J Chron Dis Vol. 35, pp. 601 to 611, 1982
Printed in Great Britain. All rights reserved

0021-9681/82/080601-11$03.00/0

MAXIMUM ATTAINABLE DISCRIMINATION AND THE UTILIZATION OF RADIOLOGIC EXAMINATIONS*

JAMES A. HANLEY and BARBARA J. MCNEIL
Department of Epidemiology and Health, McGill University, 3775 University Street, Montreal, Quebec, Canada H3A 2B4 and Department of Radiology, Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

(*Received in revised form* 1 *October* 1981)

Abstract—Discriminant Analysis and other related statistical techniques are frequently used to sort patients into those most likely and those least likely to benefit from a certain intervention. Considerable data analysis and computation are often required to arrive at the best-fitting mathematical model which translates discriminating variables or indicants into probability predictions regarding the presence or absence of disease or the likelihood of a favourable outcome. Attempts to judge how well discriminant analysis performs or to determine why it does not perform better are hampered by not knowing what is the greatest degree of discrimination theoretically possible in a data set.

In this paper we describe a method of calculating the maximum discrimination attainable in a data set and show how it can be used (1) to decide whether further model building is worthwhile, and (2) if so, to judge the discriminatory performance of any such models. We apply this tool to two previously published studies of radiologic utilization; the results provide reassurance that, at least on the basis of the presenting indicants, the patients were being adequately selected for the studies in question.

INTRODUCTION

RECENTLY there has been increased interest in developing mathematical predictions about three classes of individuals and their medical care: (1) To which subgroups of symptomatic individuals should one direct screening programs? (2) What subgroups of patients with a particular diagnostic problem are more likely to have a positive diagnostic test compared to other patients with the same problem? (3) What subgroups of patients with known disease are most likely to respond to a particular therapeutic regimen?

To develop such predictions, one usually needs to collect a large number of 'proven' cases i.e. individuals for whom the health state, test result or outcome is known, and to record for each individual in this data-base the various demographic and clinical clues or 'indicants' that would normally be available for use in predicting this 'result'. With this large data-base one can use any of a number of mathematical techniques (e.g. logistic analysis, discriminant analysis) to develop a prediction system. Ideally one then objectively assesses the performance of the prediction system by testing it on a separate independent set of individuals (the 'validation set').

Many investigators have attempted to develop discrimination or stratification rules (for a general review see Feinstein Ref. 1) for a variety of clinical problems. In all cases these rules are based on the principle that the investigator wishes to produce 'optimal stratification' or to identify 'the "best" of a group' [1]. A recent experience of ours led us to realize that the words 'optimal' and 'best' could be precisely defined from the data set itself and thereby provide an upper limit to the separation powers of any discrimination or stratification rule that might be applied to this same data-set. In brief, in a recent

*This work was supported by grants from the National Center for Health Services Research and National Center for Health Care Foundation (5 R18 HS03645) and the Hartford Foundation (79021-3H).

study [2] we were unable to definitively sort patients with neurological symptoms into those most likely and those least likely to have an abnormal computed tomography (CT) study of the head. This led us to wonder which of two situations held in this study: (1) Could we have built a more sensitive discriminant model if we had tried harder? Or, (2) Was there too little separation in the data to begin with?

In this paper we use this study on CT of the head to indicate that in this particular situation there really was too little separation in the data themselves. In the process of showing this we have developed the concept of 'maximum attainable discrimination' (m.a.d.), i.e. the maximum discrimination inherent in the data set itself. Quantification of intrinsic separability can be used in two ways. First, it can be used to decide if the inherent separability is large enough to warrant a full scale data analysis for predictive purposes. In particular, by calculating the maximum attainable discrimination before beginning an analysis, one can determine whether development of a sorting system is possible or worthwhile. Second, one can see how close any empirical sorting rule approaches the theoretical limit of separation.

METHODS

Background to this study

In a recent investigation we collected data on the presence or absence of 20 different signs and symptoms ('indicants') for 2225 consecutive patients about to undergo CT of the head [2]. Nine hundred fifty-five or 43% of these patients had an abnormal CT examination; the remaining 57% had a normal one. We tried to use the 20 indicants to predict the CT result in an individual patient with the hopes of reducing the total number of patients having CT examinations without sacrificing the discovery of a large number of patients with abnormal CT results. All of our approaches were unsuccessful, however. Our most detailed approach is summarized in Fig. 1 (left). Using a multiple regression analysis for binary outcomes [3] ('logistic regression') we found eight coefficients that were significantly different from zero; in other words, the presence of each one of these eight indicants altered the probability of an abnormal exam even after we had

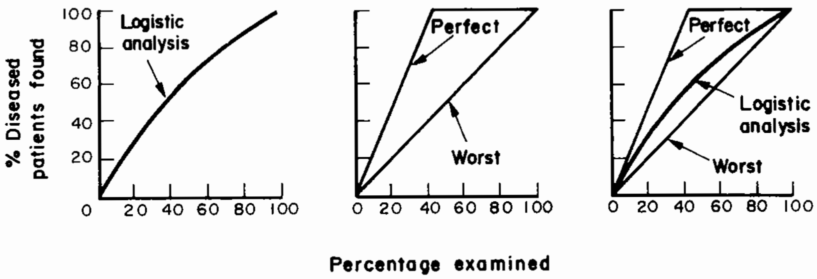

FIG. 1. Discrimination results for CT examinations of the head. The abscissa represents the percentages of the total of 2225 patients that might be examined ranging from a small portion (strict criteria) to almost all (lax criteria); the ordinate shows what percentage of the total of 955 diseased patients would be found by CT if examinations were limited to the percentage shown on the abscissa. The left plot shows the results of the logistic regression analysis described in the text. There is a close relationship between the percentage of diseased patients found and the percentage of total patients examined. For example, the analysis indicates that examining 60% of the total population of 2225 would lead to detection of about 74% of the 955 diseased patients. The middle plot gives the extremes of discrimination possible for this data set. With perfect information, since 43% of all of the patients examined actually had disease, a perfect separation technique would require examination of only 43% and this would lead to detection of 100% of diseased patients. Examining fewer than 43% would lead to a proportional decrease in the number of diseased patients found. This *perfect discrimination* is shown as the top line in the middle figure. The *worst possible discrimination* would result if there were a linear relationship between the percentage of patients examined and the percentage of diseased patients found, in other words, if the technique for examining patients ignored all information unique to a particular patient and examined patients on a purely random or indiscriminate order. The right hand figure plots the logistic analysis in relationship to the perfect discrimination and the worst discrimination lines. It is closer to the latter than to the former.

considered the presence/absence of the other seven. From these coefficients we constructed the estimated probability of an abnormal scan for each patient and evaluated the impact of limiting examinations to those for whom the probabilities were high. We did this by plotting the percentage of the 955 patients with CT-detectable disease in the population we would have found against the corresponding percentage of the entire population of patients we would have examined, considering that we would examine first individuals whose probability of disease was highest in the logistic regression and then individuals with successively lower values. Figure 1 (left) shows that although the resulting curve was convex upwards, reflecting the decreasing marginal gains from successively lower-yield examinations, the curvature was not very marked, indicating that there was only a slightly-better-than-linear relationship between the percentage of diseased patients found and the total percentage examined. Could we do better?

Extreme levels of discrimination

For the data set under consideration, because 43% of the patients had abnormal CT's, the minimum number of patients we could examine to find 100% of the diseased patients would obviously be 43% of the total—that is, examine only those who, as if one had precognition of the final health state, were diseased. In a similar way, examining fewer (say 1/2 of the 43%, or 21.5%) of the total would lead to detection of exactly the same reduced fraction (in this example, 1/2 of 100 or 50%) of the diseased patients. This is graphically displayed in Fig. 1 (center), and this curve, with an initial slope of 100:43 or 2.33:1 defines a state called *perfect discrimination.* Clearly, while this is what we would *like* to be able to do, it is in reality unattainable.

At the other extreme, *worst possible discrimination* occurs if we ignore all information unique to a particular patient and examine patients on a purely random or indiscriminate order. With this system, there is a direct 1:1 linear relationship between the percentage of abnormals detected and the percentage of the population examined; this line, with a slope of 1, is displayed as Fig. 1 (center).

The discrimination results for the CT data fall between the curves representing *perfect discrimination* and *worst possible discrimination* (Fig. 1, right). The question then becomes "How much closer to perfect discrimination could we get?" Or, said differently, how much inherent separation exists in the data set itself? What is the *maximum attainable discrimination* we could *hope* for?

RESULTS

Maximum attainable discrimination

Two populations (for example, normal and diseased) under consideration can be described according to a number of different attributes, criteria, or indices which are limited in their ability to distinguish between the two populations by *inherent* characteristics of the populations themselves. The greater the amount of relevant information considered the greater the chance of achieving separation between two groups. These basic principles lead to the identification of the maximum discrimination possible between two groups attainable by considering successively greater amounts of the information available on them.

In order to illustrate this principle consider the data base on CT of the head and assume that *for illustrative purposes* we wished to consider only 3 of the 20 indicants we assessed, i.e. headache, seizure and motor weakness. How much *inherent* separation is there in the group of patients with normal CT's compared to abnormal CT's in the data set when only these 3 variables are considered? What does a curve of percentage of patients with disease detected against percentage of patients examined look like? Where does it fit in the space defined by perfect and worst discrimination (Fig. 1 center)?

In order to derive the maximum attainable discrimination assume, as in Fig. 1, that we try to rank patients in order of decreasing probabilities of having an abnormal CT result. We will define our ranking as optimal if it guarantees that we would identify the greatest

Table 1. Yield of abnormal scans in the $2^3 = 8$ subgroups generated by the symptoms headache, seizure and motor weakness

Symptom	Symptom Complex							
Headache	−	−	−	−	+	+	+	+
Seizure	−	−	+	+	−	−	+	+
Motor weakness	−	+	−	+	−	+	−	+
1. Number in subgroup	884	495	203	86	331	139	55	32
2. Number abnormal	373	260	67	47	103	71	19	15
3. Percentage abnormal	42	53	33	55	31	51	35	47
4. Optimum/ranking (highest is [8], lowest is [1])	[4]	[7]	[2]	[8]	[1]	[6]	[3]	[5]
5. Percentage of patients with this or higher ranking	74	26	85	4	100	32	76	34
6. Percentage of total abnormals in those with this or higher ranking	80	32	89	5	100	40	82	41

percentage of abnormals for any given number of examinations performed that is, that it is close to the 'perfect' curve in Fig. 1.* Consider the above case of three symptoms ($k = 3$).

Using headache, seizure and motor weaknesses, there will be $2^3 = 8$ subgroups; Table 1 indicates the number of patients from the original 2225 who fell into each of the 8 subgroups (line 1), along with the numbers (line 2) and percentages (line 3) who were abnormal in each subgroup. Clearly the optimum ranking of these subgroups is according to the percentages in line 3: the subgroup with the highest percentage of abnormals (55%) receives the highest priority and the subgroup with only 31% abnormal the lowest (line 4). In order to display the performance of such a ranking of these 2225 patients, we determine the cumulative percentages of the 2225 who would be examined by successively relaxing the criteria for examination (line 5) and the cumulative yields (line 6), and plot them in a '% detected vs % examined' curve in Fig. 2.

*We could also argue for a different underlying principle for definition of 'optimum', and the analysis could be adapted to reflect this change. For example, we could argue that the ranking should lead us first to examine the most uncertain group, that is, those with a probability close to 50% of having an abnormal CT and then to examine those with probabilities closer to 0 and 1. In fact, Pauker has developed methods for determining the upper and lower boundaries for this 'most in need of testing' group with the greatest uncertainty [4]. The concept of maximum attainable discrimination can be used with any set of priorities, provided that they are well defined and agreed upon.

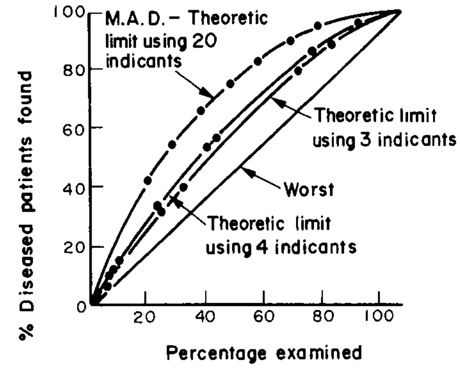

Fig. 2. Limits of detectability for CT studies of the head using 3, 4, or all 20 of the original indicants collected on each patient. The ordinate represents the percentage of diseased patients found and the abscissa the percentage of the patients examined. The light diagonal line corresponds to worst possible discrimination, similar to that seen in Fig. 1. The other three lines correspond to the theoretical limit of detectability using only three specific pieces of data on each patient, only four specific pieces of data on each patient, or all 20 pieces of data on each patient. The more information used, the greater the detectability. Because only 20 pieces of data were collected on each patient, the curve created using these data (as shown in Tables 1 and 2), is labeled the *maximum attainable discrimination curve.*

If the number of indicants used (k) is increased to 4 by adding, for example, the symptom dementia, the best ranking of the $2^4 = 16$ subgroups so generated produces a performance curve which can only 'better' than that with $k = 3$. The reasoning is as follows: adding one additional symptom has the effect of splitting each of the 8 subgroups into two smaller groups—one with patients who present with this fourth symptom and one with those who do not. Results are shown in Table 2. For example, the second highest ranked group with 495 patients has a yield of 53%. Subdividing it into those presenting with and without dementia produces subgroups with yields of 62 and 49% respectively. Similarly those in the low-yield group of 331 patients can be reranked, using this fourth symptom, into a group of 51 with a 39% yield and a remaining group of 280 which now has a yield of only 30%. This procedure can be applied to each of the eight original groups, and the 16 resulting subgroups can now be reranked as is shown in line 4 of Table 2. Rows 5 and 6 (of Table 2) can again be used to plot the performance of a ranking system based on these four symptoms.

It is now easy to see where this reasoning leads: if we use all 20 symptoms and findings, we can divide the 2225 patients into unique sets where all patients within a set have exactly the same symptoms and are thus indistinguishable from one another but are different (in at least one symptom) from patients in any other set. The subsets so formed (in our example, the 2225 patients divided into 646 different subsets) are the *finest possible partition*; one cannot distinguish them further without using a 21st indicant, their names, hospital identification number, etc. Thus, the performance curve generated by the optimum ranking of these 646 sets is 'better' than that produced by any other ranking system which uses some or all of the 20 presenting symptoms. So, by ranking on the observed proportion of positive scans associated with each of the symptom patterns, we established an upper bound on the 'separability' of this data set which any discriminant analysis of the data might achieve.

This maximum attainable discrimination associated with a ranking rule that used all 20 indicants for this group of 2225 patients is shown as the upper curve in Fig. 2. This analysis showed that examining 70% of these patients, for example, would *at the very best* detect only 92% of the disease. Moreover, when we examined those 'missed', we found they included a number of important conditions such as brain tumors. Thus, these 20 indicants provide no clearcut separability of those with normal studies from those with abnormal CT studies, especially none that would warrant further extensive data analysis or sizeable reductions in the number of patients scanned: Figure 3 plots the logistic results presented earlier in Fig. 1 with the maximum attainable discrimination results obtained for 20 indicants. The former is everywhere below the latter but only slightly so, suggesting that our logistic analysis leads to predictions only slightly below those maximally attainable by or inherent to the data themselves. This raises the question of significant differences.

Significance results

Two questions of significance may be raised. (1) What is the difference between the two curves in Fig. 3? and, (2) are the 20 indicants really helping us in our discrimination task?

The usual tests of statistical significance do not apply to the data in question (1). The real issue is "Is it worth trying other additional logistic expressions (i.e. by adding on many more combination terms) or other separation techniques to get closer to the m.a.d. curve?" The answer here depends on the investigator's goals. The theoretical data would suggest that if the *very best* we can do to find 92% of the diseased patients is to examine 70% of the original population we might as well conclude that there really is not enough separation in the data we are collecting to differentiate normals from abnormals. Either we must recognize that appropriate referral criteria exist or that we have to add additional data (e.g. 21st, 22nd, ... indicants) to the data base and thereby attempt to improve the inherent and hopefully practically attainable discrimination.

Regarding question (2)—what is the chance variation around the m.a.d. curve? In considering the m.a.d. curve it is necessary to calculate the expected chance variation

TABLE 2. YIELD OF ABNORMAL SCANS IN THE $2^4 = 16$ SUBGROUPS GENERATED BY SUBDIVIDING EACH OF THE $2^3 = 8$ SUBGROUPS IN TABLE 1 ON THE BASIS OF A FOURTH SYMPTOM (DEMENTIA)

Symptom																
Headache	−		−		+		+		−		+		−		+	
Seizure	+		−		−		+		−		+		+		−	
Motor weakness	+		+		+		+		−		−		−		−	
Rank among 8	[8]		[7]		[6]		[5]		[4]		[3]		[2]		[1]	
4th Symptom	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+
1. Number in subgroup	55	31	372	123	98	41	16	16	584	300	35	20	144	59	280	51
2. Number abnormal	30	17	184	76	42	29	8	7	214	159	10	9	39	28	83	20
3. Percentage abnormal	55	55	49	62	43	71	50	44	37	53	29	45	27	47	30	39
4. Optimum ranking among 16	[13]	[14]	[10]	[15]	[6]	[16]	[11]	[7]	[4]	[12]	[2]	[8]	[1]	[9]	[3]	[5]
5. Percentage with this or higher ranking	11	9	42	7	51	2	25	46	79	25	94	46	100	45	92	53
6. Percentage of total abnormals in those with this or higher ranking	16	13	53	11	62	3	33	57	86	33	96	57	100	56	95	64

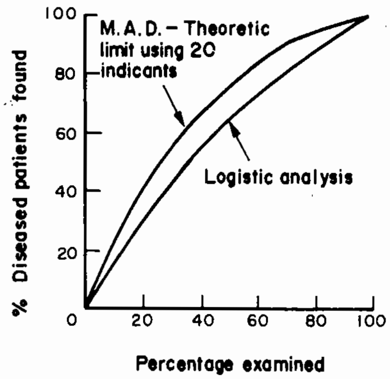

FIG. 3. Comparison of logistic results with maximum attainable discrimination results. The ordinate represents the percentage of diseased patients found and the abscissa the percentage of total patients examined. The maximum attainable curve using 20 indicants is everywhere above logistic analysis. However, the difference between these is relatively small; above 10% more diseased patients are found for a given percentage of total patients examined using maximum rather than the actual separability. This indicates that very little could be gained by further attempts using different separation algorithms.

around an 'expected' curve. In other words, it is necessary to calculate an interval within which for any given percentage of the population scanned the maximum percentage of disease detected would be most often expected to fall, if, regardless of symptoms, the probability of an abnormal result were the same for each patient instead of different for each patient. If the maximum attainable discrimination curve calculated from the CT data set lay below the upper limit of this calculated interval, then we would have to conclude that the maximum separation observed may not have resulted from information in the various symptoms but rather from capitalizing on chance fluctuations.

In order to calculate the expected variation, we used a computer simulation and for this purpose repeatedly assigned the abnormal examinations (43% of the total) over the various patient subsets in a random manner. We then constructed the maximum attainable discrimination curve produced by each such random assignment and formed the frequency distribution of the attainable maximum. We did this at only one point on the m.a.d. curve in Fig. 3: the point indicating that 70% of the original patient population would be examined.*

The results of this computer simulation are shown in Fig. 4. Note from this graph that the average maximum attainable discrimination achieved with this simulation was approximately 90%; in fact, only 2 in 1000 simulations produced m.a.d.'s of 92% or more, thus suggesting that the 92% observed in the data-set is statistically significant at the $p = 0.002$ level. However, this 2% differential is a negligible one. In other words, when the maximum discrimination was obtained by a ranking which actually used the information given by the presence or absence of the 20 symptoms, it was only greater by 2 percentage points than the average of 90% obtained from a random assignment in which each patient was given a 43% probability of having an abnormal result.

Barium enema study. Another example of maximum attainable discrimination

In another investigation [5] we recorded the presence or absence of 34 different symptoms in 802 patients on whom barium enema examinations were requested. One hundred fifty-five of these patients or 19% were subsequently shown to have an abnormal barium enema examination. Our preliminary analyses at that time failed to indicate any means of eliminating a significant fraction of patients from this examination without also failing to diagnose an equally sizeable group of patients with disease. In particular, by examining all patients who had at least one sign or symptom which was present signifi-

*For illustrative purposes in this paper, we have assumed that with a high unit cost technological examination, like computed tomography, it might be useful to reduce the number of patients scanned to 70% of the original population.

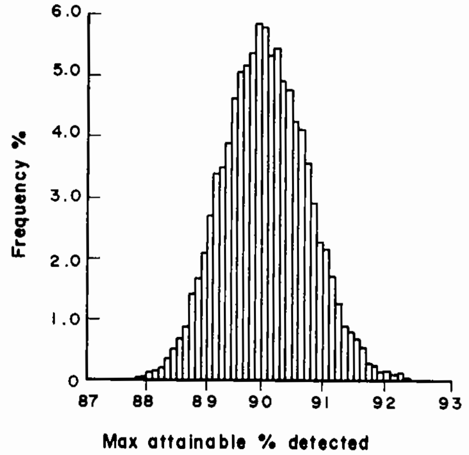

Fig. 4. Histogram showing range of variation in the maximum percentage of disease detected by examining 70% of the original patient population. On the ordinate is the frequency and on the abscissa the corresponding maximum attainable percentage detected. Simulated results obtained by randomly distributing the abnormals over all of the possible patient subgroups show that the average maximum attainable discrimination was approximately 90%. Only 0.2% of the simulations produce maximum attainable discriminations of greater than the 92% obtained on the actual data set. These results suggest that this observed 92% figure is statistically significant at the $p = 0.002$ level.

cantly more often among those with abnormal barium enemas compared to those with normal barium enemas, we were able to decrease the total number of examinations found by 36%, but at the same time we decreased the total number of diseased patients found by 23%.

In order to identify the maximum attainable discrimination of this data set using the techniques described above, we first restricted our analysis to the 19 symptoms which occurred with a frequency of 3% higher. This led to an average number of 1.76 patients per subgroup, and 65 of the 402 subgroups so formed had only a single patient in them. Under these conditions and using the techniques for developing the maximum attainable discrimination curve described above, the curve in Fig. 5, results. This curve appears high in the upper left hand part of the graph, indicating that a high percentage of abnormals could be detected by examining only a small percentage of the patients, contrary to what

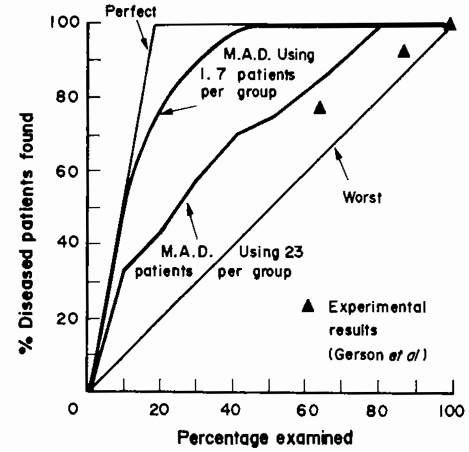

Fig. 5. Comparison of empirical results and maximum attainable discrimination results for barium enema examinations. The ordinate represents the percentage of diseased patients found, and the abscissa the percentage examined. Perfect discrimination for this data set and worst discrimination are indicated. The curved lines represent the maximum attainable using varying grouping patterns described in the text. The lower of these, corresponding to the more realistically-sized subgroups, is only slightly above the empirical results (open triangles) found in the original data collected by Gerson.

the experimental results suggested. This discrepancy arises because of the relatively low percentage of diseased patients in this group (19%) and because the instability caused by small denominators (an average of 1.76 patients per subgroup) produces an unrealistically high separation. Rather than eliminate certain symptoms to create fewer, larger subgroups, we enlarged each subgroup by a 'nearest neighbors' technique [6]. This technique ranks a subgroup with a given symptom complex by the proportion of abnormals among those with this exact set of symptoms *plus* those with these *and* any other symptoms; even though this approach produces over-lapping groups of patients, the resulting data set is more likely to give realistic and stable estimates of separability that might be expected if a still larger data base had been obtained. With this approach, each subgroup now contained an average of 23 patients, and the maximum attainable discrimination curve dropped as shown in Fig. 5. The two triangular points in Fig. 5 represent selection criteria developed at the time of the original study and indicate that while they are somewhat below the maximum attainable discrimination curve, the difference is not enormous. This suggests that, in fact, the original study had separated the patient population almost as well as theoretically possible.

DISCUSSION

Over the past few years there have been a number of studies focusing on a development of algorithms for more efficient screening, better utilization of diagnostic and ancillary services, and greater selectivity in the use of various therapeutic options. Examples include strategies for screening patients at special risk of hypothyroidism [6] or identifying carriers for Tay Sachs Disease [7], optimizing use of radiographic examinations [2, 5, 8], admitting patients to the coronary care unit [9], or the intensive care unit [10], avoiding surgical exploration of patients with cancer [11], or using medical or surgical treatment for gallstones [12].

The evaluation of such studies must consider whether the resultant algorithms were merely developed and immediately recommended for use or whether they were developed *and* then *validated* on a *separate* data set before their generalizability was suggested. After all, it is relatively easy to use a computer program which evaluates each of the indicants to choose one which is "best", then searches the remainder for the next-best one to add to the best, and so on in a stepwise manner, in essence creating finer and finer partitions of the subjects;* these if taken to their extreme would result in what we call the maximum attainable discrimination.

However it is not enough to then re-apply the algorithm which develops from this search to the very data from which it was constructed and claim that its performance there is a good measure of its performance in future patients; the claim will generally be overoptimistic. Various methods [15, 16] exist which use the original data set to obtain a more realistic estimate of the algorithm's likely future performance; however, they are at best a method of deciding whether the 'apparent' present results are so worthwhile that, even after adjustment for overoptimism, a prospective or independent validation study is worth undertaking and likely to confirm the original enthusiasm.

If an algorithm can be verified and can show that greater efficiency and selectivity of medical resources is possible, the investigator can stop. If, on the other hand a good algorithm cannot even be developed, then the investigator worries about the adequacy of his/her efforts. He must wonder whether he merely did not try hard enough to find the right separation technique or alternatively whether he was fighting an impossible battle where there really was not enough separation theoretically possible to start with. The objective of this paper was to provide a method which differentiates between these possibilities.

*A stepwise discriminant analysis [12] is the most common mathematical technique used to do this; however, the more recently available 'recursive partitioning' or 'automatic interaction detection (AID)' method [13, 14] is receving considerable attention. The latter's appeal is in its more detailed choice of successive indicants. For example, if indicant 7 is the single best discriminator, then indicant 4 may be chosen as 'second best' when indicant 7 is present, and indicant 3 if indicant 7 is absent. It can produce a quite irregular decision tree.

The methods we have described, allow us to do two things (1) to establish what the ultimate separation in a particular data set is; and (2) if the theoretically achievable separation is substantial, to judge the performance of any proposed discriminant model. If the separation is not large enough to be considered real, the analyst can be reassured of this and is spared considerable effort and time; if it is real, then he has a target against which he can measure the performance of the discriminant models he proposes.

In the examples we have presented, the maximum attainable discrimination curves provide reasonable reassurance that the patients were being adequately selected for the barium enema and CT examinations. It is obviously possible that the apparent lack of separability might have been produced by inaccuracies or misunderstandings in the recording of the data. Since all data were carefully rechecked by research assistants, however, this seems unlikely. It would be instructive to apply these methods to the data of Bell and Loop [8] to see if the remarkable discrimination they achieved was indeed real and not attributable to capitalizing on chance. If the latter were true, it might explain the contradictory results which DeSmets *et al.* [17] found when they applied the rules developed by Bell and Loop to a new population.

This paper does not address two issues. First, is the percentage of patients with *detectable* disease the desirable endpoint or label for the y-axis in Figs 1–3? Instead should the percentage of patients with *treatable* disease or the percentage of patients with *potentially curable* disease be used instead? Use of these might lead to greater inherent separation in the data set but would imply a particular value judgement about the role of diagnostic tests. Second, whatever the endpoint selected, we do not address the issues related to the choice of an ideal percentage of total examinations performed. The choice will depend upon the clinical setting, in particular the costs of missing a patient with disease as weighed against the costs of falsely classifying a normal patient as diseased.

In the Appendix, we show the computational steps required to generate the performance curve corresponding to the maximum attainable discrimination, and to calculate the various curves one would expect if the symptoms were completely uninformative. Computational costs for steps 2 and 3 are quite low; the evaluation of the expected variation will be more costly, and will depend on how efficiently each repetition can be generated and how many repetitions are used. However the entire cost is still only a small fraction of what multiple attempts at refining a mathematical algorithm would cost. A listing of the program is available upon request.

Acknowledgements—We are indebted to Susan Farwell, Joanne Polak and Susan Krock for data collection, to Colin Begg, Michael Feldstein and Kenneth Stanley for helpful suggestions, and to Gail Manyan and Irene McCammon for typing the manuscript. Part of this work was carried out while one of the authors (J.H.) was still at the Division of Epidemiology and Biostatistics, Sidney Farber Cancer Institute, Boston.

REFERENCES

1. Feinstein AR: Synchronous partition and bivariate evaluation in predictive stratification. **Clin Pharmac Ther** 13: 755–768, 1972
2. McNeil BJ, Hanley JA, Funkenstein HH, Rumbaugh C: No evidence for inappropriate utilisation of CT of the head in a tertiary care hospital. **Radiology** 139: 113–118, April 1981
3. Cox DR: **Analysis of Binary Data.** London: Methuen, 1970
4. Pauker SG, Kassirer JP: The threshold approach to clinical decision making. N **Eng J Med** 302: 1109–1117, 1980
5. Gerson DE, Lewicki AM, McNeil BJ, Abrams HL, Korngold E: The barium enema: evidence for proper utilization. **Radiology** 130: 297–301, 1979
6. Gardner MJ, Barker DJP: A case study in techniques of allocation. **Biometrics** 31: 931–942, 1975
7. Gold RJM, Maag UR, Neal JL, Scriver CR: The use of biochemical data in screening for mutant alleles and in genetic counseling. **Ann Hum Genet** 37: 315–326, 1974
8. Bell RS, Loop JW: The utility and futility of radiographic skull examination for trauma. N **Eng J Med** 284: 236–239, 1971
9. Pozen MW, D'Agostino RB, Mitchell JB, Rosenfeld DM, Guglielmo JT, Schwartz ML, Teebagy N *et al.*: The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. **Ann Int Med** 92: 238–242, 1980
10. Mulley AG, Thibault GE, Hughes RA, Barnett GO, Reder VA, Sherman EL: The course of patients with suspected myocardial infarction: the identification of low risk patients for early transfer from intensive care. N **Eng J Med** 302: 943–948, 1980

11. Brown WB: Prediction analyses for binary data. In **Biostatistics Casebook,** Miller RG *et al.* (Eds). New York: John Wiley 1980
12. Dolgin SM, Schwartz JS, Kressel HY, Soloway RD, Miller WT, Trotman B, Soloway AS, Good LI: Identification of patients with cholesterol or pigment gallstones by discriminant analyses of radiologic features. **N Eng J Med** 304: 808–811, 1981
13. Diehr P, Wood RW, Barr V, Wolcott B, Slay L, Tompkins RK: Ocult headache: presenting symptoms and diagnostic rules to identify patients with tension and migraine headache. **J Chron Dis** 34: 147–158, 1981
14. Hooton TM, Haley RW, Culver DH, White JW, Morgan WM, Carroll RJ: Joint associations of multiple risk factors with the occurrence of nosocomial infection. **Am J Med** 70: 960–970, 1981
15. Lachenbruch P, Mickey R: Estimation of error rates in discriminant analysis. **Technometrics** 10: 1–11, 1968
16. Efron B: Bootstrap methods: another look at the jackknife. **Ann Stat** 7: 1–26, 1979
17. DeSmet AA, Fryback DG, Thornbury Jr: A second look at the utility of radiographic skull examination for trauma. **Am J Roentgenol** 132: 95–99, 1979

APPENDIX

1. *Notation*

N: number of patients in data set

k: number of symptoms used to form symptom complexes

x: a k-variate vector indicating the presence/absence of symptoms in a typical patient

y: a scalar quantity indicating the result of CT examination in a typical patient ($y = 0$ to signify "normal", $y = 1$ to signify "abnormal")

$[x, y]$: a $k + 1$ variate vector formed by concatenating x and y

K: number of separate complexes or patient subgroups ($l \leqslant K \leqslant 2^k$)

$n_1, n_2, \ldots, n_K$: the numbers of patients in each of the K subgroups

x_i: a k-variate vector indicating the presence/absence of symptoms in a patient in subgroup i, $i = 1, 2, \ldots, K$

$y_1, y_2, \ldots, y_K$: the numbers of patients with abnormal CT examinations in each of these subgroups

Y: total number of patients with abnormal CT examinations; $Y = \sum_{i=1}^{K} y_i$

2. *Obtaining the* $\mathrm{n_i}$ *and* $\mathrm{y_i}$ *and forming optimum ranking*

(i) Sort the file of N $[x, y]$'s using the k different symptoms as keys.

(ii) Read the sorted $[x, y]$'s, counting the multiplicity n_i of each unique x pattern and the number of abnormals y_i seen among the n_i. Form the ratio $r_i = y_i/n_i$ for each subgroup. Write the $\{n_i, y_i, r_i\}$ vectors to another file or store in an in-core 3 × K array.

(iii) Sort the K $\{n_i, y_i, r_i\}$ vectors into descending order on the r_i key, to form an optimum ranking of the K subgroups.

3. *Plotting the performance curve*

Plot

$$\frac{\sum_{i=1}^{I} n_{(i)}}{N} \times 100 \quad \text{vs} \quad \frac{\sum_{i=1}^{I} y_{(i)}}{Y} \times 100$$

for $I = 1, 2, \ldots, K$; where $\{[1], [2], \ldots, [K]\}$ denotes a ranking of the K subgroups. The plot using the optimal ranking yields the maximum attainable discrimination (m.a.d.) curve.

4. *Expected variation in the m.a.d. curve for a given proportion* $\mathrm{p_E}$ *of examinations*

Set $N_0 = Np_E$. For each repetition of the experiment, (i) generate $y_1, \ldots, y_K$ by distributing Y successes at random among the

$$N = \sum_{i=1}^{K} n_i$$

patients. (ii) use the ratios $r_i = y_i/n_i$ and step 2(iii) to form an optimum ranking of the K subgroups. (iii) Find the largest I such that

$$\sum_{i=1}^{I} n_{(i)} \leqslant N_0.$$

(iv) Calculate

$$Y_0 = \sum_{i=1}^{I} y_{(i)}.$$

Calculate the maximum attainable percentage of disease detected:

$$p_D = \frac{Y_0}{Y} \times 100$$

and increment the appropriate class frequency of the distribution of p_D by 1.