers', *IEEE Transactions on Information Theory* **IT-14**, 55–63 (1968).
19. Ohmann, C., Lorenz, W., Ennis, M., Yang Qin, Zaczyk, R. and Schöning, B. 'Computer-aided predictions of pseudoallergic reactions to plasma substitutes', in Jesdinsky, H. J. and Trampisch, H. J. (eds), *Prognose- und Entscheidungsfindung in der Medizin*, Springer Verlag, Berlin, 1985, pp. 410–420.