COMPUTERS AND BIOMEDICAL RESEARCH 8, 173–188 (1975)

Diagnosis. II. Diagnostic Models Based on Attribute Clusters: A Proposal and Comparisons

MARIJA J. NORUSIS

University of Chicago Pritzker School of Medicine, Chicago, Illinois 60637

AND

JOHN A. JACQUEZ*

University of Michigan School of Public Health and the Medical School, Ann Arbor, Michigan 48104

Received February 20, 1974

A new discrimination procedure based on the formation of clusters of dependent attributes, and estimation of the joint probability distribution as the product of the probabilities of the disjoint clusters is proposed and investigated. The major advantages of this method are a substantial reduction of the number of probability estimates that must be made, the ability to include symptom dependencies, and the ease and flexibility of its implementation.

Comparisons with other discrimination procedures are obtained using Monte Carlo techniques. Results indicate that the proposed model is robust and may lead to gains over the independence and actuarial models, especially for small sample sizes.

INTRODUCTION

Although a variety of mathematical models for use in medical diagnosis have been proposed, few comparative evaluations have been presented. A recent investigation (*1*) considered several procedures for the estimation of the probabilities used in Bayes' theorem. When the independence, complete actuarial, optimum tree dependence, linear discriminant, and various order Bahadur models were compared in situations where parameter values were assumed known, it was found that the independence model can lead to substantial increases in misclassification rates when compared to the optimum values. Moderate symptom intercorrelations were sufficient to cause this increase.

Since Bayes' theorem, with the assumption of independence of symptoms, is frequently used in situations which violate the independence condition, alternate

* To whom inquiries should be addressed.

approaches should be considered. It would seem desirable, instead of omitting correlated variables as had been suggested (*2*), to utilize this available information in the estimation procedure. The complete actuarial model, in which each joint probability is estimated as the relative frequency of occurrence of each symptom vector, includes all symptom interdependencies. However, the need to obtain 2^{n-1} probability estimates, where n is the number of binary symptoms, requires fairly large data bases. Another disadvantage of the actuarial model is that each observation is used only for the estimation of one joint probability. Incorporation of all observations for estimation of each probability might lead to more stable estimates. Although the independence model does reduce the number of parameter estimates to n, and utilizes all observations for each estimate, the assumption of complete independence of symptoms is unduly restrictive and seldom met. It appears then that models intermediate to the actuarial and independence are needed.

The purpose of this paper is twofold: (1) to examine a new procedure for probability estimation based on the formation of clusters of symptoms, and (2) to present comparisons of this procedure with several others in situations where parameter values are assumed known, as well as in those where they must be estimated. Comparisons of the models when the data base is viewed as a sample from an underlying population allows assessment of the effect of sample size in estimation, as well as proper evaluation of the actuarial estimates since they are no longer necessarily optimum. Thus, this investigation provides extensions to the Data Base as a Universe Model considered in (*1*).

Attribute Cluster Models

Background

In many situations it is reasonable to postulate the existence of differing interrelationships between variables. For example, it would be anticipated that the association between height and weight would be quite different from that between height and eye color. In a diagnostic setting it seems natural that distinct groups of interdependent attributes may exist and that different groups occur in different diseases. Findings from the same organ system might be more closely associated with each other than with results from other systems. Thus certain observations based on renal function might be independent of those obtained from the respiratory system. Of course, numerous complex physiological interdependencies exist and must be recognized.

Meerten et al. (*3*) present correlations for symptoms which might be used in the differential diagnosis of four respiratory diseases. Their data support the hypothesis of symptom clusters, for correlations are not constant within a disease category. The substantial magnitude of the correlations indicates that independence is not a viable assumption. Prewitt (*4*) considered correlations for serum immunoglobulins.

Again, several of the correlations are quite large, with magnitudes differing between variables.

Estimation Procedure

If we allow the existence of attribute clusters, a simple method for joint probability estimation can be formulated as follows. For a given disease D let there be m mutually exclusive clusters C_i chosen so that symptoms are independent between the clusters and dependent within a cluster. The joint probability of the symptom vector x is then

$$P(x) = \prod_{i=1}^{m} P(C_i), \tag{1}$$

where $P(C_i)$ is the probability of the subset of symptoms of x found in cluster i. If all symptoms are independent, or in the present context, if each symptom forms a cluster, Eq. (1) reduces to the customary independence estimator with $m = n$.

To illustrate this procedure let us consider the simple case of 4 binary symptoms, divided equally between two clusters. Assume that symptoms 1 and 3 are in C_1, and symptoms 2 and 4 in C_2. The probability of a symptom vector $x = (1010)$, where 1 denotes the presence of a symptom, is then

$$P(x) = \prod_{i=1}^{2} P(C_i),$$

$$P(1010) = P(S_1 = 1 \text{ and } S_3 = 1) \times P(S_2 = 0 \text{ and } S_4 = 0).$$

The $p(C_i)$ estimates can be obtained directly from the data base, using

$$P(C_i) = \text{no. of persons with cluster } C_i / \text{total no. of persons.}$$

The proposed algorithm is then a combination of independence and actuarial procedures. The probabilities of each cluster configuration are actuarial estimates, the joint probability is the product of the probabilities of independent clusters.

The major advantage of the cluster procedure lies in the reduction of the number of joint probability estimates which need be obtained from a data base. For example, when the variables are dichotomous, $n = 10$, and two clusters of five symptoms each are formed, 64 ($2^5 + 2^5$) instead of 1024 (2^{10}) probabilities must be estimated. This is a substantial decrease. The actual number of estimates to be made is a function of the number and size of the clusters. However, being able to partition out even one independent attribute halves the necessary number of estimates. This reduction is of particular importance for small samples. Among the other advantages of the cluster algorithm are:

(1) inclusion of higher order symptom interactions without complex computations or estimation of many parameters;

(2) flexible implementation since different clusters can be chosen for each of the diseases under consideration;

(3) application to polychotomous as well as dichotomous data; and

(4) inclusion of observations with missing values in the estimation, as long as some of the clusters are complete.

The proposed cluster models can be derived as a special case of Bahadur's model (5). Instead of assuming that higher-order correlation parameters are zero, the assumption is made that higher-order terms factor into products of lower-order ones. Thus, higher-order correlations are included but the estimation algorithm does not entail their direct calculation.

Cluster Formation

The delineation of symptom clusters poses an interesting problem. Several alternate strategies can be considered. If there is *a priori* knowledge of underlying biological relationships that imply certain natural groupings these can be used. Otherwise, based on the collected data, clusters may be established by mathematical clustering procedures. Another possibility, one that we have explored, is to simply group the symptoms based on the magnitude of the Pearson correlation coefficients. Of course, the resulting clusters will necessarily be approximate.

Gustafson et al. (6) investigated the effectiveness of four techniques for classifying data into conditionally dependent clusters. Their research focused on subjective aggregation by humans in contrast to mathematical clustering routines. One of the salient advantages of "human clusterers" is that massive amounts of data are not needed, for groups can be formed on the basis of prior experience and knowledge. Gustafson and his associates gathered a list of 125 symptoms, physical signs, laboratory tests, and items from the history, used for thyroid diagnosis. This number was then pruned by physicians to exclude those rarely used in diagnosis. The remaining 70 attributes were given to medical students and residents who individually clustered them. Although no external criteria were available for evaluation of the accuracy of these clusters, previous studies using nonmedical data indicated that up to 91% of the variables could be identified as belonging to a cluster when the correlation was above 0.48.

Theoretical Misclassification Rates

In order to investigate the possible usefulness of the cluster algorithm, two tactics were considered. The first consisted of the examination of the theoretical (parameter values assumed known) misclassification rate, while the second required parameter estimation from samples of varying sizes. In the first situation the actuarial model is always optimum. If the clusters chosen are exact, that is, the underlying model is indeed one of groups of symptoms which are dependent within clusters and independent between clusters, the probability estimates obtained from the proposed model would be identical to those obtained from the complete actuarial procedure. Thus the resulting theoretical misclassification rates would necessarily also be

equal and optimum for both models. However, the examples we have chosen to examine are not based on the existence of perfect clusters, but instead are directed at examining the performance of our procedure in realistic situations which require its robustness. Discrimination procedures which perform poorly when parameter values are assumed known, are also unlikely to be suitable when their estimation is required. By first considering theoretical rates, it is possible to establish the best results which can be realized with use of a particular model.

The concept of partitioning attributes into independent clusters hinges on the hope that such groups do exist and can be identified. We have formed clusters based only on the magnitude of the Pearson correlation coefficients. Highly intercorrelated attributes were defined to be a cluster. Thus although higher-order correlations were present in the data only second-order correlations were used to define the clusters. Figures 1–3 illustrate the correlation patterns found in several of the disease groups.

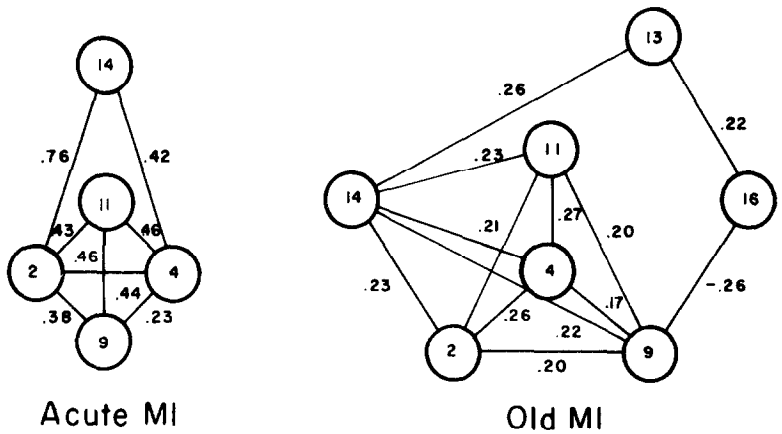

FIG. 1. Symptom correlations for acute MI and old MI. The attribute numbers correspond with those defined in Table 3 of Ref. (*1*).

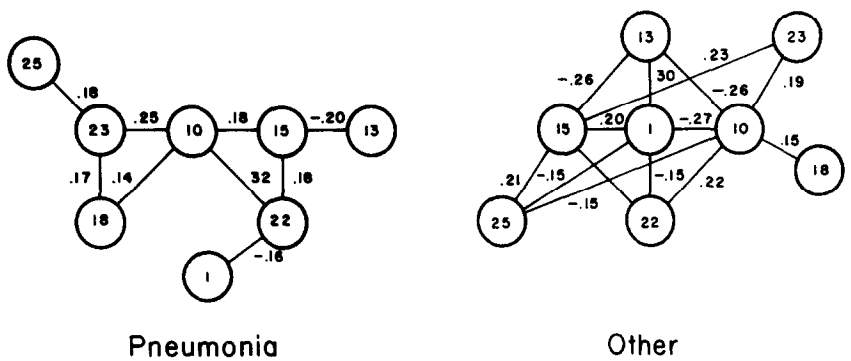

FIG. 2. Symptom correlations for pneumonia and Other. For Other, the attributes shown, except for 18, are all highly intercorrelated so not all correlations are shown.

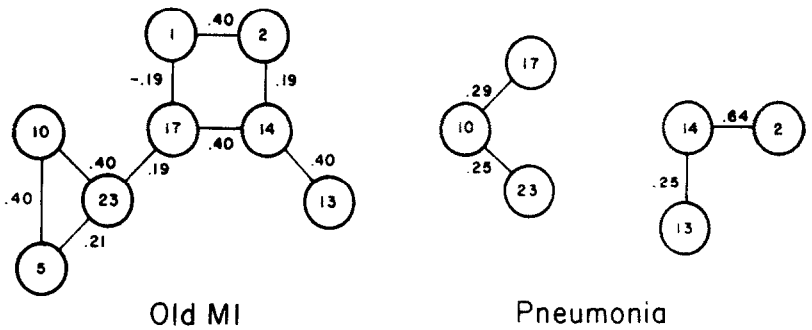

FIG. 3. Symptom correlations for old MI and pneumonia.

Only the correlations largest in magnitude have been included. (Description of the disease categories and symptoms, as well as definitions of the evaluation criteria, were reported in (*1*).)

Tables 1–3 present corresponding misclassification rates and deviations from the true, in this case actuarial, probability distribution for a variety of cluster choices. Bracketted symptom numbers indicate clusters. It should be noted that quite substantial gains over independence can be achieved with the clustering procedures. In acute MI vs old MI, the misclassification rate using the best groupings decreased from 0.192 for independence to 0.134. For pneumonia vs Other the rate went from 0.188 for independence to 0.138 for clustering and 0.123 for actuarial. When old

TABLE 1

CLUSTERING RESULTS FOR ACUTE MI VS OLD MI[a]

Disease	Clusters	Deviations		Misclassification rate
		unweighted	weighted	
Acute MI	{2,4,9,11,14}, {8,13,16}	0.0017	0.0160	0.134
Old MI	{2,4,9,11,14}, {8,13,16}	0.0031	0.0106	(0.098[b], 0.192[c])
Acute MI	{2,4,9,11,14}, {8,13,16}	0.0017	0.0160	0.136
Old MI	{2,4,11,14}, {8,9,13,16}	0.0034	0.0126	
Acute MI	{2,14}, {4,9,11}, {8}, {13}, {16}	0.0022	0.0280	0.147
Old MI	{2,4,9,11,14}, {8}, {13}, {16}	0.0033	0.0133	
Acute MI	{2,4,9,11}, {8,13,14,16}	0.0025	0.0194	0.151
Old MI	{2,4,8,11}, {9,13,14,16}	0.0037	0.0151	
Acute MI	{2,4,9,11}, {8,13,14,16}	0.0025	0.0194	0.164
Old MI	{2,4,11,14}, {8,9,13,16}	0.0034	0.0126	

[a] Symptom numbers correspond to those defined in Table 3 of (*1*).
[b] Misclassification rate for actuarial model.
[c] Misclassification rate for independence model.

TABLE 2

CLUSTERING RESULTS FOR PNEUMONIA VS OTHER[a]

Disease	Clusters	Deviations		Misclassification rate
		unweighted	weighted	
Pneumonia	{1,10,15,18,22,23}, {13}, {25}	0.0020	0.0079	0.138
Other	{1,10,13,15,22,23,25}, {18}	0.0008	0.0040	(0.123,[b] 0.188[c])
Pneumonia	{1,10,15,18,22,23}, {13}, {25}	0.0020	0.0079	0.141
Other	{1,10,13,15}, {18}, {22}, {23}, {25}	0.0020	0.0169	
Pneumonia	{10,22,23}, {1}, {13}, {15}, {18}, {25}	0.0028	0.0121	0.153
Other	{1,10,13,15,22,23,25}, {18}	0.0008	0.0040	

[a] Symptom numbers correspond to those defined in Table 3 of (*1*).
[b] Misclassification rate for actuarial model.
[c] Misclassification rate for independence model.

TABLE 3

CLUSTERING RESULTS FOR OLD MI VS PNEUMONIA[a]

Disease	Clusters	Deviations		Misclassification rate
		unweighted	weighted	
Old MI	{1,2,13,14}, {5,10,17,23}	0.0016	0.0125	0.027
Pneumonia	{1,2,13,14}, {5,10,17,23}	0.0014	0.0112	(0.022,[b] 0.046[c])
Old MI	{1,2,13,14,17}, {5,10,23}	0.0012	0.0052	0.024
Pneumonia	{1,2,13,14}, {5,10,17,23}	0.0014	0.0112	
Old MI	{1,2,13,14,17}, {5,10,23}	0.0012	0.0052	0.029
Pneumonia	{1}, {5}, {2,13,14}, {10,17,23}	0.0018	0.0156	

[a] Symptom numbers correspond to those defined in Table 3 of (*1*).
[b] Misclassification rate for actuarial model.
[c] Misclassification rate for independence model.

MI was compared to pneumonia, the rates were 0.022, 0.024, and 0.046 for actuarial, clustering, and independence, respectively. As the tables and figures indicate, "perfect" clusters are not a prerequisite for use of the cluster model. Fairly reasonable groupings including only the largest correlations give considerable improvement over independence. It is particularly interesting that use of only pairwise correlations in the grouping procedure, as well as a variety of other cluster choices led to such gains. Since the number of probability estimates to be obtained depends on the clusters chosen, it appears that cluster selection should take into consideration available sample size.

Comparisons with Other Algorithms

Introduction

Comparison of the results of the preceding section to those presented in (*1*), indicates that attribute cluster models have theoretical misclassification rates which compare quite favorably with those obtained from the independence, discriminant function, optimum tree dependence, and various order Bahadur models. The next question is then, how well can these rates be achieved in the usual setting which requires estimation from small samples. Although several procedures may have similar theoretical misclassification rates, they need not be comparable when estimation is required, for the number and complexity of the required parameter estimates may differ. For example, although the actuarial model is always theoretically optimum, the need to estimate 2^n-1 parameters from small samples seriously degrades its performance. Thus examination of a model which views the data base as a sample from an underlying population allows us to assess the effect of sample size in discrimination, as well as to include the actuarial model in all comparisons, since it is no longer necessarily optimum.

One of the outstanding difficulties inherent in any simulation is the determination of reasonable parameters. We have avoided choosing oversimplified models, such as equal marginal probabilities in a disease, absence of higher order symptom dependencies, etc. Instead we based our models on the actual data bases described in (*1*). Necessarily some clear cut results have been sacrificed. There is little doubt that if we sampled from populations which completely met the assumptions of our discrimination models, our procedures would do well. Confirmation of this result appeared less interesting than an investigation of the hazy regions into which most data presumably fall.

The disease combinations (Table 4 of (*1*)) which formed the basis for our models were (*1*) coronary artery disease and pneumonia (six symptoms), (*2*) old MI and acute MI (eight symptoms), and (3) Other and pneumonia (eight symptoms). Since all of these had substantial numbers of attribute combinations with zero probability, an undesirable characteristic for a "true" probability distribution, arbitrarily small values were substituted for zeroes. The distributions were then standardized to sum to one. This adjustment preserved almost completely the underlying correlations and probabilities.

Once a modified probability distribution was calculated, it served as the population from which data bases of varying sizes were generated. Equal prior probabilities for the two disease categories were assumed. The procedure for obtaining sample data bases was as follows:

(1) Assign a unique decimal code from one to 2^n, where n is the number of binary symptoms, to each symptom vector;

(2) store these codes in proportion to their corresponding probabilities in 20 000 memory locations, 10 000 for each disease;

(3) generate a random uniform number, multiply it by 10 000 to establish the memory location, and then select the symptom combination attached to the stored value for inclusion, and

(4) repeat (3) until the desired sample sizes and number of replicates are attained.

Simulation I. Each of the three disease combinations previously described served as a model for one of the simulations. The objective of Simulation I, based on coronary artery disease and pneumonia, was to consider the effect of incorporating attribute dependence when the two diseases are fairly separable and correlations moderate. Simulations II and III were based on less easily differentiable populations. Table 4 presents a characterization of the three models. It will be noted that the optimum actuarial and independence misclassification rates, i.e., based on true population probabilities, for Simulation I differ little.

Mean misclassification rates based on samples of size 28, 64, and 128 from each disease are displayed in Table 5. Misclassification rates were computed using

$$MR^m = \sum_{i=1}^{2^n} P^T(D_{\bar{m}})\, P^T(S_i | D_{\bar{m}}),$$

TABLE 4

MODELS FOR THE SIMULATION

A. Bahadur Goodness of Fit Ratio[a]

Simulation	Disease	Order of model							
		1	2	3	4	5	6	7	8
I	Coronary artery	0.059	0.085	0.206	0.528	0.929	1.000	—	—
	Pneumonia	0.720	0.842	0.921	0.952	0.992	1.000	—	—
II	Acute MI	0.104	0.321	0.422	0.616	0.778	0.926	0.992	1.000
	Old MI	0.252	0.436	0.616	0.779	0.923	0.978	0.999	1.000
III	Pneumonia	0.334	0.499	0.612	0.783	0.958	0.996	1.000	1.000
	Other	0.300	0.563	0.676	0.821	0.941	0.984	0.997	1.000

B. Theoretical Misclassification Probabilities

Simulation	Independence	Actuarial
I	0.066	0.054
II	0.205	0.119
III	0.207	0.140

[a] See (*1*) for description.

C. Pearson Correlation Coefficients

Simulation I

Coronary artery variable

Variable	1	2	3	4	5	6
1	1.00	0.03	0.14	0.04	0.17	0.18
2		1.00	0.23	0.08	0.11	0.25
3			1.00	0.21	0.12	0.28
4				1.00	0.12	0.12
5					1.00	0.23
6						1.00

Pneumonia variable

Variable	1	2	3	4	5	6
1	1.00	−0.11	−0.16	−0.08	−0.06	0.02
2		1.00	0.23	−0.07	−0.07	0.11
3			1.00	0.15	0.06	0.11
4				1.00	−0.03	−0.01
5					1.00	0.11
6						1.00

Simulation II

Acute MI

Variable	1	2	3	4	5	6	7	8
1	1.00	0.42	0.03	0.35	0.40	0.17	0.68	0.22
2		1.00	0.05	0.27	0.43	0.04	0.39	0.15
3			1.00	−0.05	−0.01	−0.01	0.05	0.06
4				1.00	0.42	0.11	0.27	0.09
5					1.00	0.16	0.37	0.19
6						1.00	0.26	0.11
7							1.00	0.28
8								1.00

Old MI

Variable	1	2	3	4	5	6	7	8
1	1.00	0.23	0.07	0.18	0.19	−0.09	0.23	−0.11
2		1.00	0.01	0.16	0.25	−0.13	0.22	0.10
3			1.00	0.02	0.02	0.04	−0.05	−0.04
4				1.00	0.18	−0.10	0.27	−0.25
5					1.00	0.07	0.20	0.13
6						1.00	0.26	0.22
7							1.00	0.09
8								1.00

Simulation III

Pneumonia

Variable	1	2	3	4	5	6	7	8
1	1.00	0.00	0.12	−0.10	0.04	−0.15	−0.06	0.08
2		1.00	−0.07	0.16	0.13	0.31	0.24	0.05
3			1.00	−0.24	0.04	−0.12	0.06	0.16
4				1.00	0.06	0.17	0.12	−0.16
5					1.00	−0.08	0.16	0.04
6						1.00	0.04	−0.04
7							1.00	0.16
8								1.00

Other

Variable	1	2	3	4	5	6	7	8
1	1.00	−0.024	0.29	−0.19	−0.08	0.12	−0.01	−0.10
2		1.00	−0.23	0.33	0.17	0.22	0.19	0.19
3			1.00	−0.24	−0.06	0.00	0.03	0.00
4				1.00	0.09	0.22	0.17	0.18
5					1.00	0.13	0.17	0.13
6						1.00	0.17	0.19
7							1.00	0.22
8								1.00

TABLE 5

MISCLASSIFICATION OF PROBABILITIES FOR SIMULATION I[a]

Sample size per disease	Independence	Bahadur models				Actuarial	LDF
		2	3	4	5		
28	0.072	0.120	0.125	0.156	0.162	0.136	0.078
64	0.066	0.107	0.088	0.098	0.104	0.093	0.078
128	0.063	0.073	0.074	0.077	0.074	0.070	0.074

[a] Means based on eight replicates.

where the subscript $\bar{m}$ refers to the disease having the lower posterior probability for the symptom vector S_i based on discrimination procedure m, and $P^T(S_i|D_k)$ is the population probability of the configuration S_i, given D_k. Means are based on eight replications of each experiment. That is, eight data bases with the number of cases in each disease equal to the sample size were generated. The discrimination procedures included were all six orders of the Bahadur expansion, where the first corresponds to independence and the sixth to the actuarial (using sample estimated probabilities) as well as the linear discriminant function (LDF). Besides the sample sizes of 28, 64, and 128, three additional samples of size 45, 100, and 200 were included for the actuarial versus independence comparisons. The effect of sample size is brought out more clearly in Fig. 4 in which the misclassification rates for independence, sample actuarial, and LDF models are graphed as functions of sample size.

Results indicate that for all the sample sizes best discrimination was achieved with the independence model. The difference between actuarial and independence rapidly

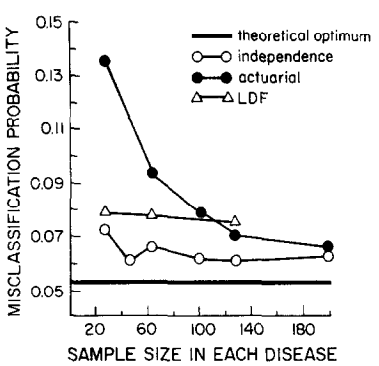

FIG. 4. Misclassification probabilities for Simulation I. Each point is the mean of eight replications of each simulation.

diminished with increasing sample size. The independence misclassification rate remained fairly constant for all sample sizes, as did the linear discriminant function rate. It is of particular interest that, for the LDF, a fourfold increase in sample size resulted in an almost negligible reduction in misclassification. The rate changed from 0.078 to 0.074. For the smallest sample size (*28*) the second- and third-order Bahadur models performed better than actuarial, but not better than independence. Increasing sample size improved results for models 2–5, but not enough to surpass independence. In summary, Simulation I suggests that for diseases which are easily separable and recognized from moderately intercorrelated attributes, independence is a tenable assumption. When the number of observations is substantially less than the number of possible symptom vectors, independence performs much better than the actuarial model. This difference decreases with increasing sample size, and for very large samples the actuarial model is always optimum.

Simulation II. The model for Simulation II, acute MI and old MI, is based on moderately separable diseases. The population misclassification rates are 0.20 and 0.12 for independence and actuarial, respectively (Table 4). The symptom correlations in acute MI range from –0.05 to 0.68, and in old MI from –0.25 to 0.27. For the simulations, samples of size 100 and 300 were considered. Besides the eight orders of the Bahadur model, two procedures based on attribute clusters were also included.

As in Simulation I, tripling of the sample size did not affect the LDF and independence methods (Table 6). The independence rate decreased from 0.219 to 0.218, the LDF increased from 0.188 to 0.189. The actuarial rate decreased from 0.197 for sample size 100 to 0.154 for sample size 300. All models were preferable to independence. For the smaller sample size the misclassification probabilities for the second cluster and the LDF models were smaller than for the actuarial model but the difference was only in the third decimal place. For 300 observations the actuarial model discriminated best. Incorporation of symptom dependencies using Bahadur's model led to better results than independence, but not better than actuarial. When 100 observations were available in each disease, the second cluster model was preferable to the first since it required fewer probability estimates. This was reversed for the larger sample size. The cluster models appeared promising, though this example did not have a broad enough range of small sample sizes to allow their adequate investigation. Overall, Simulation II indicates that the independence assumption may be detrimental in diseases which are moderately differentiable and characterized by symptom dependencies. It was of interest in comparison with Simulations I and III because the sample actuarial model did so well for a sample size of only 100.

Simulation III. The purpose of Simulation III was to provide comparisons of independence, actuarial, clustering, and LDF procedures over a wide range of sample sizes. The model for this simulation was pneumonia and Other, again moderately differentiable diseases as indicated by Table 4. The correlations were quite small, a range of 0.31– –0.24 for pneumonia, and 0.33– –0.24 for the category Other. Six

TABLE 6

MISCLASSIFICATION PROBABILITIES FOR SIMULATION II[a]

Sample size per disease	Independence	Bahadur model						Actuarial	Cluster[b]		LDF
		2	3	4	5	6	7		1	2	
100	0.219	0.208	0.212	0.202	0.211	0.205	0.210	0.197	0.206	0.189	0.188
300	0.218	0.202	0.188	0.178	0.167	0.156	0.154	0.154	0.169	0.180	0.189

[a] Means based on eight replicates.

[b] Cluster 1 was {3}, {1, 2, 4, 5, 6, 7, 8} for both diseases, cluster 2 was {1, 2, 4, 5, 7}, {3, 6, 8} for both diseases.

sample sizes were considered: 50, 100, 150, 200, 300, and 500. Not all procedures were run with each sample size since, once fairly clear trends were discerned, little would be gained by more extensive computations.

Results from this simulation (Fig. 5) are in agreement with those obtained in Simulation II. The independence model led to misclassification probabilities which were almost constant over a tenfold increase in sample size. The misclassification rate for the LDF also varied little, but a slight improvement was demonstrated from sample size 50 to sample size 500. Error rates for the actuarial model decreased very rapidly with increasing sample size. As indicated in Fig. 5, for sample size 50,

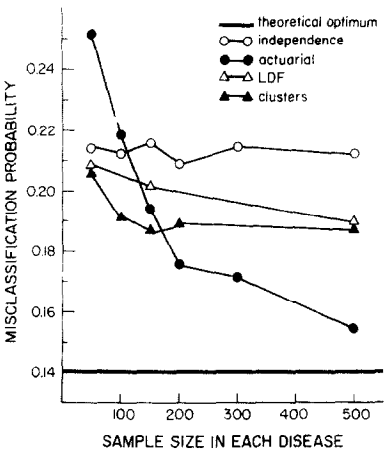

FIG. 5. Misclassification probabilities for Simulation III. The clusters used were {2,6,7}, {1}, {3}, {4}, {5}, {8} for pneumonia and {1,2,3,4}, {5}, {6}, {7}, {8}, for other.

actuarial rates were much worse than those for independence. Best results were achieved with the cluster model, although the LDF rate was very close. When a sample of size 100 was taken, the difference between actuarial and independence was slight. Best discrimination was achieved with the cluster model. By sample size 150, actuarial surpassed both independence and the LDF model, but not the clustering algorithm. This is of particular note since the clusters in this example are certainly not well delineated. For samples of size greater than 150, the actuarial model is best. However, even by size 500, the theoretical misclassification rate is not quite reached. For the clustering procedure large sample sizes did not lead to much improvement. In the present example this is to be expected, since the clusters formed were approximate. If the clusters were indeed independent, increases in sample size should yield gains even more marked than those found for the actuarial model.

Table 7 presents standard errors for the misclassification rates of Simulation III. These are sufficiently small to support the use of only eight replicates. Similar results were obtained for the other two simulations.

TABLE 7

STANDARD ERRORS FOR MISCLASSIFICATION PROBABILITIES FOR SIMULATION III[a]

Model	Sample size					
	50	100	150	200	300	500
Independence	0.0036	0.0036	0.0038	0.0050	0.0030	0.0031
Actuarial	0.0126	0.0064	0.0049	0.0048	0.0037	0.0021
Cluster	0.0054	0.0030	0.0020	0.0020	—	0.0031
LDF	0.0074	—	0.0051	—	—	0.0029

[a] Based on eight replicates.

CONCLUSIONS AND DISCUSSION

The major findings for the data bases considered are:

(1) When diseases are easily differentiable, and correlations not extremely large, the independence assumption may be acceptable. Attempts to incorporate interdependence structures in such a situation may be detrimental.

(2) The independence and LDF models are affected very little by sample size. Large increases in the number of observations do not improve discrimination.

(3) The actuarial model rapidly improves with increasing sample size, and by the time the number of observations in each disease is roughly equal to one half the number of possible symptom vectors, it performs as well as independence. When the diseases are moderately differentiable and symptom dependencies are present, the actuarial model is strongly preferable to independence for sample sizes in each disease greater than 2^{n-1}, where n is the number of binary symptoms.

(4) For very small sample sizes, the discriminant function may lead to slight gains over independence.

(5) Procedures based on clustering algorithms are very robust and may serve as suitable models for discrimination, especially when sample sizes are small.

In summary, we wish to suggest that the indiscriminate use of the independence assumption in diagnostic algorithms should be reconsidered. Both of our investigations indicate that, even in the presence of small symptom dependencies, discrimination may be hindered by inappropriate model selection. It should be emphasized, however, that for extremely small samples, even in the presence of symptom dependencies, the independence model may be a practical solution. With increasing sample size the cluster algorithm appears to be a convenient, robust strategy, especially when the sample size is large enough to allow improvements over independence, but too small to allow use of the actuarial model. Again, it should be remembered that the cluster procedure was tested in situations which did not meet the assumptions

of independent clusters. Nevertheless, gains over the other models were noted. It appears that the attribute cluster model fulfills a definite need for discrimination procedures for small samples.

ACKNOWLEDGMENT

We are grateful to Dr. H. V. Pipberger for allowing use of the data from the Veteran's Administration Cooperative Study on Automatic Cardiovascular Data Processing.

REFERENCES

1. NORUSIS, M. J., AND JACQUEZ, J. A. Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis. *Comput. Biomed. Res.* **8,** (1975), to appear.
2. NUGENT, C. The diagnosis of Cushing's disease. *In* "The Diagnostic Process" (J. A. Jacquez, Ed.), pp. 185–201. Malloy Lithographing, Ann Arbor, MI, 1964.
3. MEERTEN, R. J. VAN, DURINCK, J. R., AND DEWIT, C. Computer guided diagnosis of asthma, asthmatic bronchitis, chronic bronchitis, and emphysema. *Respiration* **28,** 293 (1971).
4. Prewitt, J. M. Experiments with statistical and quasistatistical methods in diagnosis. *In* "Computer Diagnosis and Diagnostic Methods" (J. A. Jacquez, Ed.), pp. 294–354. Charles C. Thomas, Springfield, IL, 1972.
5. BAHADUR, R. R. A representative of the joint distribution of responses to n dichotomous items. *In* "Studies in Item Analysis and Prediction" (H. Solomon, Ed.), pp. 158–168. University Press, Stanford, CA, 1961.
6. GUSTAFSON, D. H., LUDKE, R., GLACKMAN, P., LARSON, F., AND GREIST, J. H. Wisconsin computer aided medical diagnosis project—progress report. *In* "Computer Diagnosis and Diagnostic Methods" (J. A. Jacquez, Ed.), pp. 255–278. Charles C Thomas, Springfield, IL, 1972.