J Clin Epidemiol Vol. 42, No. 10, pp. 983–986, 1989
Printed in Great Britain.

0895-4356/89 $3.00 + 0.00

II-7

ACCURACY CURVES: AN ALTERNATIVE GRAPHICAL REPRESENTATION OF PROBABILITY DATA

ROBERT DETRANO

Cardiology 111-C, V. A. Medical Center, 5901 East Seventh Street, Long Beach, CA 90822, U.S.A.

(*Received in revised form* 6 *March* 1989)

Abstract—Receiver operating characteristic (ROC) curves have been frequently used to compare probability models applied to medical problems. Though the curves are a measure of the discriminatory power of a model, they do not reflect the model's accuracy. A supplementary accuracy curve is derived which will be coincident with the ROC curve if the model is reliable, will be above the ROC curve if the model's probabilities are too high or below if they are too low. A clinical example of this new graphical presentation is given.

ROC Probabilistic diagnosis Accuracy Bayes' theorem

INTRODUCTION

Receiver operating characteristic (ROC) curves are popular for evaluating the diagnostic accuracy of probability models [1–3]. Curves are constructed by using probability estimate cutpoints to calculate sensitivities and false positive rates (1 – specificity). Often two mathematical models are compared by plotting the curves for both and comparing the areas under each of them. When these areas are equal, the models are assumed to be equally accurate. If the area under one curve is greater than that under the other, the model with the larger area is claimed to be more accurate [1].

The above formulation has several pitfalls. This article deals with only one of them and suggests an improved graphical presentation of data from probability models.

ROC CURVES: PART OF THE STORY

A disease d is either present or it is not. An "exact" diagnostic standard is available but is considered too expensive or too risky to be universally applied. Therefore, imperfect tests and data from the patient's history and physical examination must be used to estimate the probability p of disease d. Thus, a probability model f is used to estimate p:

$$p = f\{x\}$$

where x represents the clinical and test information.

Given the same information $\{x\}$, two models f_a and f_b are to be compared using a group of subjects whose disease status is known.

$$p_a = f_a\{x\}$$

$$p_b = f_b\{x\}.$$

The size of the group (the number of subjects) is

$$T = D + N$$

where D is the number of patients with disease and N is the number without disease. The sensitivity and false positive rate of each model f_a and f_b are defined for given cutpoints p'_a and p'_b:

$$\mathrm{Se}_a = D_{p>p'}/D \qquad \mathrm{Se}_b = D_{p>p'}/D$$

$$\mathrm{FP}_a = N_{p>p'}/N \qquad \mathrm{FP}_b = N_{p>p'}/N.$$

By plotting sensitivity against false positive rate for the range of cutpoints between 0 and 1, an ROC curve is generated.

If the sensitivity of one model f_a is higher than that of the other, f_b for the same false positive rate, then f_a is better than f_b in discriminating diseased from nondiseased. It is often claimed that f_a is therefore more accurate than f_b, but this may not be true if accuracy is considered as the numerical proximity of the probability estimates to the actual disease prevalence in subjects with similar attributes. An example will show this.

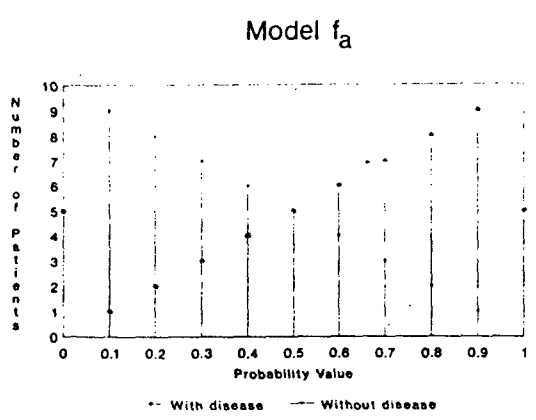
Table 1. Sensitivities and false positive rates derived from distributions of probabilities illustrated in Fig. 1

Se(f_a)	FP(f_a)	Se(f_b)	FP(f_b)
0.98	0.72	1.0	0.72
0.94	0.56	0.98	0.56
0.88	0.42	0.94	0.42
0.80	0.30	0.88	0.30
0.70	0.20	0.80	0.20
0.58	0.12	0.70	0.12
0.44	0.06	0.58	0.06
0.28	0.02	0.44	0.02

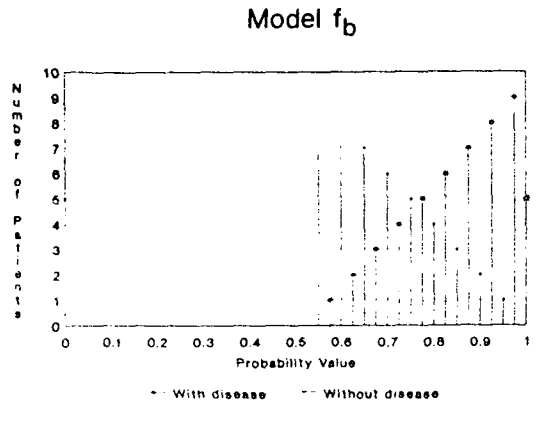
Example

Figure 1 shows the distribution of probabilities assigned to 100 subjects by two models f_a and f_b. Table 1 gives the sensitivities and false positive rates chosen so that $FP_a = FP_b$. Figure 2 shows the ROC curves obtained by plotting the data in Table 1.

Table 1 and Fig. 2 show that for most points of equal false positive rate, the sensitivity of f_b is higher than the sensitivity of f_a. If one ignores the original distributions in Fig. 1, f_b appears more accurate than f_a. A glance at Fig. 1, however is enough to show why this is not so. Almost all subjects without disease are given probabilities higher than 0.5 by model f_b. A physician using model f_b and a threshold probability of 0.45 to decide on an intervention would make erroneous decisions 45% of the time, while a physician using model f_a and the same cutpoint would be wrong only 25% of the time. Clearly, the ROC curves tell only part of the story.



(a)



(b)

Fig. 1. Distribution of probability estimates produced by two models.

THE WHOLE STORY: ACCURACY CURVES

Let us define the expected probability on an interval Δp for model f_a as:

$$\langle p_a \rangle_{\Delta p} = \sum_{T}^{T_{\Delta p}\, i=1\, p_{ai}} \Delta p.$$

Here p_{ai} is the probability estimate of model f_a for patient i and $T_{\Delta p}$ is the number of patients whose probability estimated by f_a is on the interval Δp. The summation is over the same interval Δp. The expected probability is thus the average probability according to the model f_a over an interval. Let us define accuracy as the satisfaction of the equality:

$$\langle p_a \rangle_{\Delta p} = D_{\Delta p}/T_{\Delta p} \qquad (1)$$

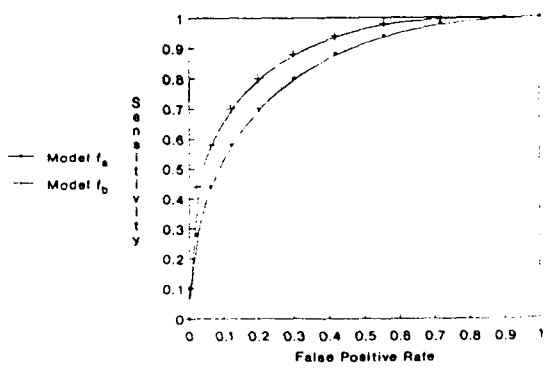


Fig. 2. ROC curves for models whose probability estimates are shown in Fig. 1.

for all intervals Δp. Equation (1) can be satisfied if and only if

$$\langle p_a \rangle_{p>p'} = D_{p>p'}/T_{p>p'} \tag{2}$$

for any p'. The motivated reader can easily prove this to be the case. Multiplying equation (2) by $T_{p>p'}$ and dividing by D gives:

$$\frac{\sum_{i=1}^{T_{p>p'}} p_{ai}}{D} = D_{p>p'}/D. \tag{3}$$

The summation is over all cases with $p_{ai} > p'$. The right side of this equation is just the sensitivity of model f_a at cutpoint p'. So:

$$\frac{\sum_{i=1}^{T_{p>p'}} p_{ai}}{D} = \mathrm{Se}_a. \tag{4}$$

A similar and analagous equation can be derived for the false positive rate:

$$\frac{\sum_{i=1}^{T_{p>p'}} (1 - p_{ai})}{N} = \mathrm{FP}_a. \tag{4a}$$

Model f_a can be considered accurate only if both equations (4) and (4a) are satisfied. When this is the case, a plot of the left side of equation (4) against the left side of equation (4a) will be coincident with the ROC curve for f_a. By plotting these quantities on the same axes as the ROC curve we can demonstrate the model's accuracy. Unlike the ROC curve, the accuracy curve can exceed the boundaries of 0 and 1 if the underlying model is not accurate.

A CLINICAL EXAMPLE

Both Bayesian algorithms assuming independence [4, 5] and discriminant functions derived using logistic regression [6, 7] have been used to estimate the probability of coronary artery disease. We have shown that when the latter are derived from the same population, they produce probabilities that fit the observed disease prevalence more closely than do the Bayesian models derived partly from the literature on diagnostic testing for coronary disease and partly from the test population [8]. Specifically, we have tested two models. The first is obtained by the sequential application of Bayes' formula:

$$p_i = \frac{\mathrm{Se}_i\, p_{i-1}}{\mathrm{Se}_i\, p_{i-1} + \mathrm{FP}_i(1 - p_{i-1})} \tag{5}$$

This was applied using an *a priori* probability p_0 obtained from literature tables of disease prevalence in subjects of different ages and genders and with various types of chest pain. The sensitivities and false positive rate (Se_i and FP_i) of the exercise electrocardiogram, exercise thallium scintigraphy and fluoroscopy were obtained from a random sample of 162 subjects (training set). These and the *a priori* probabilities p_0 were applied in equation (5) to calculate the post-test probabilities of another random sample of 141 subjects from the same population (test set). For comparison, logistic regression was applied to the training set and a discriminant function of the form

$$p_j = \frac{e^{fj}}{1 + e^{fj}} \tag{6}$$

$$f_j = \sum_{i=1}^{6} C_{ij} X_{ij} \tag{7}$$

was obtained and applied to the test set. The coefficients C_{ij} represent the various weights of the variables of age, sex, type of chest pain, exercise ECG result, thallium scintigraphy result, and fluoroscopy result.

ROC curves and accuracy curves were constructed from the Bayesian and discriminant function probabilities of the test set. These are illustrated in Fig. 3. Note that the two ROC curves are almost coincident. We have shown that the curves are not significantly different from one another [8]. The accuracy curve for the discriminant function is coincident with its ROC curve, indicating that this is an accurate model. The accuracy curve for the Bayesian method lies apart from and higher than its ROC

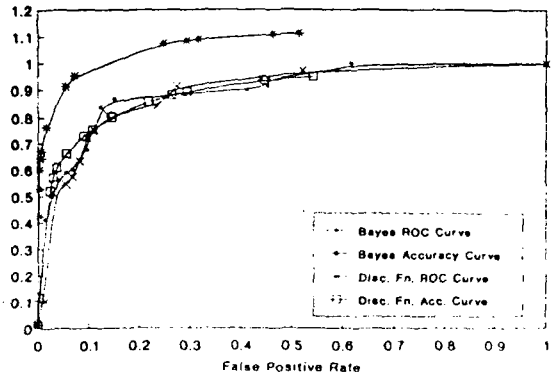


Fig. 3. ROC and accuracy curves (AC) for Bayesian and discriminant function models when applied to clinical test group of 141 subjects with suspected coronary artery disease. Vertical axis represents sensitivity for ROC curve

curve, showing that this model overestimates disease probabilities.

Using a different approach, we have already demonstrated that the discriminant function does fit the observed data better than the Bayesian model [8, 9]. The accuracy curves confirm this finding.

The accuracy of probabilistic prediction can also be compared using various indices [10]. These numbers give a general, overall picture of a model's reliability but fail to provide information concerning the relationship between accuracy and discriminatory power. The ROC curve and accuracy curve together can perform this latter function.

Various methods have been derived for comparing ROC curves [11–13]. Though we have not proposed a method of comparing accuracy curves, the examples from the simulated and clinical data support this approach.

The ROC curve is determined from the "ideal" accuracy expected for a given ranking of probabilities. Comparison of the accuracy curve with the corresponding ROC curve gives a continuous estimate of a model's deviation from the ideal. Different probability distributions may even have the same ROC curves, but their accuracy curves will tell them apart.

Acknowledgement—The editorial assistance of Ms Maggie Meyer is greatly appreciated.

REFERENCES

1. Diamond GA, Forrester JS, Hirsch M, Staniloff HM, Vas R, Berman DS, Swan HJC. Application of clinical probability to the diagnosis of coronary artery disease. **J Clin Invest** 1980; 65: 1210–1221.
2. Moise A, Diement B, Cucimetiere P, Bourassa MG. Comparison of receiver operating curves derived from the same population: a bootstrapping approach. **Comput Biomed Res** 1985; 18: 125–131.
3. Diamond GA. ROC steady: A receiver operating characteristic curve that is invariant relative to selection bias. **Med Decis Making** 1987; 7: 238–243.
4. Patterson RE, Eng C, Horowitz SF. Practical diagnosis of coronary artery disease: A Bayes' theorem nomogram to correlate clinical data with noninvasive excercise tests. **Am J Cardiol** 1984; 53: 252–256.
5. 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.
6. Pryor DB, Harrell FE, Lee KL, Califf RM, Rosati RA. Estimating the likelihood of significant coronary artery disease. **Am J Med** 1983; 75: 771–780.
7. McCarthy DM, Sciacca RR, Blood DK, Cannon PJ. Discriminant function analysis using thallium-201 scintiscans and exercise stress test variables to predict the presence and extent of coronary artery disease. **Am J Cardiol** 1982; 49: 1917–1926.
8. Detrano R, Leatherman J, Salcedo EE, Yiannikas J, Williams G. Bayesian analysis versus discriminant function analysis: their relative utility in the diagnosis of coronary disease. **Circulation** 1986; 73: 970–977.
9. Detrano R, Guppy KH, 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.
10. Hilden J, HJabbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis: III. Methods based on continuous functions of the diagnostic probabilities. **Meth Inform Med** 1978; 17: 238–246.
11. Hanley JA, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. **Radiology** 1983; 148: 839–843.
12. McNeil B, Hanley JA. Statistical approaches to the analysis of receiver operating characeristic (ROC) curves. **Med Decis Making** 1984; 4: 137–150.
13. Metz CE. Basic principles of ROC analysis. **Semin Nucl Med** 1978; 8: 238–298.