International Journal of Computer and Information Sciences, Vol. 3, No. 4, 1974

Computer Diagnosis of Goiters. The Optimal Size of Optimal Subsymptomatologies

A. Bouckaert[1,2]

Received July 1973; revised March 1974

Optimization for diagnostic recognition rate was performed for subsets of symptoms of various sizes. The diagnostic problem was the recognition and identification of thyroid diseases. Unbiased evaluation of performance was obtained and the extent of the bias in other evaluation methods was determined. Interdependence of symptoms was shown to be a negligible nuisance in the application of Bayesian inference to the present data. An optimal size of optimized subsets of symptoms was observed. A comparison with sequential diagnosis shows that the two procedures are different, although they are related, and that the optimality of subsets is sensitive to departures from their composition.

1. INTRODUCTION

In a previous part of this study[(3)] we had suggested that restricted subsets of symptoms could be used for the diagnostic screening of goiters. Our impression at the time was that it would be easier to achieve a good diagnostic accuracy by direct application of Bayes' rule of inference to subsymptomatologies rather than by the application of an empirical Bayesian rule—assuming the mutual independence of symptoms—to the whole symptomatology available.

In order to check if this impression was correct, diagnostic efficiencies will be determined in the present study for a subset of the same data, taking into account: (a) the size of the subsymptomatologies; (b) the symptom

[1] Laboratory of Physiopathology, Faculty of Medicine, Kinshasa, Zaire.
[2] Present address: Faculty of Medicine, Catholic University of Louvain, Brussels, Belgium.

composition of the subsymptomatologies; (c) the endorsement or the rejection of the hypothesis of symptom mutual independence; (d) the method of evaluation of the diagnostic efficiency.

2. MATERIAL AND METHODS

2.1. Data

The sample of 86 goitrous patients used in the previous parts of this study was retained for the present experiments. Nine physical symptoms were selected for the data matrix: (i) time elapsed since the beginning of the complaints, (ii) weight loss, (iii) nervousness, (iv) thermophobia, (v) surface of the goiter, (vi) consistency of the goiter, (vii) thrill over the thyroid region, (viii) exophthalmia, and (ix) lymphadenopathy. The patients whose record did not include all nine symptoms were discarded. Sixty-six cases were thus included in the final sample, with the following distribution of diagnoses: 14 cases of thyroid cancer, 27 cases of toxic goiter, 7 cases of nontoxic nodular goiter, 10 cases of nontoxic diffuse goiter, and 8 cases of toxic adenoma.

2.2. Probabilities

The estimation of a priori and conditional probabilities was made according to the method suggested by Bailey[2] for small-sized samples, i.e.,

$$P(D_j) = [N(D_j) + 1]/(J + k)$$
$$P(S_i/D_j) = [N(S_i/D_j) + 1]/[N(D_j) + m]$$

where:

$P(D_j)$ is the a priori probability of diagnosis D_j.
$P(S_i/D_j)$ is the conditional probability of symptom—or subsymptomatology—S_i for diagnosis D_j.
$N(D_j)$ is the number of observed cases with diagnosis D_j.
$N(S_i/D_j)$ is the number of cases where S_i and D_j are found to occur simultaneously.
J is the total number of cases.
k is the number of diagnoses.
m is the number of distinct values of S_i. For the time since the beginning of the complaints, the surface of the goiter, and its consistency $m = 3$. For all the other symptoms $m = 2$. For a subsymptomatology including, e.g., four binary symptoms, $m = 16$. For the whole symptomatology $m = 1728$.

2.3. Optimality

A subsymptomatology (SS) is defined as a subset of the set of symptoms. The number of distinct subsymptomatologies is $2^9 = 512$. There are C_r^9 subsymptomatologies with r symptoms. A subsymptomatology will be called "optimal" (OSS) if it satisfies the condition that its *diagnostic efficiency* is larger than the diagnostic efficiency of any other SS with the same number of symptoms. While our previous study[3] used information maximization as the optimality criterion, the present study will use the actual rate of correct recognitions for the same purpose.

2.4. Sizes

(a) A SS with m symptoms will be denoted by SS(m). In order to distinguish between distinct SS of identical sizes, we will write, for each m, $SS_j(m)$ with $j = 1$ to C_m^9.

(b) For each value of m the diagnostic efficiency, i.e., the rate of correct diagnostic recognition, has been determined for each $SS_j(m)$. That particular $SS_j(m)$ with maximal efficiency $f_j(m)$ is the optimal subsymptomatology OSS(m).

(c) For all values of m the following statistics were computed:

$f(O, m)$: recognition rate of OSS(m).
$f(M, m)$: average recognition rate of all $SS_j(m)$.
$f(S, m)$: standard deviation of $f_j(m)$.

2.5. Inference

In the empirical Bayesian approach conditional probabilities are computed for each symptom, and the independence assumption is made, i.e., the conditional probability of a subsymptomatology is obtained by the multiplication of the conditional probabilities of its individual constitutive symptoms. In the straightforward Bayesian approach conditional probabilities are computed directly from their frequencies in the sample.

2.6. Evaluation

Two methods were used to test the performance of each subsymptomatology.

(A) The biased method (reclassification method): The probabilities were estimated from the 66 cases and the model was used as a diagnostic decision rule for the same 66 cases.

(B) The unbiased method (leave-one-out method): In order to test the performance of the model for each case, the probabilities were estimated from the 65 other cases. Hence, since none of the 66 tested cases contributed to the computation of the values of the model by which its recognition was attempted, all 66 cases can be considered as unknown cases. The latter method is also known as the $(J-1)$ method, and it has been described by Lachenbruch[12] for the unbiased determination of performance in linear discriminant analysis.

2.7. Notation

In order to avoid possible ambiguities, the diagnostic efficiency will be denoted by $f(A, B, C, D)$ wherever necessary, with the following conventions:

A stands for the method of inference, with:

$A = E$ for the independence assumption.
$A = D$ for the nonindependence assumption.

B stands for the method of evaluation, with:

$B = B$ for the reclassification method.
$B = U$ for the $(J - 1)$ method.

C stands for the particular statistic, with:

$C = O$ for the OSS.
$C = M$ for the average efficiency.
$C = S$ for the standard deviation of the efficiencies.

D is the size of the SS.

Hence, e.g., $f(E, B, M, 5)$ is the average diagnostic efficiency of the SS of size 5, determined by the reclassifictaion method, and using the assumption of the mutual independence of symptoms.

3. RESULTS

1. All values of $f(A, B, C, D)$ for $C = O$ are given in Table I (diagnostic efficiencies of optimal subsymptomatologies).

2. All values of $f(A, B, C, D)$ for $C = M$ and $C = S$ are given in Table II (average and standard deviation of the efficiencies of all subsymptomatologies of given size).

3. It is obvious from the latter results that optimization enhances considerably the diagnostic efficiency. This is better illustrated diagrammatically. In Fig. 1 we have plotted the difference between maximal and

Table I. Diagnostic Efficiencies (%) of Optimal Subsymptomatologies of Increasing Size[a]

	Size m =	0	1	2	3	4	5	6	7	8	9
I.	$f(E, B, O, m)$	40.9	54.5	63.6	71.2	75.6	80.3	83.3	81.8	83.3	83.3
II.	$f(D, B, O, m)$	40.9	54.5	63.6	74.2	75.8	78.8	78.8	77.3	75.8	71.2
III.	$f(E, U, O, m)$	40.9	54.5	63.6	71.2	69.7	69.7	75.8	74.2	74.2	71.2
IV.	$f(D, U, O, m)$	40.9	54.5	63.6	69.7	68.2	69.7	59.1	57.6	51.5	50.0
V.	I − III	0	0	0	0	5.9	10.6	7.5	7.6	9.1	12.1
VI.	II − IV	0	0	0	4.5	7.6	9.1	19.7	19.7	24.3	21.2
VII.	I − II	0	0	0	−3	−0.2	1.5	4.5	4.5	7.5	12.1
VIII.	III − IV	0	0	0	1.5	1.5	0	16.7	16.6	22.7	21.2

[a] For explanation of the abbreviations see the text.

Table II. Average and Standard Deviation of the Diagnostic Efficiencies (%) of Subsymptomatologies of Increasing Size[a]

Size m =	0	1	2	3	4	5	6	7	8	9
$f(E, B, M, m)$	40.9	45.5	51.7	57.6	62.8	67.3	71.5	75.9	80.1	83.3
$f(E, B, S, m)$	—	5.2	5.7	5.1	5.0	5.0	4.8	3.9	2.8	—
$f(D, B, M, m)$	40.9	45.5	51.9	58.7	64.8	68.0	68.9	70.2	70.9	71.2
$f(D, B, S, m)$	—	5.2	5.6	5.2	5.0	4.3	4.3	4.1	3.5	—
$f(E, U, M, m)$	40.9	45.5	45.9	51.5	55.1	58.4	61.6	64.8	67.8	71.2
$f(E, U, S, m)$	—	5.1	10.7	8.9	7.5	6.0	5.8	5.3	5.8	—
$f(D, U, M, m)$	40.9	45.5	46.3	49.3	50.5	51.4	49.9	49.0	49.2	50.0
$f(D, U, S, m)$	—	5.1	10.8	9.5	8.1	5.8	5.0	3.6	1.9	—

[a] For explanation of the abbreviations see the text.

mean diagnostic efficiency as a function of size, using unbiased estimates and empirical Bayesian inference, i.e.,

$$g(m) = f(E, U, O, m) - f(E, U, M, m)$$

While $f(E, U, M, m)$ is closely fitted by a straight line, $f(E, U, O, m)$ reaches a plateau at $m = 3$ and begins to decrease from $m = 6$ on.

4. The results presented in Table I enable us to observe some meaningful features in relation with the problem of bias in performance evaluation.

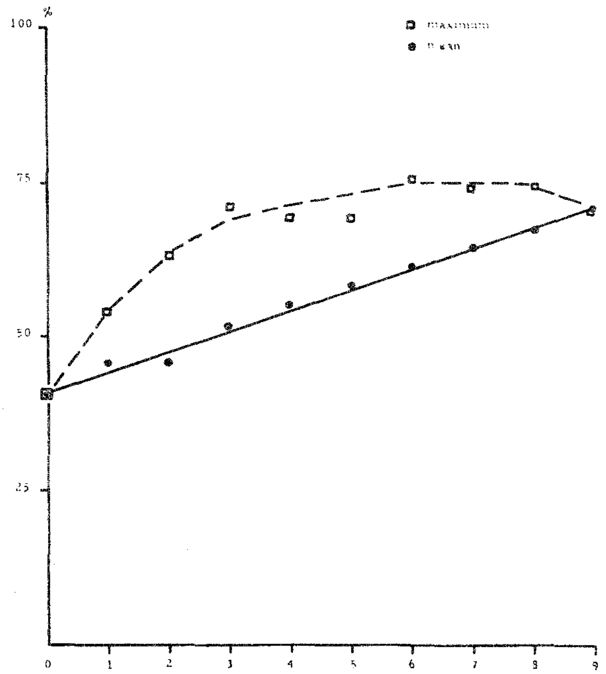

Fig. 1. Diagnostic efficiency (%) for optimal (dashed line) subsymptomatologies, and mean diagnostic efficiency (%) for random subsymptomatologies (solid line) versus size *m*.

(a) A variable bias is actually present when the reclassification method is used for the determination of diagnostic efficiency.

(b) The importance of the bias remains quite low as long as the number of symptoms is less than four. Hence, the diagnostic efficiency measured for optimal subsymptomatologies is less likely to lead to overestimation than the diagnostic efficiency measured on complete symptomatologies.

(c) When the number of symptoms is increased from four to nine the maximum extent of the bias appears to reach about 10% for the empirical Bayesian inference and 20% for the nonmultiplicative use of Bayes' rule.

(d) For biased results the rejection of the multiplicative approximation involves little advantage for subsymptomatologies of size four or less, and is actually a nuisance when more symptoms are used. With unbiased estimations the nonassumption of mutual independence of symptoms does always result in a fall of diagnostic efficiency. The importance of the fall increases with the number of symptoms, being still insignificant for m less than six, with a sharp rise to 20% thereafter.

5. The composition of all unbiased optimal subsymptomatologies for sizes from one to nine is given by Fig. 2. Two features are readily apparent:

(a) The optimality criterion can occasionnally be satisfied by more than one SS(m), as happens for $m = 5$.

(b) The first six values of m show a phenomenon of absorption of the OSS of increasing size, i.e., if we denote by OSS(i) the composition of the OSS for $m = i$, we have

$$\mathrm{OSS}(i) \subset \mathrm{OSS}(i+1)$$

where i is less than six and $\subset$ is the symbol of subset inclusion. The phenomenon of absorption has practical implications as soon as we think in terms of sequential diagnosis. This will be discussed later. However, we can already observe that:

(a) Proceeding sequentially from $m = 1$ on, we will meet the optimality criterion everywhere except at $m = 7$.

(b) At $m = 5$ two SS are optimal. Starting from the first one (5A) and going on with symptom acquisition, we will still be able to meet the optimality criterion everywhere except at $m = 7$. This is not so with the other, OSS(5B), since it is not part of any OSS($m > 5$) except at $m = 9$.

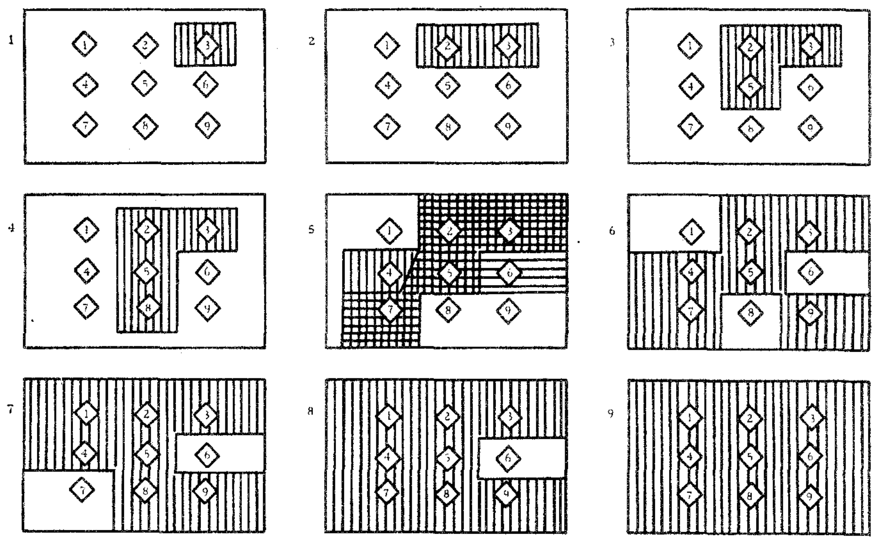

Fig. 2. Optimal subsymptomatologies for sizes one to nine. For size five, one optimal *SS* is surrounded by vertical lines and the other by horizontal lines.

4. DISCUSSION

4.1. Bias

The biased nature of the reclassification method of performance evaluation has been recognized by many authors. The extent of the bias in the reclassification method has been determined by use of the $(J - 1)$ method or by the separation of the cases under study into a training set and a test set. In all cases where the former method has been used the extent of bias of the reclassification evaluation method was found to be far from negligible. For example, the recognition rate falls from 89.1 % to 58.5 % in a study of liver diseases[6] and from 83 % to 75 % in a study of thyroid diseases.[4]

The separation method has been more widely used. Some previous results are given in Table III. In the present study the extent of the bias could be determined with accuracy by pairwise comparison of results of the reclassification and $(J - 1)$ methods. The bias is much more important for the direct than for the empirical Bayesian approach, as could well be expected. In the former case the effect of the bias is to hide the very important fall of performance observed for OSS of size larger than five. Using the empirical method the extent of the bias does not appear to be as size dependent. The bias remains at a semiconstant level of 10 % for sizes larger than four. Using the empirical method the bias is negligible for sizes up to three. The importance of the overestimation by the reclassification method is of the same order of magnitude for optimal and random subsymptomatologies.

4.2. Inference

The postulate of the mutual independence of symptoms has been subjected to much criticism.[1,7,22] Its legitimacy has been repeatedly invali-

Table III. Evaluation of Bias by Comparison of Performance on a Training Sample and New Cases[a]

Reference	Disease	Training set	New cases
Nordyke *et al.*[16]	Goiter	93.96	93.04
Fleiss *et al.*[8]	Psychiatry	56	43
Horrocks *et al.*[9]	Gastroenterology	91	84
		87.8	79.7
		88	64
Bouckaert *et al.*[5]	Goiter	83	72

[a] Results in percent correct recognition.

dated in particular cases. However, it is obvious that any other method requires that more information should be extracted from the available data, a process that can only lead to increased bias in reclassification estimates. Some previous research has been devoted to the finding of an acceptable compromise between the use of arbitrary assumptions and excessive information requirements.

Several methods have been suggested for getting rid of the interference of interdependent symptoms. However, most of them failed to take into account the biased nature of the reclassification method.[17,18] Many interesting studies have used some kind of linear discriminant analysis. It has to be kept in mind that the latter method rests on particular assumptions, e.g., multinormal distribution of variates and equality of variance-covariance matrices, that are not likely to be met. Nevertheless, these methods are not very sensitive to departures from the Gaussian hypothesis[11] and it is interesting to compare their results with the performance of the Bayesian method. Actually, for thyroid function diagnosis, by physical examination and laboratory test data, Nordyke *et al.*[16] found that the empirical Bayesian approach was superior to linear discriminant analysis. No striking difference was observed by Scheinok and Rinaldo[19] for upper abdominal pain syndromes, nor by Fleiss *et al.*[8] for psychiatric classification. In cases where quadratic rather than linear discriminant functions were used the postulate of the common covariance matrix vanishes but the information requirement increases. The problem of bias in discriminant analysis has been the subject of some recent studies.[10] The bias seems to be larger for quadratic than for linear functions.[14] The same finding was made by Michaelis[13] using medical material, whenever the bias was evaluated by the $(J - 1)$ or by the sample division method.

Several other interesting methods of dealing with the problem of interdependence among symptoms have been used in diagnostic studies. However, we will limit ourselves to the discussion of the enlightening study of Templeton *et al.*[21] on the differential diagnosis of solitary radiological pulmonary nodules. In the latter study nine diagnoses were considered. Using the empirical Bayesian approach, the diagnostic recognition rate was 67%. It rose to 95% if a linear dependence between symptoms was assumed and taken into account. This study had the merit of demonstrating clearly the real kind of trouble with reclassification estimates in the comparison of diagnostic methods using different numbers of parameters. If N is the number of patients and m the total number of symptoms, N experimental points are available for the estimation of m parameters provided there are no missing items and if linear dependence is not assumed. As soon as the hypothesis of linear dependence is used, the number of parameters needed becomes $(m/2)(m + 1)$ since we must take into account $(m/2)(m - 1)$ coefficients for the description of all first-order interactions. Since the number of observations is fixed and is mN, it is obvious that the size of the sample will rapidly become too small to allow valid determination of the values of the parameters. This conclusion was strongly supported by the results of Templeton *et al.*[21] These authors found that when the estimated parameters were used for diagnosis on the same sample from which they were computed the replacement of the independence assumption by a linear dependence model led to a rise of diagnostic efficiency from 67 % to 95 %. However, the use of these parameters for diagnosis on another sample for the same diagnostic problem allowed only 40 % of correct diagnoses to be made with the hypothesis of independence, and 27 % under the assumption of linear dependence. A fall of 95 % to 27 % is of course depressing, especially in view of the smaller loss observed with the independence hypothesis. Hence, the linear dependence model, far from improving the performance of the system, led to the nearly complete disappearance of its predictive power.

Summarizing, in the present study the complete rejection of the independence hypothesis for Bayes' rule was found to be associated with a reduction of the real recognition rate, a finding in agreement with previous studies using different models. Kanal and Chandresakaran[10] point out that "decision procedures involving estimated covariance matrices when nothing is known about the dependence or independence of the variables are non-optimal when in fact the variables are independent."

4.3. Optimal Size

As was pointed out, it is probable that not all the information contained in clinical records can really be put to use for recognition purposes. In the present study a glance at Fig. 1 shows clearly that six out of nine symptoms are actually enough to give us all the information really available. The fact that the diagnostic efficiency actually *decreases* when the size of the OSS increases beyond a certain point is also of interest. There is no drastic reduction of the diagnostic efficiency after the optimal size is reached, as would be the case for some theoretical models.[10] Nevertheless, it is important to point out that although it is impossible to do better with nine symptoms than with six, it is quite possible to do worse. This statement is of course valid only for *optimal* subsymptomatologies. The solid line in Fig. 1 shows that there is a clear relationship between the size of the SS and its *average* diagnostic efficiency. Hence, one has to make a sharp distinction between two basically different situations:

(a) When the subsymptomatology used is an optimal one (i.e., an OSS) the *optimum optimorum* is already reached with two-thirds of the symptoms. To increase the size of the OSS still further would only be harmful.

(b) When the SS are composed of symptoms drawn randomly from the collection of nine, any new symptom added to a subset already used will lead to enhanced performance.

The contrast between the two statements above is, by itself, a justification of the optimization of the subsymptomatologies.

The optimal size of the OSS depends also on the kind of inferential rule used. While there is no important loss of performance for empirical Bayesian decisions when the size of the OSS is larger than six, the loss is much more pronounced when the multiplicative assumption is not made. In the latter case the optimum optimorum is reached for sizes 3–5 and performance deteriorates rapidly thereafter, the accuracy obtained with nine symptoms being less than that achieved by proper selection of one single symptom.

Hence, the assumption of Bouckaert[3] that the direct application of Bayes' rule to diagnosis of goiters would allow subsymptomatologies to be found with increased diagnostic efficiency is certainly not supported by the present findings. What appears clearly here is that if there is actually an optimal size of OSS, this has very little to do with the problem of symptom independence. The empirical Bayesian OSS's show an optimum optimorum even without the problem of statistical validity reduction with size. The optimum optimorum is sharper for direct Bayesian inference: This is probably at least partially an effect of the limited statistical validity of esitmates, but since the latter procedure does not prove to be superior to the former before the optimum is reached and is clearly less efficient beyond the optimal size, there is no reason to give particular consideration to this nonoptimal decision procedure. In another context, the observation of Mount and Evans[15] that the increase of the size of the training set could not be followed by an increase of the diagnostic efficiency beyond a certain limit seems completely in agreement with the present results. However, the suggestion of these authors that in place of further sampling further investigation into the structure of symptom interaction may prove more profitable is not supported by this study with the present data using an unbiased estimation of performance.

The present study has many points of similarity with the study carried out by Scheinok and Rinaldo[19] for the differential diagnosis of upper abdominal pain. In the latter study 11 physical signs were used for a biased estimation of Bayesian diagnosis using the multiplicative rule and Bailey's correction for small samples. The peak diagnostic efficiency was found for a SS of size nine. As in the present study, the diagnostic efficiency decreased from size nine to size 11 on, and the extent of the loss was of the same order of magnitude: The diagnostic efficiency of OSS(9) was 58.33% vs. 57% for OSS(12). In the present study we found the peak diagnostic efficiency at OSS(6) with 83.3% and we found again 83.3% for OSS(9). The peak phenomenon was more readily apparent when unbiased estimates were used: In this case the peak occurred at OSS(6) with 75.8 % vs. 71.2 % for OSS(9). In this case, thus, as in the quoted study of abdominal syndromes, the optimum optimorum occurs at two-thirds of the maximal size. The diagnostic efficiency for larger sizes either decreases steadily or remains unchanged, according to the method of evaluation.

Other authors have tried to determine the influence of the size of the subsymptomatology on the diagnostic efficiency. However, subsymptomatologies were usually not optimized at each stage. It is thus best to compare their results with the results of Table II, where average efficiencies are shown. Mount and Evans[15] studied simulated symptomatologies of size 100. The shape of the diagnostic efficiency curve showed a plateau from size 60 on, symptoms being generated by simulation from an arbitrary matrix of conditional probabilities. If the conditional probabilities were generated first, then determined on a training set and used for recognition on an independent set, the plateau did not appear before size 80 for a training sample of 1000 cases, and no plateau was observable if the size of the training set was less than 1000. Thus, with small training sets the shape of the performance curve approached that of a straight line, a finding analogous to ours.

4.4. Design

The theory of optimal subsymptomatologies did not develop for research into the structure of the recognition process. Its main aim was to allow the design of optimal small-sized examination structures, either for screening purposes or as a response to the challenge of medical manpower shortage. We are able to show here that the examination time can be appreciably shortened without loss of accuracy, provided the composition of the examination is designed optimally. A second problem now arises due to the fact that the queuing process generated by examination time consumption is a stochastic process. The time imparted to examination will thus be sometimes larger, sometimes shorter, than the fixed time allowed to the OSS. If we assume that the patient input follows a statistical distribution, then two kinds of questions arise:

(a) If the OSS(m) type of examination is completed before the next patient shows up, what is the best thing for the examiner to do?

1. To stop the examination and wait for the next patient?

2. To increase the size of the examination: (i) by picking an ($m + 1$)th symptom at random and using it as a supplementary and supposedly useful predictor? or (ii) by increasing optimally the size of the OSS, i.e., by performing OSS($m + 1$)?

(b) If the patient is unable to go through the full course of the OSS(m), the actual number of symptoms investigated will be $m - 1$ or $m - a$ according to each case. What will be the effect of such an omission? Or, more precisely:

1. What effect has the random deletion of one of the m components of OSS(m) on diagnostic efficiency?

2. What is the effect of the replacement of OSS(m) by OSS($m - 1$)?

In both cases (a) and (b), questions 2 are closely related to the problem of the analogy between the OSS and sequential diagnosis. This matter will be discussed later. We will first deal with the questions of kind 1, related to the effects of addition or deletion of symptoms at random.

Since there is a wide range of variation for a given value of the size m between the diagnostic efficiencies of various subsymptomatologies, it quite often happens that some of the OSS for m have a better diagnostic efficiency than some nonoptimal subsymptomatology for $m + 1$. If no preliminary planning of investigations has been made, let us suppose that the diagnostician observes that the time allowed to him for his investigations offers him the possibility of pushing ahead the examinations one step ahead of what was originally contemplated. If he does not know in advance the diagnostic efficiencies of all $SS_j(m + 1)$, can he still be confident about the fact that the supplementary information obtained by adding one supplementary symptom at random will, on the average, increase its diagnostic efficiency? The answer can be found in Table IV. It is obvious that:

(a) Any observation is always better than no observation at all.

(b) The gain in diagnostic efficiency obtained by adding at random one, two, or even three supplementary symptoms to an OSS is either rather low (negligible) or, usually, negative. In our previously introduced terminology, this means that

$$f(x, x, M, m + 1)$$

is not superior to

$$f(x, x, O, m)$$

The complementary problem can be stated as follows: m being fixed, it can happen that it will unfortunately not be possible to observe all m elements of OSS(m) either because it would take too much time and thus we are forced to decrease the value of m, or because some element of OSS(m) is not observable for certain reasons. While in the first case the problem is one of symptom omission, in the second case it is a problem of symptom replacement. Here, too, the diagnostician does not know the diagnostic

Table IV. Mean Diagnostic Efficiency M Obtained when A Symptoms Outside the Optimal Subsymptomatology are Added to the Latter[a]

$m =$	0	1	2	3	4	5	6	7	8
$A = 0$									
M	40.9	54.5	63.6	71.2	69.7	69.7	75.8	74.2	74.2
$A = 1$									
M	45.5	54.9	64.1	66.2	68.2	70.8	70.2	69.7	71.2
S	5.1	9.5	4.08	2.47	1.5	4.2	1.7	6.4	—
$A = 2$									
M	45.9	58.0	62.1	66.1	68.3	69.2	71.7	71.2	—
S	10.7	7.2	4.6	3.2	3.7	1.8	2.3	—	—
$A = 3$									
M	51.5	58.5	62.8	67.3	69.2	70.8	71.2	—	—
S	8.9	5.3	5.0	4.4	2.0	3.1	—	—	—

[a] The standard deviation of the resulting diagnostic efficiency S is also given. The size of the optimal subsymptomatology is m.

efficiencies of all $SS_j(m)$ or $SS_j(m-1)$ and thus he deletes or replaces symptoms at random. The critical values are

$$f(x, x, O, m) - f(x, x, O_m - 1, m - 1)$$

for deletion [$O_m - 1$ stands for the $SS_j(m-1)$ obtained by deletion of one element of OSS(m): hence, only a subset of the $SS_j(m-1)$] and

$$f(x, x, O, m) - f(x, x, M', m)$$

for replacement [M' stands for the average of all $f_j(m)$ excluding $SS_j(m) =$ OSS(m)]. The effects of replacement and deletion are summarized in Table V. The fact that replacement is an inefficient procedure was already suggested by Table IV results. Since the random addition of one element to OSS(m) leads usually to a decrease of the diagnostic efficiency, it is quite probable that the random replacement of one element of OSS(m) by another will lead to an even worse efficiency. The two problems, while related, are not identical, however. OSS($m - 1$) is not always a subset of OSS(m). Similarly, the $SS_j(m-1)$ obtained by random deletion of an element of OSS(m) is not equivalent to OSS($m - 1$). Table V shows that the effects of deletion with or without replacement of the lost symptom are not very different, considered from the point of view of diagnostic efficiencies. The mean of the differences between these two procedures is 1.32%, with standard deviation 2.28% (not significant).

Table V[a]

m	OSS(m)	DEL(OSS(m))	REPL(OSS(m))
1	54.5	40.9	44.3
2	63.6	49.2	51.3
3	71.2	55.0	59.3
4	69.7	61.3	60.0
5	69.7	61.5	60.5
6	75.8	60.1	63.7
7	74.2	65.8	65.2
8	74.2	66.9	67.0

[a] m is the size of the optimal subsymptomatology.
OSS(m) is the diagnostic efficiency (%) of the optimal subsymptomatology of size m.
DEL(OSS(m)) is the diagnostic efficiency (%) obtained on the average by random omission of one of the elements of the optimal subsymptomatology of size m.
REPL(OSS(m)) is the average diagnostic efficiency of a subsymptomatology of size m containing $m - 1$ elements of the optimal subsymptomatology of size m and one element randomly selected outside this optimal subsymptomatology.

These results point to a conservative policy in diagnostic planning as the best safeguard against performance deterioration. Thus, if the observation of one item out of an OSS is impossible, it is better to be satisfied with what is left than to replace at random the missing item by an item that does not belong to the OSS. When all elements of an OSS have been observed it is better to be satisfied with this rather than to try to fill some extra time that could be available by increasing the number of observed items by random selection of an extra item outside the OSS. Since we are still convinced of the interest in designing small optimized subsets of symptoms for use in medical screening, it seems wise to reduce to a minimum the role played by the personal opinions of the examiner in the application of the planning in the field.

4.5. Sequences

At first, the process of OSS construction may appear to be similar to sequential diagnosis. In both cases, indeed, the problem at hand is one of designing an optimal subset of symptoms for a given size of the subset. The difference between the two concepts is that sequential diagnosis (SD) is subjected to a constraint on the selection of the items. For a given size m the subsymptomatology must always be a subset of the subsymptomatology selected for size $m + 1$. SD can only grow by absorption, while this is not so for OSS, where no inclusion is required for subsets of increasing size.

Moreover, in Part IV of this study we made a sharp distinction between two kinds of SD. In the first kind the subsymptomatologies for a given size m are determined by the results obtained at the $(m - 1)$th step. For the second variety all subsymptomatologies generated for the same size are identical. It was then shown that for the particular sample considered the performance of the two systems was the same. It is only in the second kind of SD that the construction of OSS becomes a related procedure. The construction of the OSS becomes identical with the planning of the second kind of SD as soon as the growth of the OSS is accomplished by absorption at all steps. Of course, this will preclude the use of some very powerful OSS's just because they do not include OSS's of smaller sizes. Hence the optimality criterion will not be met everywhere in a SD. Considering things this way, the problem of SD can be expressed as follows: nine points must be attained sequentially, each point being a symptom. At each point a certain amount of diagnostic efficiency is given as a payoff to the diagnostician. The latter is expected to maximize his payoff. He is not allowed to bypass any point or to cross the same point more than once. Since there are 362,880 distinct pathways, a solution by enumeration of all the payoffs is an impressive task. An approximate suboptimal solution is obtained by allowing at each step the OSS to be incremented by one single element. In this case the SD is shown in Fig. 3. It is obvious that it would be meaningless to allow the SD to proceed further than the third step, if the number of steps is used as a stopping rule. The procedure for SD construction used for Fig. 3 is obviously related to well-known sequential tests, e.g., stepwise multiple regression.

The SD described here differs from that reported in Part IV of this study by the very early appearance of the plateau. The two situations are not comparable, however: In Part IV, 12 (vs. 9) symptoms were used for 86

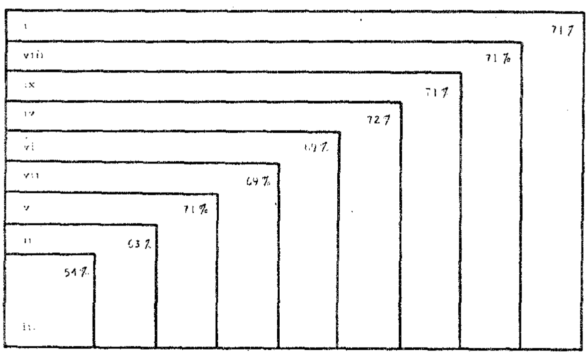

Fig. 3. Sequential diagnosis by stepwise absorption into optimal subsymptomatologies. Symptoms are referenced by roman numerals (see Section 2.1).

(vs. 66) cases. We suspect, moreover, that part of the difference is accounted for by the fact that in the present results unbiased performance estimates were used.

ACKNOWLEDGMENTS

The present study was made possible by the courtesy of the Data Processing Unit of the Zairese Ministry of Finance, whose responsibles allowed the author to use their computational facilities.

REFERENCES

1. J. A. Anderson and J. A. Boyle, "Computer diagnosis: statistical aspects," *Brit. Med. Bull.* **24**:230–235 (1968).
2. N. T. J. Bailey, "Probability methods of diagnosis based on small samples," in *Mathematics and Computer Sciences in Biology and Medicine* (H.M.S.O., London, 1965).
3. A. Bouckaert, "Computer diagnosis of goiters. III—Optimal subsymptomatologies," *J. Chron. Dis.* **24**:321–327 (1971).
4. A. Bouckaert, "Computer diagnosis of goiters. I—Classification and differential diagnosis," *J. Chron. Dis.* **24**:299–310 (1971).
5. A. Bouckaert, J. De Plaen, J. A. Kapita, and S. Ditu, "Statistical analysis of symptoms for the differential diagnosis of goiters," *Ann. Soc. Belge Med. Trop.* **52**:113–126 (1972).
6. F. Burbank, "A computer diagnostic system for the diagnosis of prolonged and undifferentiating liver diseases," *Amer. J. Med.* **46**:401–415 (1969).
7. M. F. Collen, L. Rubin, and L. Davis, "Computers in multiphasic screening," in *Computers in Biomedical Research*, Vol. III (Academic Press, New York, 1965).
8. J. L. Fleiss, R. L. Spitzer, J. Cohen, and J. Endicott, "Three computer diagnosis methods compared," *Arch. Gen. Psychiat.* **27**:643–649 (1972).
9. J. C. Horrocks, A. P. McCann, J. R. Staniland, D. J. Leaper, and F. T. de Dombal, "Computer-aided diagnosis: description of an adaptable system, and operational experience with 2034 cases," *Brit. Med. J.* **2**:5–9 (1972).
10. L. Kanal and B. Chandrasekaran, "On dimensionality and sample size in statistical pattern recognition," *Pat. Recog.* **3**:225–234 (1971).
11. S. Koller, J. Michaelis, and E. Scheidt, "Untersuchungen an einem diagnostischen Simulationsmodell," *Meth. Inform. Med.* **11**:213–227 (1972).
12. P. A. Lachenbruch, "Estimation of error rates in discriminant analysis," Ph.D. dissertation, UCLA., Los Angeles (1965).
13. J. Michaelis, "Zur Anwendung der Diskriminanzanalyse für die medizinische Diagnostik," Habilitationsschrift, Medizinische Fakultät, Mainz (1972).
14. D. H. Moore, "Evaluation of five discrimination procedures for binary variables," *J. Amer. Stat. Ass.* **68**:399–404 (1973).
15. J. F. Mount and J. W. Evans, "Computer-aided diagnosis. A simulation study," in *Fifth I.B.M. Symposium* (Endicott, New York, 1963).
16. R. A. Nordyke, C. A. Kulikowski, and C. W. Kulikowski, "A comparison of methods for the automated diagnosis of thyroid dysfunction," *Comput. Biomed. Res.* **4**:374–389 (1971).
17. C. A. Nugent, H. R. Warner, J. T. Dunn, and E. H. Tyler, "Probability theory in the diagnosis of Cushing's syndrome," *J. Clin. Endocrin.* **24**:621–627 (1964).

18. P. Scheinok, "Symptom diagnosis, Bayes's theorem and Bahadur's distribution," *Biomed. Comput.* **3**:17–28 (1972).
19. P. A. Scheinok and J. A. Rinaldo, "Symptom diagnosis: optimal subsets for upper abdominal pain," *Comput. Biomed. Res.* **1**:221–236 (1967).
20. P. A. Scheinok and J. A. Rinaldo, "Symptom diagnosis: a comparison of mathematical models related to upper abdominal pain," *Comput. Biomed. Res.* **1**:475–489 (1968).
21. A. W. Templeton, C. Jansen, J. Lehr, and R. Hufft, "Solitary pulmonary lesions," *Radiology* **89**:605–613 (1967).
22. J. M. Vanderplas, "A method for determining probabilities for correct use of Bayes's theorem in medical diagnosis," *Comput. Biomed. Res.* **1**:215–220 (1967).