AUTOMATED DIAGNOSIS IN MULTIPHASIC SCREENING

DANKWARD KODLIN and MORRIS F. COLLEN
PERMANENTE MEDICAL GROUP
OAKLAND, CALIFORNIA

1. Introduction

While efforts at mathematizing the medical diagnostic process continue [10] the current state of the art appears to be characterized by two observations:

(1) in applications where the "correct" diagnosis can be established (for example, by surgery or autopsy) the accuracy of diagnostic algorithms is comparable but not superior to the performance of experts [14], [16];

(2) different analytical techniques give similar results [6], [7], [9].

One is thus tempted to argue that if experts can effectively compete with Bayes' theorem (at least, with that version which assumes the independence of symptom variables [16]) or if experts can weigh the evidence as effectively as a discriminant function, a good case for the exploration of relatively simple decision schemes can be made. This would seem to apply, in particular, to medical areas in which no confirmation of physicians' diagnoses is routinely available and where the major purpose of "automated" diagnosis is to maximize agreement with "routine clinical diagnosis" rather than agreement with "ultimate authority." The so-called multiphasic screening [3] as practiced in the Kaiser Foundation Medical Care Program, provides a typical example; here responses to a battery of several hundred medical questions are recorded in addition to measurements from a standard series of laboratory tests. At the conclusion of such an examination, the patient sees a clinician who reviews the findings and records diagnostic impressions on a check list containing some two hundred diagnoses. This setting is thus quite different from the typical application area of "computer diagnosis" mentioned above where a relatively small set of variables are considered for a small set of mutually exclusive diagnoses as they can be defined in narrow specialty fields. The magnitude of the task inherent in multiphasic screening would seem to make computational simplicity a feature of extreme virtue.

For these reasons, we continue to be interested in diagnostic schemes involving a limited number of dichotomized variables (YES-NO Questions and/or Tests) such as the likelihood ratio method of Neyman [5], [12], [13] and the simple

Partial support came from the National Center for Health Services Research and Development (Grant HS00288) and the Kaiser Foundation Research Institute.

scoring procedure developed in conjunction with the Cornell Medical Index questionnaire by van Woerkom and Brodman [1], [2], [15]. We shall compare the performance of these two methods on an example of considerable medical interest, the diagnosis of coronary heart disease in multiphasic screening, and discuss the utility aspects of various decision alternatives in this context.

2. Material and methods

For some 26,000 persons who took the multiphasic screening examination within one year, we have records on "variables" (about 700 questions and some 50 "objective" tests ranging from X-ray films to chemical body fluid determinations) and "diagnoses," the latter recorded by the "followup" physician. The average number of diagnoses per person is close to two, about half the value recorded in medical settings that deal with the typical "office patient" [2] rather than with persons seeking a health checkup. We envision a computerized diagnostic system that will, for the ith diagnosis or at least for the ith set (if related diagnoses have to be combined) examine a predetermined "relevant variable set" V_i and on the basis of a critical region R_i make a decision concerning that diagnosis. Consequently, a "healthy" person will be one whose response pattern is such that he falls within all negative regions.

Selection of the relevant variable sets is accomplished on a "learning sample," consisting of half the number of cases with diagnosis D_i and a group of 1000 noncases. The remaining cases and another group of 1000 noncases are retained as a "validation sample." With the aid of the learning sample, we select the "best" relevant variable set V_i from an "initial variable list" L_i prepared by clinical judgment. The choice is made on the basis of the likelihood ratio (cases/noncases) for each item on the list L_i, by taking, for practical reasons, the eight "best" variables only. Since tests (T) are much more expensive than questions (Q) a limited number of combinations of "best" Q and T are also made up, such as "best $6Q$ + best $2T$," and so forth.

In the following, we shall concentrate on a particular D_i, the condition coronary heart disease (chd). For this group (made up of the three diagnoses angina pectoris, ischemic heart disease, and myocardial infarction) the initial variable list contained some fifty questions and six tests. On the basis of the learning sample (336 cases and 1000 noncases) the eight best questions are those on chest pain and shortness of breath with likelihood ratios between 5 and 3 and the best two tests appear to be EKG (most abnormalities) and serum cholesterol (upper five percentile) with likelihood ratios of 3.4 and 3. Two analytical methods, the likelihood ratio method θ and the scoring method β are now applied to the learning sample.

2.1. *Likelihood ratio method θ.* Briefly, the method [5], [13] orders the observed response patterns of the eight dichotomized variables by the pattern likelihood ratio θ and investigates the two types of classification errors at various

cutting levels of this array. Following medical custom, we plot sensitivity $1 - \alpha$ and specificity $1 - \beta$ for each level to obtain a performance curve such as Figure 1. Amongst the 256 possible patterns only some 90 occur in our learning sample. When the validation sample (316 cases and 1000 noncases) is classified according to a given positive region derived from the learning sample, some of the remaining patterns appear; these "unknown" patterns are allocated, under current practice, to the positive region. With sample sizes of the magnitude indicated, the performance curve from the validation sample is noticeably, but not excessively, worse than the curve derived from the learning sample.

2.2. *Scoring method* β. The method uses the same learning and validation samples and considers the same variables. While, with β, there is no need, for practical reasons, to be restrictive on the number of variables, we have found that no improvement results from inclusion of seven variables in addition to the eight used with the θ method and for this reason we present results that are identical as to type and number of variables considered.

Each positive response is scored by the relative likelihood deviation

$$s_i = \frac{p_i - P_i}{\sqrt{P_i}}, \tag{2.1}$$

where p_i is the observed frequency of positive response to the ith variable amongst cases and P_i is the observed frequency in the general patient population. The quantity is summed over all positive responses giving the total score

$$\beta = \sum_{\substack{i=1 \\ \text{YES}}}^{\kappa} \frac{p_i - P_i}{\sqrt{P_i}} \tag{2.2}$$

for a person having κ positive (YES) responses. From an empirical distribution of β amongst cases of the learning sample a number of critical β values are taken and a person in the validation sample is diagnosed as CHD if his β value is larger than the critical value under consideration.

The scoring procedure is similar to, but not identical with, that of van Woerkom and Brodman [15], who sum over all positive responses within the whole set of variables (150 questions), while we restrict ourselves to the relevant set V_i.

Since, approximately,

$$s_i = \sqrt{p_i}(\sqrt{\theta_i} - \sqrt{1/\theta_i}), \tag{2.3}$$

where θ_i is the likelihood ratio (cases/noncases) for the ith variable and furthermore since the term in brackets is roughly equal to $\log \theta$ (in the θ range from one to ten), one would expect that a score

$$s_i' = \sqrt{p_i} \log \theta_i, \qquad \lambda = \sum s_i', \tag{2.4}$$

would give results similar to those obtained with β. It also remains to be seen if inclusion of negative responses in β or λ would improve the performance.

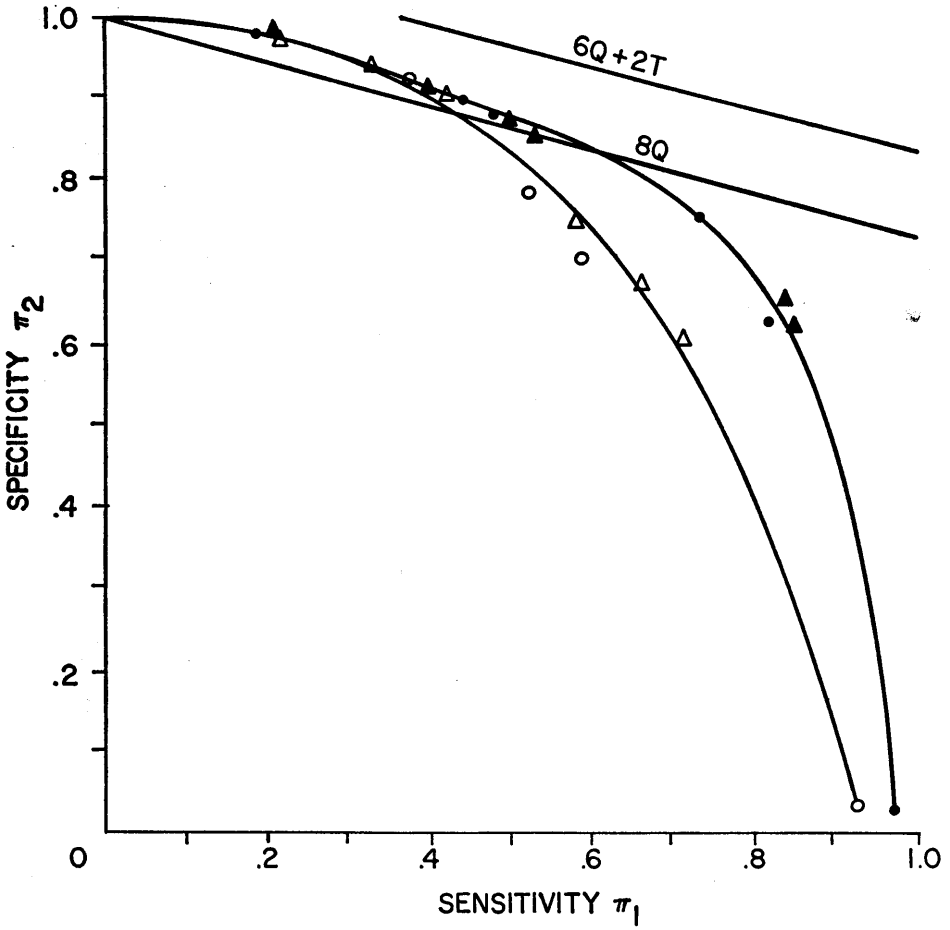

FIGURE 1

Performance of automated diagnosis of coronary heart disease on a validation sample (316 cases and 1000 noncases).
Comparison of two analytical techniques: likelihood ratio method (○, ●) *versus* scoring method (△, ▲).
Comparison of two variable sets: "8 Questions" (○, △) *versus* "6 Questions plus 2 Tests" [EKG and cholesterol] (●, ▲).
Upper straight line is the cost indifference line for "6 Questions plus 2 Tests"; lower straight line is the cost indifference line for "8 Questions."
Note that the methods have essentially the same performance and that, for the cost matrix assumed, only the "8 Questions" procedure would be cost advantageous (against "no screening") and only at a sensitivity of less than 0.5 where the corresponding curve crosses its indifference line.

3. Results

From Figure 1, it appears that θ and β have similar performance and corresponding plots for a number of different disease categories give no indication of consistent differences. Furthermore, the two techniques are comparable as to speed.

Differences between "questions alone" and "questions plus tests" are clearcut, the tests helping to improve specificity at sensitivity levels between 0.5 and 0.9. Even so, the performance is modest in comparison with other diagnostic achievements; for instance, for X-ray diagnosis of tuberculosis [10], a 94 per cent specificity can be achieved at a sensitivity level of 80 per cent. This may merely be an indication that diagnosis of tuberculosis is a relatively easy task. However, we believe that the performance curve for chd can be improved by reviewing the total medical record. The physician's diagnosis, as currently used, comes from a check list employed by him at the time of the followup examination and sometimes is at variance with the medical record. Such a review by an independent observer is time consuming. At least for the case of diabetes, where it has been carried out on a sampling basis, improvement of the automated diagnostic scheme was clearcut.

4. Decision theory

That the screening problem must be cast into the mold of decision theory has been clearly recognized for some time. Flagle [8] has reviewed this aspect of screening in a preceding Symposium and articles on utility and diagnosis are beginning to appear in the medical literature [11]. As Flagle [8] has pointed out, the utilities to be estimated for use in the decision process are *ad hoc*. That is to say, they do not only depend on the obvious factors such as the severity of the disease in question, but on the particular setting in which the subsequent *therapeutic* effort takes place. For this reason, we have decided to consider a cost matrix slightly more elaborate than that underlying the "regret" analysis of Flagle; the matrix in Table I contains the cost for each cell with its three components, the screening cost s, the "workup" or "referral" costs w (that is, costs of additional tests and physician time devoted to a patient declared positive by the screen), and the therapeutic costs T_{tp} for the true positive and T_{fn} for the false negative, both of which have to be carried, at least in part, by a "prepaid" medical care system such as ours.

For the three strategies:

S_1: "do nothing,"
S_2: "refer all,"
S_3: "refer those screened positive,"

we then have the respective costs

$$C_1 = P(c_{21} - c_{22}), \tag{4.1}$$

$$C_2 = P(c_{11} - c_{21}) + (1 - P)w, \tag{4.2}$$

$$C_3 = P(c_{21} - \pi_1 e) + (1 - P)(c_{12} - \pi_2 w), \tag{4.3}$$

where P is the prevalence of the condition, π_1 the sensitivity, and π_2 the specificity, the remaining symbols being cost items defined in Table I.

TABLE I

ASSUMED COST MATRIX FOR CORONARY HEART DISEASE

T: treatment costs for true positive (tp), false positive (fp), false negative (fn), and true negative (tn); w: "workup" or referral cost; s: screening cost.

Note that $e = c_{21} - c_{11} = T_{fn} - T_{tp}$, "pure treatment differential" and $w = c_{12} - c_{22}$; T values of \$100 and 200 are conjectural as is $w = 10$; $s = 1$ applies to the cost of EKG, Collen, Kidd, Feldman, and Cutler [4].

		Physician's diagnosis +	Physician's diagnosis −
Screen	+	$T_{tp} = 100$ $w = 10$ $s = 1$ $c_{11} = 111$	$T_{fp} = 0$ $w = 10$ $s = 1$ $c_{12} = 11$
Screen	−	$T_{fn} = 200$ $w = 10$ $s = 1$ $c_{21} = 211$	$T_{tn} = 0$ $w = 0$ $s = 1$ $c_{22} = 1$

Equating (4.1) and (4.3) and solving for π_2, we obtain

$$\pi_2^* = 1 + \frac{s}{w}\frac{1}{1 - P} - \frac{e}{w}\frac{P}{1 - P}\pi_1, \tag{4.4}$$

the line of indifference between S_1 and S_3.

Equating (4.2) and (4.3) and solving for π_2, we obtain

$$\pi_2^* = \frac{e}{w}\frac{P}{1 - P} + \frac{s}{w}\frac{1}{1 - P} - \frac{e}{w}\frac{P}{1 - P}\pi_1, \tag{4.5}$$

the line of indifference between S_2 and S_3. Equations (4.4) and (4.5) are thus parallel lines, the slope of which is determined by the prevalence P and the ratio e/w. The two intercepts differ by the first term only and a plot of (4.5) quickly eliminates S_2 as a competitive strategy.

In Figure 1, two lines of cost indifference (S_1 *versus* S_3, equation (4.4) have been drawn, the upper one for a screening strategy involving only the cost of EKG with $s = 1$ (see Table I) and the lower one for $s = 0$, assuming that cost

of questions is negligible. Since the performance curve rises only above the question line, but not above the EKG line, one would conclude that screening by questions is cost advantageous while screening by questions and EKG is not. The weak link in this argument is, of course, the uncertainty as to the magnitude of the value of e, the excess treatment cost due to delayed disease recognition.

5. Optimum screening level

Setting the derivative of (4.3) with respect to π_1 equal to zero, one obtains the relation

$$f_0' = -\frac{e}{w}\frac{P}{1-P}, \tag{5.1}$$

where f_0' is the derivative of the performance curve, the empirical function $\pi_2 = f(\pi_1)$, at the minimal cost.

Since (5.1) is identical to the slope of (4.4), one may locate the optimum screening level as the point at which the tangent of the performance curve is parallel to the indifference line. For Figure 1, this optimum is sensitivity π_1 between 0.2 and 0.3.

The expression (5.1) can be used to bring out the value system implied in the choice of particular screening levels [11]. For instance, for the case of X-ray screening for tuberculosis, Lusted [11] shows a preference for the point $\pi_1 = 0.8$ and $\pi_2 = 0.94$. At this point, his performance curve has a slope of -0.2. With a prevalence $P = 5 \times 10^{-4}$ from (5.1) we have $e/w = 400$.

It is interesting to note how much divergence can be produced on casual approach to this question. Rubin, Collen, and Goldman [13], discussing the optimal choice of a screening level, consider the function $U = a\pi_1 + b\pi_2$, and locate the pair π_1, π_2 which maximizes this function. The coefficients a and b are not defined, but from (4.3) it is clear that $a = Pe$ and $b = (1 - P)w$. Thus, maximization of U indeed minimizes the cost of the screening strategy. The writers remark that for "conditions such as tuberculosis a would probably be set several magnitudes higher than b." However, P values of the order of 5×10^{-4} still apply with population estimates of annual incidence rates of newly reported cases standing at 3×10^{-4}. Therefore, if we would take $w = 10$, and thus, $e = 4000$, then $a = 2$ and $b = 10$, indicating that divergence of opinion can be of the order of several magnitudes.

6. Discussion

The cost matrix postulated in Table I can be criticized for understating the benefits of screening. For instance, one may postulate that in the absence of screening facilities, a certain fraction ϕ of healthy persons would come for a "conventional" checkup, perhaps involving the cost w. Under these conditions, S_1, the "do nothing" strategy, would imply the cost

$$C_1 = P(c_{21} - c_{22}) + \phi(1 - P)w, \tag{6.1}$$

giving the indifference line (S_1 *versus* S_2)

$$\pi_2^* = 1 + \frac{s}{w}\frac{1}{1-P} - \phi - \frac{e}{w}\frac{P}{1-P}\pi_1. \tag{6.2}$$

With ϕ values of the order of 0.1, the two indifference lines of Figure 1 would have to be displaced downwards by 0.1. As a consequence of this, screening with $6Q + 2T$ could now be justified on a cost basis for sensitivity levels below 0.5 and screening with $8Q$ for sensitivity levels below 0.65. The optimal screening level would remain at $\pi_1 \doteq 0.3$ at which π_2 is the same for these two alternatives (see Figure 1). In this case, as can be seen from equation (4.3), the "cheap" screen ($8Q$) is still saving the amount s in relation to the "expensive" screen ($6Q + 2T$). Once again, we wish to stress the speculative nature of these impressions, in the face of uncertainty about e/w and current reliance on diagnostic entries, as mentioned in Section 3. In general, we feel, however, that decision theoretical formalisms such as those presented, are useful in the way they point towards crucial items of information that are extremely difficult to acquire.

7. Summary

Multiphasic screening, the application of a large battery of questions and laboratory tests to large numbers of persons, has reached an advanced stage of automation, and rapid classification techniques that provide diagnostic categorization are of considerable interest in this field. With the example of automated diagnosis of coronary heart disease, we find the performance of two methods (likelihood ratio and scoring technique) in terms of characteristic curves to be comparable. Notions of test selection strategy are discussed in relation to the same disease although the precise nature of the cost matrix is speculative.

◇ ◇ ◇ ◇ ◇

The authors wish to acknowledge contributions from Dr. Robert Feldman and programming work by Mrs. L. Lo and Mr. J. Standish.

REFERENCES

[1] K. Brodman and L. S. Goldstein, "The medical data screen. An adjunct for the diagnosis of 100 common diseases," *Arch. Environ. Health*, Vol. 14 (1967), pp. 821–826.

[2] K. Brodman and A. J. van Woerkom, "Computer-aided diagnostic screening for 100 common diseases," *J. Amer. Med. Assoc.*, Vol. 197 (1966), pp. 901–905.

[3] M. F. Collen, "Periodic health examinations using an automated multitest laboratory," *J. Amer. Med. Assoc.*, Vol. 195 (1966), pp. 830–833.

[4] M. F. Collen, P. H. Kidd, R. Feldman, and J. L. Cutler, "Cost analysis of a multiphasic screening program," *New England J. Med.*, Vol. 280 (1969), pp. 1043–1045.

[5] M. F. Collen, L. Rubin, J. Neyman, G. B. Dantzig, R. M. Baer, and A. B. Siegelaub, "Automated multiphasic screening and diagnosis," *Amer. J. Pub. Health*, Vol. 54 (1964), pp. 741–750.

[6] J. M. DICKEY, "Estimation of disease probabilities conditioned on symptom variables," *Math. Biosciences*, Vol. 3 (1968), pp. 249–265.

[7] S. FELDMAN, D. F. KLEIN, and G. HONIGFELD, "A comparison of successive screening and discriminant function techniques in medical taxonomy," *Biometrics*, Vol. 25 (1969), pp. 725–734.

[8] C. D. FLAGLE, "A decision theoretical comparison of three procedures of screening for a single disease," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 4, pp. 887–901.

[9] G. A. GORY and G. O. BARNETT, "Experience with a model of sequential diagnosis," *Comp. Biomed. Res.*, Vol. 1 (1968), pp. 490–507.

[10] J. A. JACQUEZ (editor), *Proceedings of the Second Conference on the Diagnostic Process*, held at the University of Michigan, 1971, Fort Lauderdale, Charles C Thomas, in press.

[11] L. B. LUSTED, "Decision-making studies in patient management," *New England J. Med.*, Vol. 284 (1971), pp. 416–424.

[12] J. NEYMAN, *First Course in Probability and Statistics*, New York, Holt, 1950, Chapter 5.

[13] L. RUBIN, M. F. COLLEN, and G. E. GOLDMAN, "Frequency decision theoretical approach to automated medical diagnosis," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1967, Vol. 4, pp. 867–886.

[14] A. W. TEMPLETON, K. BRYAN, R. WAIRD, J. TOWNES, M. HUQUE, and S. J. DWYER, "Computer diagnosis and discriminate analysis decision schemes," *Radiology*, Vol. 95 (1970), pp. 47–55.

[15] A. J. VAN WOERKOM and K. BRODMAN, "Statistics for a diagnostic model," *Biometrics*, Vol. 17 (1961), pp. 299–318.

[16] H. R. WARNER, A. F. TORONTO, and L. G. VEASY, "Experience with Bayes' theorem for computer diagnosis of congenital heart disease," *Ann. N.Y. Acad. Sci.*, Vol. 115 (1964), pp. 558–567.