COMPUTERS AND BIOMEDICAL RESEARCH 8, 156–172 (1975)

Diagnosis. I. Symptom Nonindependence in Mathematical Models for Diagnosis

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

The consequences of the simplifying assumption of independence of symptoms are examined by considering a data base of cardiovascular disease patients. A mathematical model based on Bahadur's expansion (*16*) is used for quantification of nonindependence. It is shown that small symptom dependencies are sufficient to cause a substantial increase over the minimum misclassification rate. Incorporation of symptom interactions by use of Fisher's linear discriminant function, optimum tree dependence models (*21*), and Bahadur's expansion is also discussed.

INTRODUCTION

The assumption underlying the most frequently used mathematical models for diagnosis is that of independence of symptoms. Bayes' theorem, with the joint probability distribution estimated by the product of the marginal probabilities of the symptoms, has been applied to the differential diagnosis of numerous diseases (*1–2*). The inappropriateness of the independence assumption has been widely recognized, but small data samples, as well as confidence in the robustness of Bayes' theorem, has lessened the severity of most objections.

The consequences of the independence assumption in the presence of varying degrees of symptom dependence have not been considered. The apparently satisfactory results reported with use of independence have fostered a security which may not be warranted. Although independence may lead to "good" discrimination, are even better results possible? What of the situations in which poor results are obtained? May we attribute some of them to lack of independence?

Of the statistical problems outstanding in diagnosis, that of symptom interactions is one of the most serious (*3*). Some of the most useful results may come from investigations of problems related to the conditional dependence of attributes. Although clinicians are forced to incorporate symptom interdependencies into their

* To whom inquiries should be addressed.

TABLE 1

JOINT DISTRIBUTION OF TWO SYMPTOMS (S_1, S_2)

		Disease 1 S_1				Disease 2 S_1	
		Present	Absent			Present	Absent
S_2	Present	0.5	0	S_2	Present	0	0.5
	Absent	0	0.5		Absent	0.5	0

diagnostic thought processes, most mathematical models disregard, or utilize only to a very limited extent, the information present in correlated observations. Often this information is not even recognized. Table 1 presents a simple example of the importance of the joint distribution of two variables (*4*). Although both of the symptoms have equal marginal probabilities in each of the two disease states, considering them jointly permits perfect discrimination. The usual methods of analysis, based only on marginals, would have excluded these variables as unimportant. Elashoff et al. (*5*) studied the selection of 2 out of p dichotomous items in a classification problem. They found that discrimination could be increased by choosing positively correlated items and decreased by choosing negatively correlated ones. These results were not predicted from the analysis of the same problem with normally distributed variables. These two studies indicate that attempts to choose only uncorrelated symptoms may be detrimental to the overall effectiveness of a classification system.

Though most authors conscientiously draw attention to the lack of independence in their data, few have considered alternate approaches. Warner, Toronto, and Veasy (*6*) grouped mutually exclusive symptoms so that not more than one symptom in any set may be presented. Nugent (*7*), in the diagnosis of Cushing's syndrome, examined the frequency of simultaneous occurrences of signs and symptoms by calculating χ^2 for all possible pairs. The two symptoms which were significantly correlated with the others were then removed from the analysis. The drawbacks to eliminating correlated symptoms have already been considered. Brunk and Lehr (*1*) presented a method, based on Gram–Schmidt orthogonalization, which substitutes the weaker assumption of linear relationships for that of complete independence. An alternate procedure considered by Hills (*8*) focuses on the reduction of the number of attributes to be used in a diagnostic system. By selecting in a stepwise algorithm a small subset of symptoms, their joint probability can be estimated directly from the data, without using the independence assumption.

A somewhat different strategy is to apply Bayes' theorem as if independence was present, and then to adjust the posterior probabilities for the dependencies present. Such an approach was considered by Mosteller and Wallace (*9*) in their analysis of

the disputed authorship of the *Federalist Papers*. By assuming that the frequencies of words are approximately normally distributed, the mean difference between the logarithm of the odds factor under dependence and independence was obtained. For the Federalist data only moderate adjustments to the log odds were needed. These results are, however, of limited applicability since they require bivariate normality and hence consider only pairwise interactions. Lincoln and Parker (*10*) attempted to circumvent the problem of nonindependence in another way. Let $S = S_1 \cap S_2 \cap \cdots \cap S_n$, where S_i is the ith attribute and D_k is the kth disease, then

$$P(S|D_i) = P(S_1|D_i)P(S_2|S_1 \cap D_i) \cdots P(S_n|S_1 \cap S_2 \cap \cdots \cap S_{n-1} \cap D_i). \quad (1)$$

In the Lincoln–Parker approach (1) is approximated by

$$P(S_1|D_i)P(S_2|S_1 \cap D_i) \cdots P(S_{n-1} \cap D_i).$$

Having made a rather arbitrary assumption, which depends, among other things, on the arbitrary ordering of the symptoms, this procedure was applied to a data base of forty patients and ten diseases. Results from such an investigation are necessarily inconclusive. Monte Carlo comparisons of several discrimination procedures for binary variables have been presented by Gilbert (*11*), Smith (*12*), and Moore (*13*).

The Present Investigation

We have approached the problem of attribute nonindependence from several directions. The common emphasis underlying all of our strategies was examination of the problem in as realistic a milieu as possible. Whenever tenable oversimplifying models and assumptions have been avoided. The objective of our study was threefold: (1) to establish the consequences of the independence assumption when the joint-probability distribution is assumed known, (2) to evaluate several established models for joint-probability estimation which incorporate variable interactions, and (3) to propose and investigate a new model for diagnosis based on the formation of attribute clusters. The results of (3) and extensions of (1) and (2) to the case where the data is viewed as a sample from an underlying population are presented in (*14*).

The Quantification of Nonindependence

Normal theory multivariate analysis owes much of its relative theoretical simplicity to the fact that the n-variate normal distribution is completely specified by $n(n+3)/2$ parameters (n means, n variances, and $\binom{n}{2}$ covariances). To characterize an n-variate binary distribution $2^n - 1$ parameters are necessary. Thus, to allow examination of the conditions under which the independence model may result in an increased misclassification rate, it is necessary to describe the data bases in terms of $2^n - 1$ parameters. (For simplicity we have chosen to restrict ourselves to binary

distributions.) It is, of course, highly desirable to work with parameters which are easily interpretable. Though various models (*15*) are available for parametrization of multivariate binary distributions, with meaningfulness in mind, we have selected the framework of Bahadur's (*16*) expansion for quantification of nonindependence.

Bahadur's Expansion

Consider a set of n dichotomous variables, and let an observation pattern be represented by $x = (x_1, x_2, \ldots, x_n)$ where each $x_i = 0$ or $x_i = 1$, and X is the set of all points x. Let $p(x)$ be the joint probability distribution on X. A description of $p(x)$ can be obtained in terms of $2^n - 1$ independent parameters.

For each $i = 1, \ldots, n$, let $\alpha_i = p(x_i = 1)$, or equivalently,

$$\alpha_i = E_p(x_i), \qquad 0 < \alpha_i < 1,$$

where E_p denotes the expected value when $p(x)$ is the joint probability distribution. Let

$$Z_i = (x_i - \alpha_i)/(\alpha_i(1 - \alpha))^{1/2},$$

and

$$\begin{aligned} r_{ij} &= E_p(Z_iZ_j), \\ r_{ijk} &= E_p(Z_iZ_jZ_k), \\ &\cdots \\ &\cdots \\ r_{12\cdots n} &= E_p(Z_1Z_2\cdots Z_n). \end{aligned}$$

In the sequel the parameters r_{ij} will be termed second order correlations, the parameters r_{ijk} third order correlations, and so on.

Define $p'(x_1, \ldots, x_n)$ as the joint probability distribution of the x_i's when (1) the x_i's are independently distributed and (2) they have the same marginal distribution as $p(x_1, \ldots, x_n)$, that is,

$$p'(x_1, \ldots, x_n) \equiv \prod_{i=1}^{n} \alpha_i^{x_i} (1 - \alpha_i)^{1-x_i}.$$

Then Bahadur has shown that:

Theorem. For every x in X,

$$p(x) = p'(x)f(x) \tag{2}$$

where

$$f(x) = 1 + \sum_{i}^{n} \sum_{<j}^{n} r_{ij}Z_iZ_j + \sum_{i}^{n} \sum_{<j}^{n} \sum_{<k}^{n} r_{ijk}Z_iZ_jZ_k + \cdots + r_{12\cdots n}Z_1Z_2\cdots Z_n. \tag{3}$$

Among the advantages of Bahadur's expansion are the presence of the r_{ij} which are Pearson correlation coefficients, a familiar measure of association, as well as

the simplification of (2), to independence when correlation parameters of all orders are zero.

In order to ascertain the relative importance of different order correlation terms, an index suggested by Bahadur can be considered. Let $\delta^2_{[N]}$ be the squared norm of the nonconstant part of $f(x)$ in (3). That is, $\delta^2_{[N]} = \|f(x) - 1\|^2$. Then

$$\delta^2_{[N]} = \sum_{i}^{n} \sum_{j>i}^{n} r^2_{ij} + \sum_{i}^{n} \sum_{j>i}^{n} \sum_{k>j}^{n} r^2_{ijk} + \cdots + r^2_{12\ldots n} = \delta_2{}^2 + \delta_3{}^2 + \cdots + \delta_n{}^2.$$

The ratios $\delta_j{}^2/\delta_{[N]}$ provide a measure of the relative magnitude of jth order terms. If for some m the ratio

$$\left(1 + \sum_{j=2}^{m} \delta_j{}^2\right) \Big/ \left(1 + \sum_{j=2}^{n} \delta_j{}^2\right),$$

is close to 1, an approximation to $p(x)$ including only terms of order less than or equal to m, is likely to be good.

The Data Base

In order to assess the effect of the independence assumption on the misclassification probabilities in a realistic setting, we have examined nine disease pairs with varying symptom combinations. Data were obtained from the Veterans Administration Cooperative Study of Automatic Cardiovascular Data Processing (*17*). Information was available on 498 items obtained from each of 1308 patients seen at five V.A. hospitals. Cases were accepted into the study only when objective evidence

TABLE 2

DISEASE CATEGORIES INCLUDED IN THE V.A. STUDY

Disease entity	Number of cases	Criteria used for selection
1. Acute myocardial infarct	373	Diagnostic ECG, enzyme elevation
2. Old myocardial infarct with coronary insufficiency	143	Documented history
3. Angina pectoris	207	Substernal pain, precipitated or aggravated by exertion, relieved by nitroglycerine
4. Pneumonia (with or without pleural involvement)	190	Pulmonary consolidation on X-ray
5. Other diseases[a]	271	X-ray, paracentesis, etc.
6. Multiple diagnoses	124	Presence of more than one disease from categories 1–5

[a] This includes small (less than 55) numbers of patients in the following disease categories: hypertensive CV, arteriosclerotic CV, rheumatic valvular heart disease, pericarditis, pleurisy, pulmonary embolism, spontaneous pneumothorax, trauma to the chest, hiatus hernia, upper GI, gall bladder, pancreatitis, unusual types of heart disease, chest pain due to unusual causes.

for a diagnosis existed, or when the total findings allowed a definitive diagnosis. Table 2 presents the various disease groups and the criteria for selection. The patient characteristics and disease combinations included in our analyses are pre-

TABLE 3

ATTRIBUTES USED IN THE ANALYSIS

1.	Retrosternal pain—present
2.	Retrosternal pain—past
3.	Right anterior chest pain—present
4.	Constriction tightness—past
5.	Stabbing, knife-like pain—present
6.	Duration: few minutes—present
7.	Duration: few minutes—past
8.	Duration: several hours—present
9.	Duration: several hours—past
10.	Duration: days—present
11.	Radiation: left arm—past
12.	Numerous seizures daily—present
13.	Precipitated by physical exertion—present
14.	Precipitated by physical exertion—past
15.	Precipitated by respiration—present
16.	Dyspnea on exertion
17.	Cough and expectoration
18.	Hemoptysis
19.	Fatigue
20.	Lagging of either hemithorax
21.	Absence of breath sounds
22.	Moist basilar rales
23.	Signs of consolidation
24.	Temperature elevation—present
25.	Temperature elevation for more than 2 days
26.	White blood count $\geq$ 10 000
27.	Pain, lateral chest—present
28.	Constriction, tightness—present
29.	Precipitated by emotional factors—present
30.	Peripheral edema—present

sented in Tables 3 and 4. Variables were selected not because of *a priori* knowledge of correlations, but based on the magnitude of the chi-squared values. Since the primary interest was methodological, that is, what is the impact of the independence assumption for a fixed set of attributes, the optimality of the subsets chosen was not a major consideration.

The Model

The model chosen for our investigation corresponds to the Type 1 (Data Base as the Universe) situation considered by Jacquez (*18*). That is, the assumption is

TABLE 4

DISEASE COMBINATIONS AND ATTRIBUTES USED FOR ANALYSES

Diseases	Attributes[a]
1. Old MI vs acute MI	(2), 4, 8, 9, 11, 13, 14, 16
2. Angina vs acute MI	(2), 6, 7, 13, 14, 16, 19, 26
3. Old MI and angina vs acute MI	2, 4, 6, 13, (14), 16, 19, 26
4. Angina vs pneumonia	1, 5, 10, 13, 15, 17, 23, 24
5. Acute MI vs pneumonia	1, 5, 10, 12, 15, 17, 22, 24
6. Old MI vs pneumonia	1, 2, 5, 10, 13, 14, 17, 23
7. CAD[b] vs pneumonia	3, 5, 15, 17, 21, 23
8. Other[c] vs pneumonia	1, 10, 13, 15, 18, 22, 23, 25
9. CAD vs Other	2, 5, 10, 15, 20, 21, 27, 28

[a] Attribute numbers refer to Table 3. Parentheses indicate that analyses were done with and without the attributes in parentheses.

[b] Coronary artery disease refers to the combined category of acute MI, old MI, and angina.

[c] Category 5 in Table 2.

made that there exists a fixed population for which disease incidences and conditional probabilities are known and are identical with those of the data base. Though such a situation is unlikely to exist for most applications, the results obtained from a model of this nature are valuable. By viewing the data as a population, we have been able to examine a wide variety of probability and correlation structures with a relative degree of simplicity. Type 1 models should also approximate quite well situations in which the data base is large when compared to the possible number of symptom configurations and diseases. Discrimination procedures which perform poorly when parameter values are known are also unlikely to be suitable when parameters must be estimated.

In a Type 1 model the complete actuarial (full multinomial) procedure is always optimum. That is, classification should be based on the relative frequency of occurrence of the different symptom vectors. When the data base is treated as a sample from an underlying population, the actuarial model need not be the best tactic. Use of actuarial estimates obtained from a sample would tend to give poorer results than those indicated here. Misclassification probabilities would be higher for all procedures if cases from outside the sample were to be classified. This situation is considered in (*14*).

Discrimination Models

Multivariate binary distributions can be approximated by a variety of statistical models intermediate to the n parameter independence model and the $2^n - 1$ parameter actuarial model. The basic aim of such models is to obtain good probability

estimates without requiring estimation of many parameters. In diagnostic applications this is of particular importance, since data bases are usually quite small. Estimation of many parameters from few observations usually results in poor probability estimates. Procedures which incorporate all observations for estimation of each probability are also desirable since the number of observations available for each estimate increases.

Besides the independence and actuarial models we also considered the following models which incorporate symptom dependencies in differing ways.

(1) *Fisher's linear discriminant function (19).* This model incorporates second-order correlations in a multivariate normal framework. The assumption of equal variance–covariance matrices within the populations is also required. Although binary data do not meet these requirements, the robustness of the linear discriminant function (LDF) to deviations from both normality and equality of covariance matrices, as well as its frequent application to classification problems, makes it a plausible model. (The quadratic discriminant function which does not assume equal covariance matrices can also be considered.)

Frequently it is not enough to just classify an individual; an estimate of risk, or the probability of belonging to a disease state, is required as well. Suppose for n variables $x_1, \ldots, x_n$ and two populations with prior probabilities p, and $1-p$, the multivariate distributions are given by $f_1(x)$ and $f_2(x)$. The probability that an individual with symptom vector x falls in disease category D_1, by Bayes' rule is

$$P(D_1|x) = \frac{P(x|D_1)P(D_1)}{\sum_{k=1}^{2} P(x|D_k)P(D_k)}$$

$$= \frac{1}{1 + ((1-p)/p)(f_1(x)/f_2(x))}.$$

If f_1 and f_2 are multivariate normal with equal covariance matrices,

$$\frac{(1-p)f_1(x)}{pf_2(x)} = e^{-\left(\alpha + \sum_{i=1}^{n} \beta_i x_i\right)}$$

and

$$P(D_1|x) = \left(1 + e^{-\alpha - \sum_{i=1}^{n} \beta_i x_i}\right)^{-1}, \tag{4}$$

where the β_i are the coefficients of the linear discriminant function. Equation (4) is the multivariate logistic function. Though multivariate normality is a sufficient condition for the multivariate logistic function to hold, Truett, Cornfield, and Kannel (*20*) point out that a much weaker condition, namely, that the linear compound

$$\hat{Y} = \hat{\alpha} + \sum_{i=1}^{n} \hat{\beta}_i x_i$$

be univariate normal, is also sufficient. Thus, Fisher's LDF can be used with Bayes' theorem to obtain estimates of posterior probabilities for belonging to different populations.

(2) *Optimum tree dependence models.* This procedure (*21*) is concerned with best approximating an n-variate distribution P by a product of $n-1$ two-variate components. The "best" approximation P' is defined to be that which minimizes the closeness measure

$$\begin{aligned} I_{P-P'} &= \sum P_i \log P_i - \sum P_i \log P_i \\ &= I_P - (I_a + I_b + \cdots + I_n), \end{aligned}$$

where $P' = P_a \cdot P_b \cdots P_n$, and I_a is the information of the ath component distribution. Thus, it is necessary to maximize the information in each of the component distributions.

Consider a probability distribution

$$P(x) = \prod_{i=1}^{n} P(x_{m_i} | x_{m_{j(i)}}), \qquad 0 \leqslant j(i) \leqslant n,$$

where $(m_1, \ldots, m_n)$ is an unknown permutation of the integers 1–n, and $P(x_i|x_0)$ is defined to be $P(x_i)$. Define the mutual information of two variables x_i and x_j by

$$I(x_i, x_j) = \sum_{x_i} \sum_{x_j} P(x_i, x_j) \log \left[\frac{P(x_i, x_j)}{P(x_i) P(x_j)} \right].$$

This will be the weight assigned to the "branch" connecting variables x_i and x_j. The problem then becomes that of constructing an approximating distribution (tree) which has maximum branch weight. Since branch weights are additive, the maximum weight branch tree can be constructed branch by branch. By calculating $I(x_i, x_j)$ for all $\binom{n}{2}$ pairs, and then ranking each branch, the tree is formed by first choosing the largest branch and then adding in order successive branches which contain at least one new variable.

The optimum tree dependence model provides an alternate approach for including pairwise variable interactions in probability estimation. No assumption on the distribution of the x_i are needed. The appeal of this procedure is in its ability to utilize many large symptom dependencies. The major drawbacks are the inclusion of only pairwise interrelationships as well as the selective inclusion of only the largest.

(3) *Models based on Bahadur's expansion.* Besides providing a model for the examination of attribute nonindependence, Bahadur's expansion easily lends itself to approximation of joint probability distributions. By considering only correlations up to some order m, where m is less than the number of symptoms, various approximations to the probability distribution can be calculated. The disadvantages of this procedure are its limitations to binary data, as well as the possibility of negative probability estimates. Its major advantage is the inclusion of interactions of order greater than two in a noniterative algorithm.

Evaluation Criteria and Procedures

For evaluation and comparison of the approaches, three criteria were used.

1. Mean absolute deviation,

$$MD_k{}^m = \sum_{i=1}^{2^n} \frac{|P^m(S_i|D_k) - P^T(S_i|D_k)}{2^n},$$

where $P^m(S_i|D_k)$ is the probability estimate of symptom vector S_i given D_k, based on model m, and $P^T(S_i|D_k)$ is the true probability of the configuration S_i.

2. Weighted absolute deviation,

$$WD_k{}^m = \sum_{i=1}^{2^n} |P^m(S_i|D_k) - P^T(S_i|D_k)|P^T(S_i|D_k).$$

3. Misclassification rate,

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

where the subscript $\bar{m}$ refers to the disease having the lower posterior probability for S_i based on model m, and $P^T(D_{\bar{m}})$ is the prior probability for $D_{\bar{m}}$.

Since the problem under consideration involves both estimation and classification, the evaluation measures were also of two types. It is not necessarily true that a method which provides better probability estimates will also be better in classification. Indices one and two measure the deviations of the estimated probabilities from the actuarial probabilities, which in the present model are assumed true. Both weighted and unweighted deviations were included, since it is desirable that procedures give good probability estimates not only for the frequently occurring vectors but also for the less likely combinations.

For classification purposes the following procedure was followed. First probability estimates for $P(S_i|D_k)$ were obtained from the previously described models. Then, using Bayes' rule and assuming equal priors, posterior probabilities for belonging to each of the two disease categories were calculated. Assignment was made to the population with the higher posterior. At times small negative probability estimates may be obtained with truncated Bahadur models. Two strategies were considered: (1) negative probability estimates were replaced by zero while the non-negative ones were unaltered, (2) negative probability estimates were again replaced by zero, but the rest of the probabilities were restandardized to add to one. Both approaches led to very similar results. All values presented here are based on the first approach.

RESULTS

In order to characterize our data base, Tables 5 and 6, respectively, contain the distributions of the second-order correlation coefficients and the Bahadur index for the relative magnitude of various order correlations. Examination of Table 5

TABLE 5

DISTRIBUTION OF SECOND ORDER COEFFICIENTS

Disease	S^a	Proportion of $\lvert r_{ij}\rvert$ <0.33	0.33–0.67	>0.67	Disease	S^a	Proportion of $\lvert r_{ij}\rvert$ <0.33	0.33–0.67	>0.67
Acute MI	1	0.71	0.25	0.04	Angina	4	0.96	0.04	0.00
Old MI		1.00	0.00	0.00	Pneumonia		1.00	0.00	0.00
Acute MI	1[b]	0.81	0.19	0.00	Acute MI	5	0.96	0.04	0.00
Old MI		1.00	0.00	0.00	Pneumonia		1.00	0.00	0.00
Old MI and angina	3	0.86	0.14	0.00	CAD	9	1.00	0.00	0.00
Acute MI		0.86	0.11	0.04	Other		0.79	0.21	0.00
Old MI and angina	3[b]	0.90	0.10	0.00	Old MI	6	0.89	0.11	0.00
Acute MI		0.90	0.10	0.00	Pneumonia		0.96	0.04	0.00
Angina	2	0.79	0.21	0.00	Other	8	0.96	0.04	0.00
Acute MI		0.86	0.11	0.04	Pneumonia		1.00	0.00	0.00
Angina	2[b]	0.81	0.19	0.00	CAD	7	1.00	0.00	0.00
Acute MI		0.90	0.10	0.00	Pneumonia		1.00	0.00	0.00

[a] Refers to attribute combinations in Table 4.
[b] Table 4 attributes without parenthesized attribute.

TABLE 6

GOODNESS OF FIT RATIOS FOR SELECTED DISEASES

$$\left(1+\sum_{j=2}^{m}\delta_j^2\right)\Big/\left(1+\sum_{j=2}^{n}\delta_j^2\right)$$

Disease	S^a	Order of the Bahadur approximation $m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
Acute MI	1	0.0432	0.1465	0.2422	0.4491	0.6642	0.8862	0.9881	1.000
Old MI	1	0.2195	0.4072	0.5932	0.7604	0.9204	0.9782	0.9995	1.000
CAD	7	0.1682	0.2045	0.4241	0.9786	0.9998	1.000	—[b]	-
Pneumonia	7	0.7072	0.8271	0.9133	0.9471	0.9913	1.000	—	-
Old MI	6	0.0015	0.0031	0.0117	0.0621	0.2309	0.5501	0.8698	1.000
Pneumonia	6	0.1062	0.1886	0.3228	0.4664	0.7037	0.9193	0.9957	1.000
CAD	9	0.0024	0.0042	0.0136	0.0658	0.2987	0.7573	0.9981	1.000
Other	9	0.1182	0.3707	0.4836	0.7689	0.8941	0.9661	0.9943	1.000
Pneumonia	8	0.2540	0.3800	0.5225	0.7382	0.9455	0.9969	0.9998	1.000
Other	8	0.2073	0.3947	0.5224	0.7138	0.9009	0.9763	0.9983	1.000

[a] Number refers to the disease category attributes in Table 4.
[b] Indicates that fewer than eight attributes were used.

indicates that most of the symptom pairs are moderately correlated. Few correlations exceed 0.33. Based on the Bahadur indices of Table 6, it can be seen that the contributions of correlation coefficients of order greater than two are not negligible and should probably be included for probability estimation.

Misclassification rates based on probability estimates obtained from independence, Bahadur, actuarial, and linear discriminant function models are presented in Table 7. It should be noted that the Bahadur model of order one corresponds to independence while the nth order model, where n is the number of symptoms, corre-

TABLE 7

MISCLASSIFICATION PROBABILITIES

Diseases	Order of the Bahadur approximation								Discriminant function	% Increase independence to actuarial
	1	2	3	4	5	6	7	8		
Old MI vs acute MI	0.192	0.169	0.154	0.124	0.105	0.098	0.098	0.098	0.156	95
Old MI vs acute MI (without 2)[a]	0.194	0.155	0.138	0.132	0.117	0.117	0.117	—[b]	0.150	65
Old MI and angina vs acute MI	0.210	0.218	0.210	0.175	0.158	0.157	0.157	0.157	0.204	34
Old MI and angina vs acute MI (without 14)[a]	0.212	0.206	0.195	0.181	0.177	0.176	0.175	—[b]	0.216	21
Angina vs acute MI	0.198	0.194	0.179	0.158	0.146	0.145	0.145	0.145	0.174	37
Angina vs acute MI (without 2)[a]	0.182	0.184	0.174	0.157	0.154	0.154	0.154	—[b]	0.173	18
Angina vs pneumonia	0.040	0.040	0.040	0.022	0.022	0.021	0.021	0.021	0.051	86
Acute MI vs pneumonia	0.057	0.071	0.060	0.038	0.024	0.021	0.021	0.021	0.063	175
CAD vs other	0.258	0.272	0.265	0.247	0.242	0.242	0.242	0.242	0.263	7
Old MI vs pneumonia	0.046	0.036	0.025	0.034	0.025	0.024	0.022	0.022	0.045	116
Other vs pneumonia	0.188	0.167	0.146	0.126	0.123	0.123	0.123	0.123	0.167	53
CAD vs pneumonia	0.055	0.053	0.047	0.047	0.046	0.046	—[b]	—	0.061	20

[a] Refers to attributes in Table 3. These were found to be highly correlated with other symptoms.
[b] Fewer than eight attributes were used.

sponds to the complete actuarial (full multinomial) model. Comparison of misclassification rates from Table 7 suggests that the independence assumption can seriously degrade results obtained from Bayes' rule. The largest percent increase in misclassification rates when independence was substituted for the actuarial model occurred for the acute MI vs pneumonia categories. The rate rose from 0.021 to 0.057, an increase of 175%. However, since these misclassification rates are quite small, the difference might not be as important as that found for old MI vs acute MI, a

shift from 0.098 to 0.192, or for Other vs pneumonia, an increase from 0.123 to 0.188. The smallest increase was 7% for CAD vs other, a combination which was poorly distinguishable at outset. For the categories considered in Table 7, there was roughly an average increase of 60%. It is important to recall that the symptom combinations chosen were not highly intercorrelated. Deliberate attempts to violate the independence assumption could, presumably, uncover even more striking illustrations.

When only pairwise interactions are included in probability estimation, results are not uniformly better than independence. In three of the twelve disease combinations, use of a second-order Bahadur approximation led to results worse than independence, although only in one category, acute MI vs pneumonia, is this difference large. When optimum dependence trees were constructed, little gain over independence was noted. Table 8 presents misclassification rates and the measure of

TABLE 8

MISCLASSIFICATION RATES FOR TREE DEPENDENCE MODELS

Diseases	$I_{p-p'}$		Misclassification rate	
	Tree	Independent	Tree	Independent
Acute MI vs old MI and angina	0.276	0.706	0.205	0.210
	0.404	0.775		
Old MI vs pneumonia	0.416	0.673	0.057	0.046
	0.356	0.533		
Old MI vs acute MI	0.393	0.923	0.182	0.192
	0.883	1.065		
Acute MI vs pneumonia	0.240	0.430	0.050	0.057
	0.632	0.729		
Angina vs pneumonia	0.113	0.181	0.036	0.040
	0.480	0.595		
Pneumonia vs Other	0.583	0.710	0.185	0.188
	0.403	0.593		

closeness, $I_{p-p'}$, for tree models as well as independence. Although the former approximated the probability distributions better than independence, a corresponding decrease in misclassification rates was not obtained. The proportion misclassified was almost identical for both models. For the discriminant function there was an average increase of 58% over actuarial. There was no discernible pattern for its performance. In one-third of the cases it was somewhat preferable to independence, in the remaining it was about equal or somewhat worse.

In eleven of the twelve combinations when Bahadur approximations of order greater than two were considered, monotonically decreasing misclassification rates were noted. In old MI vs pneumonia the rate rose for the fourth order, although it

was still better than independence. By fourth order, the results based on the expansion were definitely superior to independence and close to actuarial values.

Table 9 contains weighted absolute deviations for the $n - 1$ orders of the Bahadur model. It can be observed that the deviations decrease almost monotonically with increasing order of the model. By the fourth order approximation the deviations have been markedly reduced. Similar patterns were found for the unweighted deviations.

TABLE 9

WEIGHTED DEVIATIONS FOR SELECTED DISEASES

Disease	S^a	Order of the Bahadur approximation						
		1	2	3	4	5	6	7
Acute MI	1	0.0438	0.0178	0.0172	0.0077	0.0033	0.0008	0.0002
Old MI	1	0.0168	0.0146	0.0091	0.0061	0.0025	0.0009	0.0001
Acute MI	2	0.0405	0.0065	0.0102	0.0022	0.0009	0.0002	0.0000
Angina	2	0.0170	0.0099	0.0082	0.0041	0.0016	0.0005	0.0001
Acute MI	3	0.0400	0.0075	0.0121	0.0022	0.0012	0.0003	0.0001
Old MI and angina	3	0.0096	0.0050	0.0038	0.0021	0.0011	0.0004	0.0001
Angina	4	0.0288	0.0052	0.0018	0.0006	0.0000	0.0	0.0
Pneumonia	4	0.0156	0.0072	0.0061	0.0022	0.0016	0.0002	0.0001
Old MI	6	0.0397	0.0129	0.0223	0.0187	0.0129	0.0068	0.0014
Pneumonia	6	0.0276	0.0106	0.0033	0.0024	0.0020	0.0005	0.0000

[a] Number refers to the disease category attributes in Table 4.

DISCUSSION

The results from the first phase of our investigation support the conjecture that the independence assumption can substantially decrease the effectiveness of a Bayesian classification system. Usually this is not recognized, for true misclassification probabilities are not known and a basis for comparison is unavailable. A probability of correct classification in the range of 80% is often acceptable to many investigators. The gain which may still be realized with use of other models is obscured by the relatively satisfactory results obtained with independence.

Examination of models which include only pairwise dependencies (LDF, second-order Bahadur, and optimum dependence trees) suggested that higher-order interactions should be incorporated into estimation procedures for all three second-order models led to unpredictable results when compared to independence. When the order of the Bahadur approximation was equal to or greater than half the number

of attributes, results were definitely superior to independence and usually quite close to optimum actuarial rates.

It is of interest to note that the effect of the independence assumption was not constant over the disease categories, and it did not reflect the differences in correlation. We would like to postulate a mechanism which might aid in explaining these results. The basic requirement for discrimination to be possible is that corresponding symptom vectors in the populations to be differentiated have unequal probabilities. When diseases have similar joint probability distributions, discrimination is necessarily poor. Easily separable diseases are often characterized by (high, low) or (low, high) probability pairs. That is, a vector which occurs with high probability in one of the diseases, occurs with low probability in the other. For example, let us consider the data of Collen et al. (*22*) who examined the response of 230 patients with a clinical diagnosis of bronchial asthma, and 517 asthma-free patients, to six dichotomous questions. Of the nonasthmatics, 68% reported having none of the six symptoms, while only 6.9% of the asthmatics evidenced no symptoms. Fourteen percent of the asthmatics had all the symptoms, while none of the nonasthmatics had all of the symptoms. Assuming that the only information available for classification is whether an individual has none, all, or between one and five of the symptoms, the symptom–disease matrix in Table 10 could be constructed. Assuming equal priors

TABLE 10

ABRIDGED SYMPTOM DISEASE MATRIX FOR DATA OF COLLEN ET AL. (*22*)

Number of symptoms	Asthma	Nonasthma
0	0.07	0.68
1–5	0.79	0.32
6	0.14	0.00

for both populations, the optimum misclassification probability is 0.195. Thus, over 80% of all patients could be correctly classified on the basis of only one decision rule: classify as nonasthmatic unless one or more symptoms are present.

Let us briefly consider the effect of attribute nonindependence in such a situation. In order for the classification rules to be altered an extremely large change in the joint probability distribution is needed. Under independence the P (no symptoms/nonasthma) is estimated as 0.576, while the probability estimate for no symptoms given asthma is 0.0055. Thus, the (high, low) ranking for nonasthmas and asthma is maintained. Preservation of this ordering is the only prerequisite for the classification rule to remain unchanged. Grossly inaccurate probability estimates may leave the misclassification rate unchanged. In diseases which have most of the probability "concentrated" in a few symptom vectors, the range in which probability estimates

may fall without altering misclassification rates is quite broad. Thus in easily differentiable diseases we would expect departures from independence to result in a rather small absolute increase in misclassification rate.

Our data confirm this hypothesis. All four of the easily discriminable disease categories (numbers 4–7 in Table 7) had an absolute increase of only 1–3% when independence and actuarial misclassification rates were compared. Examination of the corresponding correlation matrices indicated that these diseases were not characterized by a better fit to the independence model than the other combinations considered. Inspection of the joint probability distribution supported the conjecture that there would be several very discriminating vectors. For example, when CAD and pneumonia were considered, it was found that 72% of all CAD patients exhibited none of the symptoms, while this was true for only one percent of the pneumonia patients. Similar symptom vector pairs with large differences in probability were also found for the other easily differentiable categories.

For moderately differentiable diseases (misclassification rates in the range of 10–20%), the large differences between vector probabilities which characterize easily distinguishable diseases are diminished. Thus moderate perturbations of probability estimates may result in changes of the classification rule, which leads to increases in the misclassification rate. Again, Table 7 indicates that fairly large absolute increases in misclassification rate have been noted for the moderately separable categories. When diseases are poorly differentiable at outset, the problem is not as interesting, for it is only selection of different attributes that can improve discriminability. If the differences between the probabilities for a given vector are slight, reversal of the classification rule would not be expected to have much effect on the overall misclassification rate.

It is important to realize that good discrimination does not require that large differences in the joint probability distributions be present. Theoretically, it is possible to obtain good discrimination even when the absolute differences in the joint probability distributions are small. For example one disease may have one-half of all possible symptom vectors nonzero, while the other has these vectors with zero probability. Thus no large probability vectors need be present in either. However, whether this is a situation which is often encountered in realistic settings remains to be seen.

In summary, we have found that the independence assumption, even when symptom correlations are small, can lead to an increase in misclassification rates when compared to the optimum rates. In the present case where parameter values are assumed known, the actuarial model is optimum. For easily differentiable diseases the impact of the independence assumption may be less severe than for moderately separable diseases. Results reported in the literature usually focus on diseases for which good classification results have been obtained, and thus could not recognize how deleterious the independence assumption can be for the more challenging situations.

ACKNOWLEDGMENT

The authors thank Dr. H. V. Pipberger for use of the data from the Veterans Administration Cooperative Study on Automatic Cardiovascular Data Processing.

REFERENCES

1. LUSTED, L. B. "Introduction to Medical Decision Making." Charles C Thomas, Springfield, IL, 1968.
2. ROSS, P. Computers in medical diagnosis. *CRC Crit. Rev. in Rad. Sci.* **3,** 197 (1972).
3. ANDERSON, J. A., AND BOYLE, J. A. Computer diagnosis: Statistical aspects. *Brit. Med. Bull.* **24,** 230 (1968).
4. MEEHL, P. E. Configural scoring. *J. of Cons. Psych.* **14,** 165 (1950).
5. ELASHOFF, J. D., ELASHOFF, R. M., AND GOLDMAN, G. E. On the choice of variables in classification problems with dichotomous variables. *Biometrika* **54,** 668 (1967).
6. WARNER, H. R., TORONTO, A. F., AND VEASY, L. G. Experience with Bayes' theorem for computer diagnosis of congenital heart disease. *Ann. N.Y. Acad. Sci.* **115,** 558 (1964).
7. NUGENT, C. The diagnosis of Cushing's disease. *In* "The Diagnostic Process" (J. A. Jacquez, Ed.), pp. 185–209. Malloy Lithographing, Ann Arbor, MI, 1964.
8. HILLS, M. Discrimination and allocation with discrete data. *Appl. Stat.* **16,** 237 (1967).
9. MOSTELLER, F., AND WALLACE, D. F. "Inference and Disputed Authorship: The Federalist." Addison-Wesley, Reading, MA, 1964.
10. LINCOLN, T. F., AND PARKER, R. D. Medical diagnosis using Bayes' theorem. *Health Services Res.* **2,** 34 (1967).
11. GILBERT, E. S. On discrimination using qualitative variables. *J. Amer. Stat. Assoc.* **63,** 1399 (1968).
12. SMITH, J. A. "Statistical Procedures for Diagnosis Based on Binary Variables," Technical Report 33, Dept. of Statistics, Stanford University, Stanford, CA, 1972.
13. MOORE, D. H. Evaluation of five discrimination procedures for binary variables. *J. Amer. Stat. Assoc.* **68,** 339 (1973).
14. NORUSIS, M. J., AND JACQUEZ, J. A. Diagnosis. II. Diagnostic models based on attribute clusters: A proposal and comparisons. *Comput. Biomed. Res.*, **8** (1975), to appear.
15. COX, D. R. The analysis of multivariate binary data. *Appl. Stat.* **21,** 113 (1972).
16. BAHADUR, R. R. A representation 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.
17. PIPBERGER, H. V., KLINGEMAN, J. D., AND COSMA, J. Computer evaluation of statistical properties of clinical information in the differential diagnosis of chest pain. *Meth. Inf. Med.* **7,** 79 (1968).
18. JACQUEZ, J. A. Algorithmic diagnosis: a review with emphasis on Bayesian methods. *In* "Computer Diagnosis and Diagnostic Methods" (J. A. Jacquez, Ed.), pp. 374–393. Charles C Thomas, Springfield, IL, 1972.
19. ANDERSON, T. W. "An Introduction to Multivariate Statistical Analysis." Wiley, New York. 1958.
20. TRUETT, J., CORNFIELD, J., AND KANNEL, W. B. A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.* **20,** 511 (1967).
21. CHOW, C. K., AND LUI, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory* **IT 14,** 462 (1968).
22. COLLEN, M. F., RUBIN, L., NEYMAN, J., DANTZIG, G. B., BAER, R. M. AND SIEGELAUB, A. B. Automated multiphasic screening and diagnosis. *Amer. J. Pub. Health* **54,** 741 (1964).