International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease

Robert Detrano, MD, PhD, Andras Janosi, MD, Walter Steinbrunn, MD, Matthias Pfisterer, MD, Johann-Jakob Schmid, DE, Sarbjit Sandhu, MD, Kern H. Guppy, PhD, Stella Lee, MS, and Victor Froelicher, MD

A new discriminant function model for estimating probabilities of angiographic coronary disease was tested for reliability and clinical utility in 3 patient test groups. This model, derived from the clinical and noninvasive test results of 303 patients undergoing angiography at the Cleveland Clinic in Cleveland, Ohio, was applied to a group of 425 patients undergoing angiography at the Hungarian Institute of Cardiology in Budapest, Hungary (disease prevalence 38%); 200 patients undergoing angiography at the Veterans Administration Medical Center in Long Beach, California (disease prevalence 75%); and 143 such patients from the University Hospitals in Zurich and Basel, Switzerland (disease prevalence 84%). The probabilities that resulted from the application of the Cleveland algorithm were compared with those derived by applying a Bayesian algorithm derived from published medical studies called CADENZA to the same 3 patient test groups. Both algorithms overpredicted the probability of disease at the Hungarian and American centers. Overprediction was more pronounced with the use of CADENZA (average overestimation 16 vs 10% and 11 vs 5%, p <0.001). In the Swiss group, the discriminant function underestimated (by 7%) and CADENZA slightly overestimated (by 2%) disease probability. Clinical utility, assessed as the percentage of patients correctly classified, was modestly superior for the new discriminant function as compared with CADENZA in the Hungarian group and similar in the American and Swiss groups. It was concluded that coronary disease probabilities derived from discriminant functions are reliable and clinically useful when applied to patients with chest pain syndromes and intermediate disease prevalence.

(Am J Cardiol 1989;64:304–310)

Despite Osler's axiom that "medicine is a science of uncertainty and an art of probability,"[1] the applicability of probability analysis to the diagnosis of common diseases is still uncertain. Part of this uncertainty is due to the difficulty in obtaining clinical data from very large numbers of patients. Such data are needed to derive accurate probability models that could be applied universally. Diamond et al[2] were the first to circumvent this paucity of large data collections. By calculating weighted averages of sensitivities and specificities obtained from a review of published studies on diagnostic testing, and by using tables of pretest probabilities also from those studies, they pooled knowledge obtained from several thousand patients to define a probability algorithm for the diagnosis of coronary artery disease (CAD). Because published reports rarely give complete distributions of clinical and test variables, the investigators had to assume that these variables were independent of one another. This assumption of independence created errors in their predictions. Furthermore, the reported sensitivities and specificities are affected by bias and other methodologic problems[3-5] that may actually lead to further errors when probabilities are based on them. We undertook this project to determine if a probability algorithm derived from the clinical and test characteristics of a relatively small group of 303 patients could accurately predict CAD probabilities in study samples drawn from various ethnic populations with different clinical characteristics. This algorithm was compared with the algorithm of Diamond et al.[2]

METHODS

Reference group for derivation of the probability model: The reference group used to derive the model consisted of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic between May 1981 and September 1984. No patient had a history or electrocardiographic evidence of prior myocardial infarction or known valvular or cardiomyopathic disease. All 303 patients provided a history and underwent physical examination, electrocardiogram at rest, serum cholesterol determination and fasting blood sugar determination as part of their routine evaluation. Historical data were recorded and coded without knowledge of noninvasive or angiographic test data. In addition, after giving informed consent, the patients underwent 3 noninvasive tests as part of a research protocol. The results of these tests (exercise electrocardiogram, thallium scintigraphy and cardiac fluoroscopy) were not interpreted until after the coronary angiograms had been read. These tests were analyzed and the results recorded without knowledge of the historical or angiographic results. Work-up bias was therefore not present.[3] The mean age of these patients was 54 years; 206 were men. Angiograms were interpreted by a cardiologist without knowledge of other test data. Further details of this data collection are described elsewhere.[6]

Clinical and test variables: The 4 clinical variables were age, sex, chest pain type (typical anginal, atypical anginal, nonanginal, asymptomatic[6]) and systolic blood pressure. Routine test data collected included serum cholesterol, fasting blood sugar >120 mg/dl and electrocardiographic results at rest (classified as (1) normal; (2) ST-T-wave abnormality [T-wave inversions or ST depression >0.05 mV or both]; or (3) probable or definite left ventricular hypertrophy by Estes' criteria).

Noninvasive tests were exercise electrocardiogram, exercise thallium scintigraphy and fluoroscopy for coronary calcium. Exercise data collected included maximal heart rate, exercise-induced angina, slope of the peak exercise ST segment (upsloping, flat or downsloping), exercise-induced ST-segment depression (where 1 mm = 0.1 mV), exercise thallium scintigraphic defects (fixed, reversible or none). The fluoroscopic data consisted of the number of major vessels that appeared to contain calcium.[7] Data for all of these 13 variables were entered into a computerized database.

Derivation of the algorithm: The algorithm was derived by applying logistic regression to the 13 clinical and test variables against the angiographic variable of the presence or absence of a >50% diameter narrowing (dependent variable). Our object was to make the algorithm relevant to clinical situations in which data might be present for only certain combinations of the 13 variables. Derivation of the most complete algorithm would require applying logistic regression to all 8,191 possible combinations of these 13 variables. To simplify this calculation, only "clinically relevant" combinations were allowed. Combinations were considered clinically relevant if they included age, sex and chest pain type, and,

in the case when exercise electrocardiography was performed, maximal heart rate and exercise-induced angina. Because ST-segment depressions induced by exercise are sometimes difficult to interpret, clinically relevant combinations do not necessarily include them. However, when the slope of the ST segment could be calculated, the actual depression also was required. In all, there were 352 such relevant combinations and, therefore, 352 logistic regression calculations were performed.

Testing the algorithm: The 352 subsets of regression coefficients were stored and indexed in a computer data file. A computer program was written to read clinical and test data of test patients and match the data available with the appropriate subset of coefficients from this data file. The program then computes the ith patient's disease probability, P_i, using the formula: $P_i = e^{f_i}/(1 + e^{f_i})$, where f_i is the linear combination of this patient's data, using the appropriate subset of coefficients.

Bayesian method: The Bayesian method based on published medical studies can be applied using a computer program called CADENZA.[8] We obtained this program from its authors and transformed all test patient data to fit the input documentation of this computer program. The program uses weighted averages of sensitivities and specificities from published medical works. The pretest probabilities in the program also were derived from published studies.[9] The sensitivities, specificities and pretest probabilities are sequentially substituted in Bayes' theorem. The resultant equation is applied first to pretest probabilities based only on clinical and routine test data (age, sex, chest pain type, history of diabetes and hypertension, serum cholesterol, electrocardiogram at rest and so on) and then consecutively to resulting posttest probabilities from the previous calculation.[10]

Test group data: Three test groups were analyzed using the 13-variable discriminant function and the Bayesian algorithm. These groups included subjects without a prior catheterization or evidence of prior infarction or valve disease, whose CAD status was therefore unknown but whose angiograms were performed to determine the presence and severity of disease. These test groups were (1) 200 patients at the Veterans Administration Medical Center in Long Beach, California; (2) 425 patients at the Hungarian Institute of Cardiology; and (3) 143 patients at 2 Swiss university hospitals.

In these test groups, noninvasive test results were not withheld from the treating physician, and might have influenced the decision to perform coronary angiography; therefore, workup bias or angiographic referral bias was present.[3] Clinical data and noninvasive test results were recorded before (and therefore without knowledge of) the coronary angiography results.

All available data that could be used by either algorithm were collected from the patient records for the test groups. The data included age, sex, chest pain characteristics (as seen earlier), systolic blood pressure at rest, history of hypertension, smoking history, history of

TABLE I Clinical Characteristics of the Study Groups

Study Group	Mean Age (yrs)	Men (%)	Angina (%)	Disease* (%)	MVD† (%)	SBP‡
Long Beach (n = 200)	59	97	66	74	62	135
Hungary (n = 425)	48	71	43	38	63	132
Basel (n = 85)	55	86	73	85	69	139
Zurich (n = 58)	54	93	74	74	40	119
Cleveland (n = 303)	54	68	53	46	60	132

* Disease is defined as >50% diameter narrowing; † MVD = multivessel disease defined by >50% diameter narrowing in >1 vessel; ‡ SBP = mean systolic blood pressure.

TABLE II Sensitivities and Specificities of Exercise Testing in the Study Groups

	1-mm ST Depression		Exercise Angina		Thallium Defect	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Long Beach (%)	54 (81/149)	73 (37/51)	72 (83/115)	63 (20/32)	90 (27/30)	45 (5/11)
Hungary (%)	67 (109/162)	86 (225/263)	59 (95/162)	89 (233/263)	94 (15/16)	59 (10/17)
Basel (%)	29 (21/72)	85 (11/13)	56 (40/71)	69 (9/13)	70 (50/71)	54 (7/13)
Zurich (%)	53 (23/43)	67 (10/15)	30 (13/43)	80 (12/15)	1 patient	1 patient
Cleveland (%)	66 (92/139)	73 (119/164)	55 (76/139)	86 (144/164)	73 (101/138)	79 (129/163)

diabetes, family history, electrocardiogram at rest, serum cholesterol and fasting blood sugar. The exercise variables included medications at the time of the exercise test, duration of the exercise test, peak achieved heart rate, heart rate at rest, peak exercise systolic blood pressure, exercise-induced angina or hypotension or both, exercise-induced ST depression relative to rest, exercise-induced ST slope, exercise-induced R-wave change, radionuclide ejection fraction and wall motion abnormalities at rest and during exercise, and exercise thallium results (as seen earlier). Coronary angiograms were considered abnormal if there was >50% luminal narrowing of any major epicardial vessel. Histories, physical examinations and all noninvasive tests were performed within 6 weeks before the date of the coronary angiogram. A description of the individual test groups follows.

Long Beach Veterans Administration Medical Center: This group was drawn from all consecutive subjects undergoing cardiac catheterization at the Veterans Administration Medical Center in Long Beach between 1984 and 1987. After excluding those with prior infarction, valvular disease and prior catheterization, there were 200 test group subjects.

Hungarian Institute of Cardiology: This group was drawn from all patients undergoing catheterization at the Hungarian Institute of Cardiology in Budapest between 1983 and 1987. Patients with prior infarction or valvular disease were excluded. The remaining 425 subjects made up the Hungarian test group.

Swiss universities: This group was drawn from all subjects undergoing cardiac catheterization at the university hospitals in Zurich and Basel, Switzerland, in 1985. The aforementioned exclusion criteria were applied. Of the 143 Swiss patients, 58 underwent catheterization in Zurich and 85 in Basel.

The clinical characteristics of the subjects in the 3 test groups and the rationalization for combining the groups from the Swiss universities is given in the Results section. The observed sensitivities and specificities of the exercise electrocardiogram and exercise thallium scintigraphy also are given in the Results section.

Evaluation of probabilities: The reliability of a probability estimate reflects its numerical proximity to the actual disease prevalence in subjects with similar clinical and test data. If the estimates are reliable, the mean disease probability in a test group will be the same as the disease prevalence in the group. Thus, by subtracting the prevalence of disease from the mean or expected probabilities, and dividing this difference by the standard deviation, we get an overestimation index that is a measure of how much a model over- or underpredicts disease probability. Because this is a mean of an assumedly normally distributed difference divided by its standard deviation,[11] comparison of models is simplified. This can be done using the Student t test with a standard deviation of 1.0.

Finer detail can be obtained by sorting the probability estimates in ascending order and then dividing them into quintiles of probability.[12] The expected probabili-

| | Thresholds | | | | | |
| | 0.4 | | 0.5 | | 0.6 | |
	CDF	CADENZA	CDF	CADENZA	CDF	CADENZA
Hungarian	74 ± 2	71 ± 2	77 ± 2[†]	74 ± 2	82 ± 2[†]	76 ± 2
Only those 0.2–0.8	58 ± 3	56 ± 3	64 ± 3	61 ± 3	73 ± 3[†]	64 ± 3
Long Beach, California	78 ± 3	77 ± 3	79 ± 3	77 ± 3	79 ± 3[†]	75 ± 3
Only those 0.2–0.8	64 ± 5	61 ± 5	66 ± 5	60 ± 5	65 ± 5[†]	57 ± 5
Swiss	82 ± 3	82 ± 3	81 ± 3	81 ± 3	78 ± 3	79 ± 3
Only those 0.2–0.8	73 ± 6	77 ± 5	70 ± 6	73 ± 6	66 ± 6	70 ± 6

* Percent ± standard error of percent; [†] p <0.05 vs CADENZA (McNemar's test).
CDF = Cleveland discriminant function.

ties in each quintile are compared with the prevalences in that quintile and the differences computed. These differences are a measure of the overestimation per quintile, and reflect the reliability of low, intermediate and high estimates.

A probability estimate will be clinically useful if it accurately classifies patients as diseased or not diseased. A probability algorithm will be useful if its probability estimates are clinically useful over an appropriate range of probability thresholds.

We agreed that the most relevant probability thresholds for making decisions concerning angiography or therapy lie between 0.20 and 0.80 for subjects with chest pain syndromes. Therefore, the percentage of correct classifications was calculated over this range (0.2 < p < 0.8) for both algorithms in the 3 test groups.

Subjects whose clinical and test data are concordant will generally have very high or very low probability estimates from any algorithm. We agreed that a clinician would probably not need a probability estimate for clinical decision making in these cases. Such estimates would instead be most useful for cases where patient data are discordant. These patients would have intermediate probability estimates by most algorithms; therefore, the percent correct classification rate was recalculated ignoring all subjects for whom probability estimates from both the discriminant function and CADENZA were out of the range (0.2 < p < 0.8).

To compare the correct classification rates for the 2 algorithms, we used the McNemar's test.

RESULTS

Table I lists the demographic and clinical characteristics of the various patient study groups. Table II lists the sensitivities and specificities of 1 mm = 0.1 mV exercise-induced ST depression, exercise-induced angina pectoris and an abnormal thallium scintigram (fixed or reversible defect or both).

The 2 Swiss groups are very similar with respect to age, sex and symptoms. Because of their small size and their similarity, they were combined into a single group.

Reliability: Overestimation indexes for the probability estimates were significantly higher for CADENZA than for the discriminant function in the American test group (6.1 vs 2.0, p <0.001) and in the Hungarian group (10.4 vs 5.6, p <0.001). In the Swiss group, CADENZA slightly overestimated disease probability, whereas the discriminant function underestimated it. Figure 1A, B and C shows that both models tend to overestimate intermediate disease probabilities (second, third and fourth quintiles), with CADENZA causing the most overestimation. At the Swiss universities, where disease prevalence was highest, both models underestimated low probabilities. In all but the second quintile of this group, the probability estimates derived by CADENZA had a larger absolute error than those obtained using the discriminant function.

Clinical utility: The percentage of patients who are correctly classified will depend on the accuracy of the model and on the overall disease prevalence in the test group. Models that overestimate disease will cause more erroneous diagnoses at low prevalence, but will classify patients correctly at high prevalence. Figure 2A, B and C bears this out. The Cleveland discriminant function more accurately classified patients at the Hungarian Institute, where there was a low disease prevalence. These differences were statistically significant between probability thresholds of 0.4 and 0.7 (p <0.05). In the American group, where disease prevalence was higher, the discriminant function also classified patients more correctly than did CADENZA, but the difference was not significant except at thresholds around 0.6. In the Swiss group, which had the highest disease prevalence, CADENZA resulted in a significantly higher rate of correct classification at thresholds of at least 0.70 (p <0.05).

It is interesting to compare the percentage of correct classifications for the 2 algorithms in the 3 groups at specific thresholds. Table III is such a comparison. Thresholds of 0.4, 0.5 and 0.6 were used because we thought them to be appropriate for many clinical decisions. The table gives the percentage of correct classifi-

cations by the 2 algorithms (1) when all patients are included; and (2) when those for whom both algorithms produced very low (≤0.2) or very high (≥0.8) probabilities are excluded. Excluding these latter subjects may be appropriate because this exclusion leaves primarily those patients with discordant results for whom clinical

decisions are more difficult and the use of a probability algorithm is more relevant. Both algorithms performed less well when these exclusions were made. The discriminant function performed moderately better than CADENZA, both with and without the exclusion of subjects with extreme probabilities.
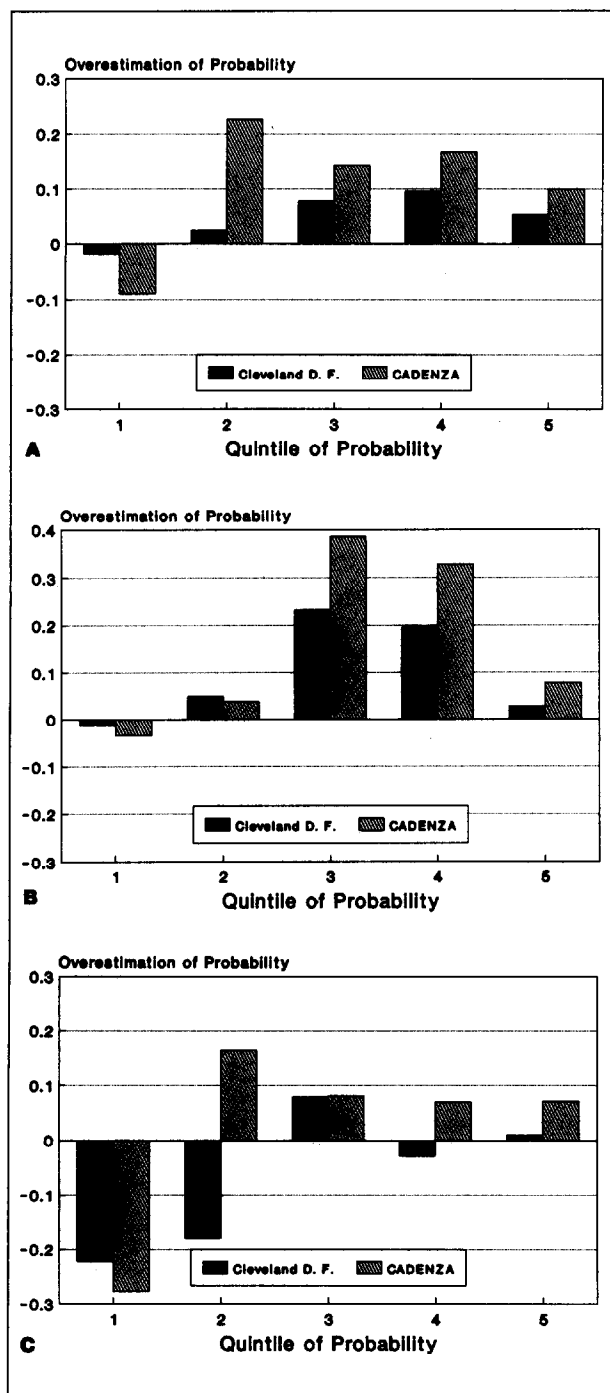

FIGURE 1. Overestimation by quintiles for the American (A), Hungarian (B) and Swiss (C) test groups. Bar heights are calculated by subtracting disease prevalence from the average estimated probabilities in each quintile.
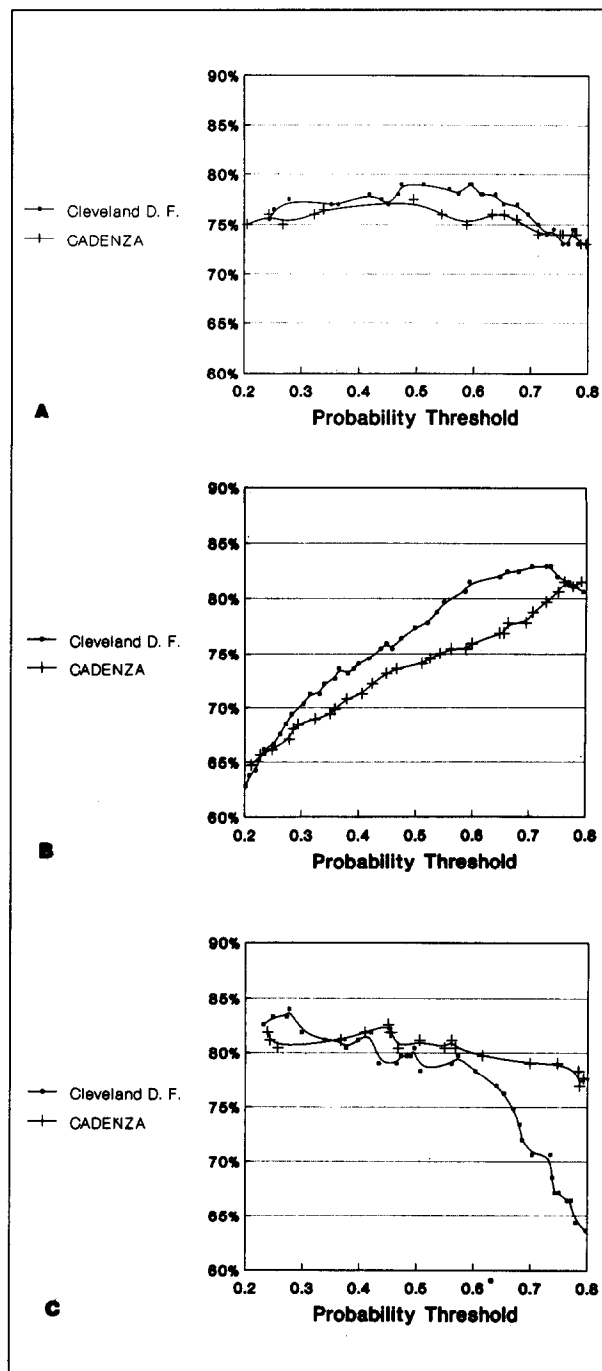
FIGURE 2. Percentage of patients correctly classified versus probability thresholds in the American (A), Hungarian (B) and Swiss (C) test groups. Vertical distances between the curves are significant only for thresholds near 0.6 (American), 0.5 to 0.7 (Hungarian) and near 0.7 (Swiss).

DISCUSSION

The results reported herein illustrate the comparative accuracies of probabilities derived by applying a discriminant function based on clinical and test data from a relatively small group of patients and those derived from a Bayesian algorithm based on published medical studies. The superiority of the discriminant function is most clear at an institution (Hungarian Institute of Cardiology) in which the disease prevalence in an angiographic cohort was relatively low (38%). Its superiority is less clear in the American group, in which the more accurate probabilities of the discriminant function did not produce a significantly higher number of correct diagnoses. In the Swiss group, in which the disease prevalence was highest, the CADENZA algorithm produced a higher percentage of correctly classified patients than did the discriminant function, though its overall reliability was similar.

Cornfield et al[13] noted that the interdependency of symptoms generates overconfidence when algorithms based on published medical studies are used, such as the Bayesian algorithm we used here for comparison purposes. This overconfidence is reflected in the over- and underestimation of disease probability that is evident in Figure 1A, B and C.

There is another potential source of overconfidence in the algorithm based on published studies, namely the bias implicit in the conduction, reporting, review and publication of the results of research studies on diagnostic testing.[14] Sensitivities and specificities that are high are more likely to be published than those that are low. When these are applied in Bayes' theorem to subjects in laboratories in which falsely positive and negative results are more common, the resulting after-test probabilities will be too high or too low, as reflected in our results.

Limitations of the method: Because the Cleveland discriminant function was derived from a population referred for angiography, its best performance is expected in such a population. Clinicians select patients for angiography because their symptoms and test results merit this procedure. This bias is present in the great majority of reports on diagnostic testing[3,15,16] and is, in our opinion, impossible to avoid. Aside from the angiographic result, the only other referent standard that has been proposed is the use of epidemiologic endpoints.[17] These have been justly criticized as being inadequate[17,18] due to brief periods of follow-up. For these reasons, we consider angiographic populations suitable for evaluating testing and algorithms such as the one presented here.

When the discriminant function derived in this research is compared with the algorithm derived from medical studies, its improved accuracy is most evident in the Hungarian test group, which had the lowest disease prevalence (38%). In the Swiss test group with high disease prevalence (81%), the accuracy of the discriminant function is similar to the accuracy of the standard algorithm. The results of this investigation would support the use of our discriminant function for predict-

ing the disease probability of subjects in whom the pretest probability is between 0.2 and 0.7. The algorithms based on published studies can be accurately applied when the pretest probability is between 0.7 and 1.0. Such tactics would reduce the number of unnecessary normal coronary angiograms with fewer "missed" patients with severe disease. However, we caution against the application of such algorithms in subjects with very low probability of disease, such as those seen in screening clinics.

As in all previous reports on this topic, angiograms were read by visual assessment of the percent diameter narrowing. Although this was done without knowledge of test results, it is likely that the clinical angiographers making these assessments were aware of clinical symptoms, and it is even more probable that they knew the ages and sexes of their patients. Although visual assessments have well-known drawbacks,[19] they are standard practice in most institutions and were used in the groups from which both models were derived. It was therefore decided that they presented a more appropriate referent standard than more refined endpoints, such as quantitative coronary angiography.

The algorithm derived from medical studies had the apparent advantage of having more data available to it. Despite this, it did not outperform the discriminant function in any of the 3 test groups. We[20] and others[21] have noted that Bayesian algorithms tend toward greater overestimation when more data become available to them. This paradoxic increase in error may account for some of the differences between the 2 algorithms.

REFERENCES

1. Osler W. In: Dean WB, ed. Aphorisms from His Beside Teachings and Writings. New York: Henry Shuman, 1950:125.
2. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. N Engl J Med 1979;300:1350–1358.
3. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercised testing for coronary artery disease: groups, analysis and bias. Am J Cardiol 1980;46:807–812.
4. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982;248:2467–2470.
5. Hlatky HA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati R. Factors affecting sensitivity and specificity of exercise electrocardiography: multivariable analysis. Am J Med 1984;77:64–71.
6. Detrano R, Yiannikas J, Salcedo EE, Rincon G, Go RT, Williams G, Leatherman J. Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. Circulation 1984;69:541–547.
7. Detrano R, Salcedo EE, Hobbs RE, Yiannikas J. Cardiac cinefluoroscopy as an inexpensive aid in the diagnosis of coronary artery disease. Am J Cardiol 1986;57:1041–1046.
8. Diamond GA, Staniloff HM, Forrester JS, Pollock BH, Swan HJC. Computer-assisted diagnosis in the noninvasive evaluation of patients with suspected coronary artery disease. JACC 1983;1:444–455.
9. Diamond GA, Forrester JS, Hirsch M, Berman DS, Vas R, Swan HJC, Staniloff HM. Application of probability analysis to the clinical diagnosis of coronary artery disease. J Clin Invest 1980;65:1210–1221.
10. Diamond GA, Pollock BH, Work JW. Clinical decisions and computers. JACC 1987;9:1385–1396.
11. Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med 1978;17:227–237.

12. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fed Proc 1962;21(suppl II):11.*

13. Cornfield J, Dunn RA, Batchlor CD, Pipberger H. Multigroup diagnosis of electrocardiograms. *Comput Biomed Res 1973;6:97–120.*

14. Diamond GA. Monkey business. *Am J Cardiol 1986;57:471–475.*

15. Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froelicher V. Factors affecting the sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. *Am J Med 1988;84:699–710.*

16. Gianrossi R, Detrano R, Mulvihill D, Lehmann K, Dubach P, Colombo A, McArthur D, Froelicher V. Exercise-induced ST depression in the diagnosis of coronary artery disease: a meta-analysis. *Circulation 1989, in press.*

17. Detrano R, Froelicher V. A logical approach to screening for coronary artery disease. *Ann Intern Med 1987;106:846–852.*

18. Detrano R, Nourijah P, Froelicher V. Screening for coronary artery disease. *Ann Intern Med 1987;107:594–595.*

19. Zir LM, Miller SW, Dinsmore RE, Gilbert JP, Harthorne JW. Interobserver variability in coronary angiography. *Circulation 1976;53:627–632.*

20. Detrano R, Guppy K, Abbassi N, Janosi A, Sandhu S, Froelicher V. Reliability of Bayesian probability analysis for predicting coronary artery disease in a veterans hospital. *J Clin Epidemiol 1988;41:599–605.*

21. Russek E, Kronmal RA, Fisher LD: The effect of assuming independence in applying Bayes' theorem to risk estimation and classification in diagnosis. *Comput Biomed Res 1983;16:537–552.*