The Robustness of the "Binormal" Assumptions Used in Fitting ROC Curves

JAMES A. HANLEY, PhD

The binormal form is the most common model used to formally fit ROC curves to the data from signal detection studies that employ the "rating" method. The author lists a number of justifications that have been offered for this choice, ranging from theoretical considerations of probability laws and signal detection theory, to mathematical tractability and convenience, to empirical results showing that "it fits!" To these justifications is added another, namely that even if an alternative formulation based on another underlying form (e.g., power law) or model (e.g., binomial, Poisson, or gamma type distributions) were in fact correct, the binormal fit differs so little from the true form as to be of no practical consequence. Moreover, the small lack of fit is unlikely to be demonstrated in practice: it is obscured by the much larger variation that can be attributed to sampling of cases. In addition, even if a very large sample of cases could be studied, the small number of rating categories used does not permit seemingly very different models to be distinguished from one another. *Key words:* binormal assumptions; ROC curves; signal detection theory; rating method. **(Med Decis Making 8:197–203, 1988)**

The binormal form[5] is the one most commonly used to formally fit a receiver operating characteristic (ROC) curve to data from a signal detection study employing the "rating" method. It describes a curve with the same functional form as that implied by two underlying Gaussian distributions for the decision variable (hence "binormal"). The curve is described by two adjustable parameters: the distance between the two distributions (standardized by the standard deviation of the "signal + noise" distribution) and the ratio of their standard deviations. These parameters can be estimated from the slope and intercept of a straight-line fit to the observed ROC points plotted on binormal deviate paper.[5] Since 1969, a formal method has been available for estimating the parameters and their uncertainty.[2]

Some new users of ROC methodology refer to the binormal form as "the ROC curve generated by two overlapping normal distributions." However, from the beginning, several authors have continued to emphasize that "while two distributions determine a ROC curve, a ROC curve does not fully determine the underlying distributions," i.e., any monotonic transformation of the decision-variable axis yields generally different underlying distributions but the same ROC curve.[4,8,12,13] In practice, however, since the distributions are "latent" and thus, their form unobservable, the binormal ROC curve is often inaccurately described as "having arisen from two Gaussian distributions" rather than "from two distributions which can, by some monotonic rescaling of the decision axis, be transformed into two Gaussian ones."

Several ROC forms have been catalogued and described by Egan.[4] They are of two types. In one type, which Swets termed "algebraic,"[13] the curve can be described by an equation—without beginning with underlying distributions. The simplest example is the power ROC, in which the true-positive (TP) and false-positive (FP) proportions are linked by the equation $TP = FP^k$, with one parameter k ranging from 0 (perfect discrimination) to 1 (chance performance). The second type of ROC curve is generated by specifying two overlapping (but latent) probability distributions. The binormal one is the commonest example, although many other pairs of distributions are possible.

In spite of the wide choice of distributions, none of these competitors to the binormal form has ever become popular. In the case of algebraic forms, this is not surprising, since it is more difficult to estimate their parameters by conventional statistical methods. Moreover, the estimation process does not yield a measure of the uncertainty of these estimates, and so it is difficult to make statistical inferences (the latter is now less of a problem with the use of "sample reuse" methods such as jackknifing and bootstrapping[3,7]).

The same general estimation method (maximum likelihood) can be used to fit all "distributional" forms. However, the full computational procedure has been laid out for only two of these, the binormal and the "bi-logistic."[11] Like the Dorfman and Alf procedure, fitting the bi-logistic ROC by maximum likelihood involves an iterative procedure, but the computations require fewer specialized statistical functions. However, it has been passed over in favor of the Dorfman and Alf procedure published the next year.

Received March 12, 1987, from the Department of Epidemiology and Biostatistics, McGill University, Montréal, PQ, Canada, H3A 1A2. Revision accepted for publication November 5, 1987. Supported by an operating grant from the Natural Sciences and Engineering Council of Canada.

Table 1 • Justifications of the Binormal Form*

The Gaussian distribution is a natural one
. . . many of the random variables describing natural phenomena may be considered to be the sum of a large, relatively constant number of other independent, random variables; . . . since we often believe that sensory events are composed of a multitude of similar, smaller events, the Central Limit Theorem might be invoked to justify the Gaussian assumption[5] [pp 54–58]
Other distributions can be approximated by Gaussian ones
. . . the binomial, Poisson, hypergeometric, and chi-squared distributions can, under certain conditions, be closely approximated by the normal distribution[5] [p 58]
The decision axis can be transformed to produce Gaussian distributions
. . . any monotonic transformation of the decision-variable axis yields generally different underlying distributions but the same ROC curve[4,12]
Other ROC forms look "almost straight" on binormal paper
. . . the plot of Power-Law ROCs in binormal coordinate shows that they are nearly straight lines[4] [p 100]
Empirical results showing that "the binormal form fits"
. . . illustrative results from one observer are shown. The ROC curve is not very different from the curve that we expect if the underlying distributions were Gaussian[5] [p 185–187]
. . . it is a highly robust, empirical result, which is now substantiated in dozens of diverse applications, that the empirical ROC is very similar in form to a theoretical ROC derived from normal probability distributions. In practice, in other words, the ROC curve is adequately described by a straight line when plotted on a binormal graph[15] [pp 5 and 30]
Mathematical tractability and convenience
. . . it has the convenient property that all possible binormal ROC curves are transformed into straight lines if plotted on "normal deviate" axes [Metz,[10] quoting Green and Swets[5]]
. . . it is relatively easy to fit by eye and is easily fitted by statistical techniques that give estimates of the slope and intercept of the binormal ROC[15] [p 31]

*Quotes have been edited.

A number of justifications had been offered for the binormal form. These are grouped under six main headings in table 1. The claim that "the binormal model fits" has recently been considerably strengthened by Swets' survey of empirical ROC forms in experimental psychology and in several practical fields.[14] In addition, his other recent paper,[13] dealing with the models implied by certain ROCs, rules out some on purely theoretical grounds.

One must add to this list of reasons for the popularity of the binormal form one very practical one, namely the availability of specialized computer programs. Formally fitting any form requires matrix inversion and the use of several statistical subroutines; thus, the computations cannot easily be programmed by naive users. Dorfman made widely available a FORTRAN program to carry out the estimation and curve fitting of the binormal form, and a complete listing of a revised version of this program has been included in the text by Swets and Pickett.[15] Also, a comprehensive computer package (including procedures to compute statistical power and to estimate the parameters of correlated binormal ROC curves[9]) has been made available by Metz.[10]

With this many justifications or reasons for using it, the binormal model is likely to continue to be used. However, many of the justifications for the binormal form are either pragmatic or else too technical to evaluate, and it is difficult to assure the less sophisticated user that the binormal form will serve as well as or better than other models. Such a user could legitimately ask: what if the underlying decision scale *cannot* be transformed to produce Gaussian distributions? What if some *alternative* form or pair of distributions are the (unobserved) truth? What if the good empirical fits are merely a consequence of sample sizes that are too small to distinguish the binormal from other forms?

To answer these questions, we assessed the effect of fitting the binormal form to a ROC curve that we knew had arisen from another form. We wished to determine whether, with samples large enough to separate pure lack of fit from sampling variability, the misspecification of the ROC form had any serious consequences.

Materials and Methods

DATA

As described below, we studied each of the major alternative ROC forms discussed by Egan; by varying the parameters of each form, we examined several ROC curves within each form. Depending on the form, we used four to seven rating categories. In order to minimize sampling variation and allow real lack of fit to show itself, for each ROC curve we used a sample of 10,000 noise-only (n) and 10,000 signal-plus-noise (s + n) observations. The frequencies in the different rating

categories were designed so that the observed points fell exactly on the true ROC curve (see below).

POWER ROCS. We chose six rating categories by having five FP operating points of 0.09, 0.25, 0.49, 0.64, and 0.81. The corresponding TP operating points were calculated using the equation $TP = FP^k$, with the values of the parameter k varying in the six different datasets from 0.02 to 0.75. The frequencies in each rating category for the "n" and "s+n" distributions were calculated by taking differences of successive FPs and TPs.

ROC CURVES BASED ON BINOMIAL DISTRIBUTIONS. The frequencies in seven rating categories were taken proportional to the seven binomial probabilities $B(6,\pi)$, with the parameter for the distribution of signal + noise ratings (π_{s+n}) always larger than that for the noise-only rating (π_n). Some 45 different sets of rating data were generated by varying the parameter pairs π_n and π_{s+n} within the range $0.1 \leq \pi_n < \pi_{s+n} \leq 0.9$ in increments of 0.1.

ROC CURVES BASED ON POISSON DISTRIBUTIONS. The integers from 0 to infinity were grouped into (0,1), (2,3), (4,5), (6,7) and (≥8) to simulate five rating categories. The frequencies in these five rating categories were taken proportional to the Poisson probabilities $P(\mu_{s+n})$ for the distribution of signal + noise ratings and $P(\mu_n)$ for the noise-only ratings. Some 15 different sets of rating data were generated by varying the parameter pairs μ_n and μ_{s+n} within the range $1 \leq \mu_n < \mu_{s+n} \leq 6$ in increments of 1. An additional 36 datasets were generated using $0.1 \leq \mu_n < \mu_{s+n} \leq 0.9$.

ROC CURVES BASED ON CHI-SQUARED DISTRIBUTIONS. The chi-squared range was divided into six "rating categories," with the cutoffs varying depending on the locations of the n and s+n distributions. The frequencies in the six rating categories were taken proportional to chi-squared distributions with ν_{s+n} and ν_n degrees of freedom. Some 23 different datasets were generated in the range $1 \leq \nu_n < \nu_{s+n} \leq 10$.

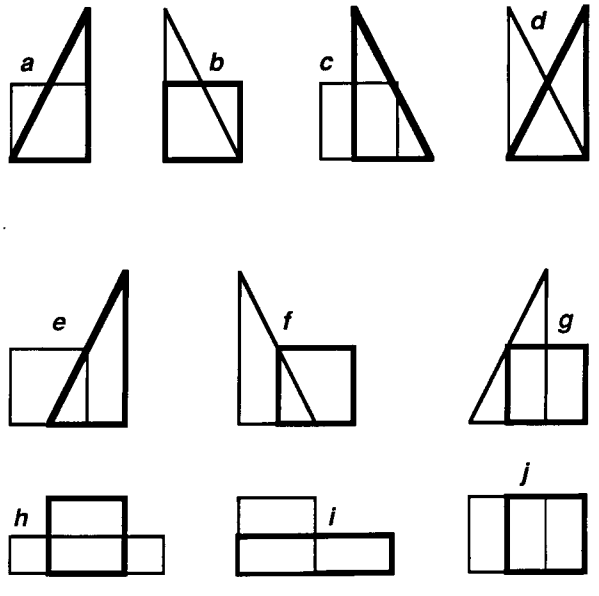

FIGURE 1. The overlapping rectangular and triangular distributions used to produce ten ROC curves. In each pair, the shape of the "signal plus noise" distribution is shown using bolder lines.

ROC CURVES BASED ON GAMMA DISTRIBUTIONS. The scale parameter was set equal to 1. Datasets were created by varying a pair of location parameters over the range 1 to 10. Using the procedure already described for the chi-squared distributions and using the same cutoff points, 26 datasets were created.

ROC CURVES BASED ON RECTANGULAR/TRIANGULAR DISTRIBUTIONS. Ten datasets were created by overlapping rectangular and triangular distributions, as shown in figure 1. Five rating categories were used in each case.

We omitted logistic-based ROC curves because it had been pointed out by Birdsall[1] that they are difficult to distinguish from binormal ROCs.

ANALYSIS

We used the computer program ROCFIT,[10] kindly supplied by Charles Metz, to fit a two-parameter ROC curve to each dataset. To judge the fit, we used the following measures:

1. The maximum discrepancy, in the TP or FP direction, between the fitted ROC points and the true ones. As an example, the discrepancies from fitting a binormal curve to the power curve with $k = 0.1$ are shown in table 2. The largest discrepancy, 0.0034, was found at the second TP (correct TP value 0.8705). [We could also have used the Euclidean distance between the true and fitted (TP,FP) points (which can be as much as $\sqrt{2}$ times larger, but, as will become clear below, it would not have materially changed the conclusions.]

   We compared two series of four or five points rather than two entire curves, for two reasons. First, two of Egan's forms (Poisson, binomial) give rise to ROC curves that can be somewhat arbitrarily drawn between the discrete cutpoints, so that the cutpoints are the only natural points at which to compare these curves with the binormal fits. Second, if two smooth ROC curves are within 0.01 of each other at four or five points in the unit square, then they must also be close at all the other places where we didn't calculate them.

2. The significance level of the chi-square goodness-of-fit statistic (in the example shown, it was $p = 0.03$). This statistic,[2] based on the discrepancies between the observed (or in our case, the true) and the fitted counts in each category, is compared with a χ^2 distribution with three fewer degrees of freedom than there are rating categories. While one could argue that in this study we know what model was used to generate each dataset, what we wished to know was whether the only test-of-fit procedure available in practice can detect that the model that generated the data was not the binormal one.

Table 2 • Fit of Binormal Model to Data from Power-Law Curve $TP = FP^{0.1}$ with 10,000 Noise Only and 10,000 Signal + Noise Observations

	Rating Category					
	1	2	3	4	5	6
Response data						
Noise trials	1,900	1,700	1,500	2,400	1,600	900
Signal + noise trials	209	228	252	606	845	7,860
Observed operating points						
False positive	0.0900	0.2500	0.4900	0.6400	0.8100	1.000
True positive	0.7860	0.8705	0.9311	0.9563	0.9791	1.000
Fitted operating points (binormal model)						
False positive	0.0918	0.2480	0.4878	0.6393	0.8114	1.000
True positive	0.7854	0.8739	0.9332	0.9565	0.9772	1.000
Discrepancies (absolute)						
False positive	0.0018	0.0020	0.0022	0.0007	0.0014	
True positive	0.0006	0.0034	0.0022	0.0002	0.0019	
Maximum discrepancy		0.0034				
Goodness of fit chi-square	9.15					
Degrees of freedom	3					
Significance level	0.03					

3. For the power law curves, where the area under the true ROC curve could be calculated analytically, the difference between this area and the area under the fitted curve.

We summarized the results still further by reporting, for each family of distributions, the worst discrepancy, as calculated in 1, and the number of datasets that showed a statistically significant lack of fit, i.e., $p < 0.05$.

Results

The main results are summarized in table 3. It can be seen that except for the power law and rectangular/triangular forms, the binormal form fitted exceptionally well, with discrepancies occurring only in the third decimal place of the TP and FP values. In the two power law forms where the binormal model showed statistically significant lack of fit, the discrepancies were

Table 3 • Fits of Binormal Model to Various ROC Curves

	Parameters	*n*	Areas	Maximum Error of Fit*	No. with Significant Lack of Fit†
ROC curves					
Power law	k: 0.02, 0.10, 0.20, 0.33, 0.50, 0.75	6	0.57–0.98	0.0035	2
Binomial	Noise: π = 0.1–0.8 Signal + noise: π = 0.2–0.9	45	0.66–0.99	0.0019	0
Poisson	Noise: μ = 1–5 Signal + noise: μ = 2–6	15	0.62–0.98	0.0004	0
	Noise: μ = 0.1–0.8 Signal + noise: μ = 0.2–0.9	36	0.53–0.84	0.0002	0
Chi-squared	Noise: ν = 1–9 Signal + noise: ν = 2–10	23	0.75–0.99	0.0018	0
Gamma	Noise: a = 1–9; b = 1 Signal + noise: a = 2–10; b = 1	26	0.77–0.99	0.0015	0
Rectangular/triangular	See figure 1	10	0.50–0.96	0.0302	7

*The error in each curve was the biggest $|TP_{true} - TP_{fitted}|$ or $|FP_{true} - FP_{fitted}|$ among the operating points.
†The fit is measured by a chi-squared type of statistic comparing actual and fitted frequencies.

entirely negligible, and can be shown only on graphs with very high resolution (see fig. 2). In all six power law curves, the area under the fitted curve was within 0.0006 of the true area.

The largest discrepancies were seen in the ROC curves constructed by overlapping rectangular and triangular distributions. Figure 3 shows the true and fitted curves.

Discussion

Egan has described in considerable detail the ROC curves that arise from different distributions and forms; he assessed how close one of them (the power law ROC) is to the binormal model by plotting it on binormal paper. Swets[13] has recently plotted the logistic, power law, and threshold models on binormal axes, but one can only judge by eye how nonlinear they appear. Our approach was different in that we took the data implied by several models, converted them to the tabular format typical of rating method data, and produced formal binormal fits, along with derived indices. We then compared these binormal-based fits with the original ROC curves. The results reported here indicate that the binormal-based fits are certainly good enough for all practical purposes.

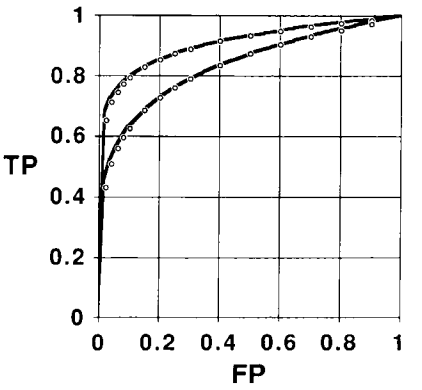

FIGURE 2. Two power-law ROC curves, of the form $TP = FP^k$, in which the binormal model showed a statistically significant lack of fit ($p < 0.05$). The true power-law curves are shown as solid lines (upper curve $k = 0.1$, lower curve $k = 0.2$). The open circles lie on curves fitted by the binormal model. The lack of fit is negligible, and due only to the large sample sizes used.

The only serious exception we were able to document was in four ROC curves generated by rectangular/triangular distributions (see fig. 3). In all four, the true curve approaches FP = 0 abruptly, implying what Swets calls a threshold model; in two of the four (*c* and *j*),

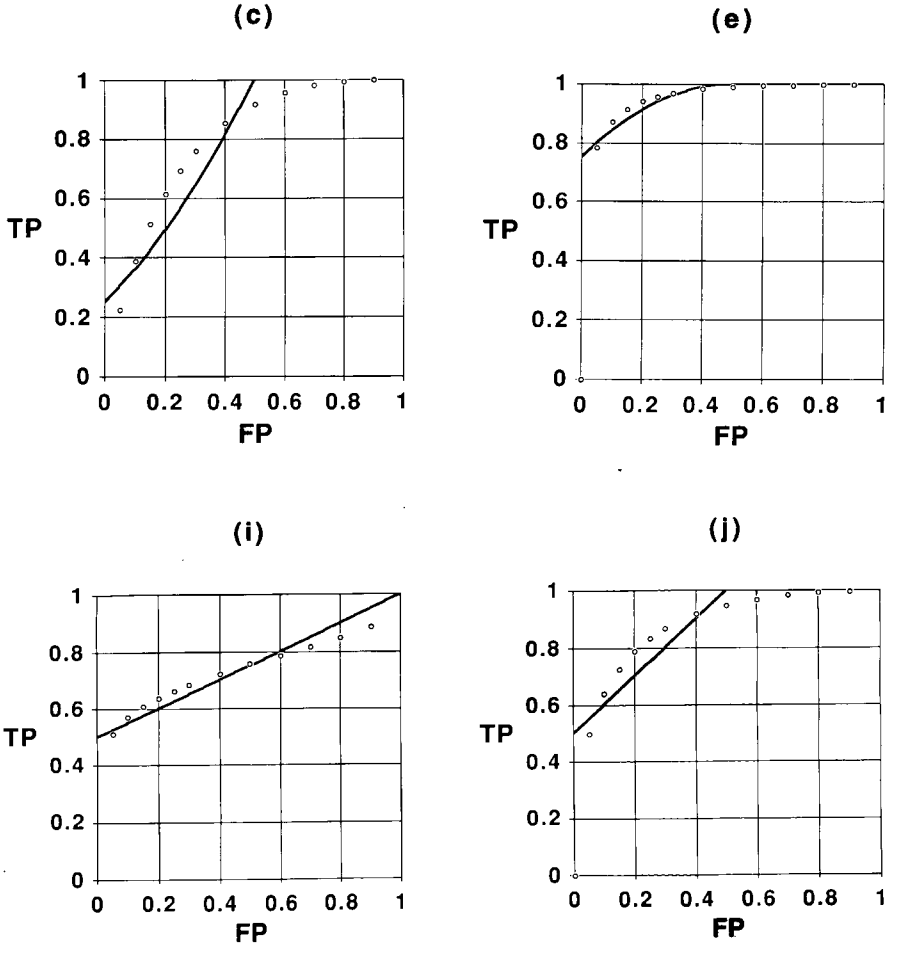

FIGURE 3. Four ROC curves, based on rectangular/triangular distributions, to which the binormal model showed a statistically significant lack of fit ($p < 0.05$). The four curves, correspond to the paired distributions *c*, *e*, *i*, and *j* of figure 1. The open circles represent the curves fitted by the binormal form.

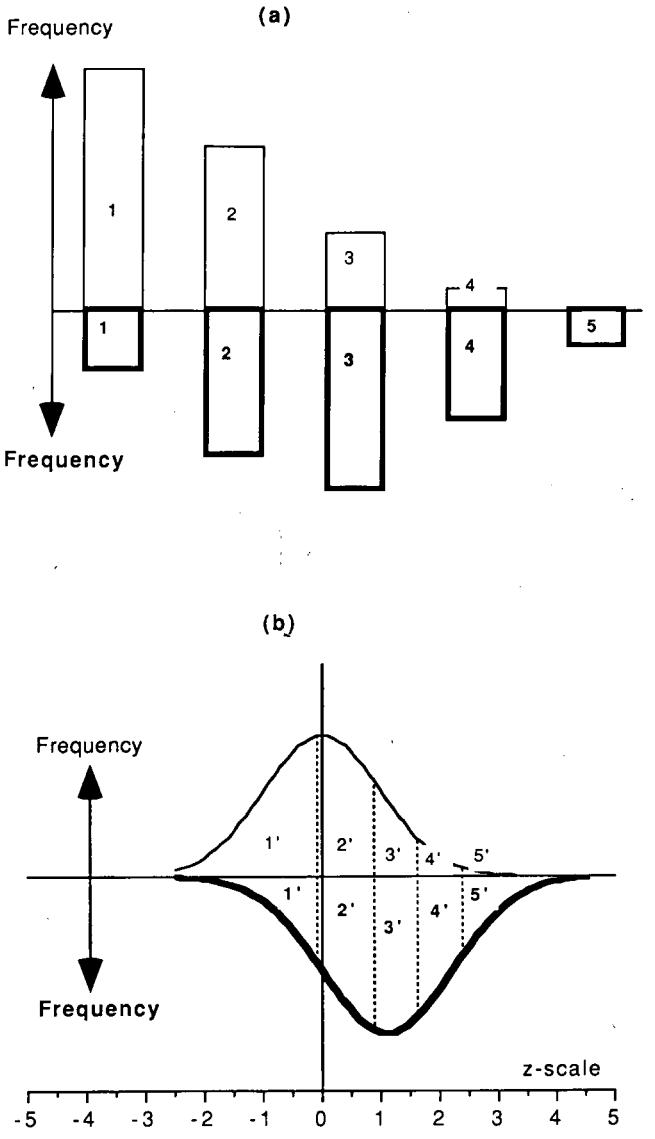

FIGURE 4. One explanation of why the binormal form fits rating data so well. (a) The observed distributions of ≈ 57,000 non-cases (*upper distribution, with lighter lines*) and ≈ 2,500 cases (*lower distribution, with bolder lines*) of postsurgical wound infections across the "simplified risk index" developed by Haley et al.[6] (*b*) The binormal distributions fitted to the data in (*a*); again, the upper distribution (*lighter line*) represents non-cases and the lower one (*heavier line*), cases. The fitting procedure consists of matching as closely as possible the five areas marked 1 to 5 with those marked 1′ to 5′ and those marked **1** to **5** with those marked **1′** to **5′**.

it does likewise at TP = 1. One could argue that this form is not unrealistic, in that there may often be some sample stimuli about which there is no diagnostic doubt. The poor fit to the curve in panel *c* was to be expected, in view of the curve's concave upwards form. The binormal curve fitted to the curve in panel *i* is also notable, since it too is partly concave and crosses the positive diagonal. Although the binormal form is not necessarily always strictly convex upwards, it usually is in practice. Again, the anomaly is in the ROC region of least interest. Moreover, Swets argues that such ROC curves are "irregular" and that they are not observed in practice.

One other index that we might have used to compare the true and fitted curves is the optimum (TP,FP) point, obtained, as described by Swets and Pickett[15] (pp 40–42), through the slope (β) of the ROC curve. If we use this index with the six power law curves (which are the most tractable mathematically), we find that in the β range 0.5 to 2.0 [excluding one point where the FP was 0.5 (50%)], the Euclidean distance between the true and fitted optimum (TP,FP) points is not more than 0.017. Thus, although this index produces a larger discrepancy than the "maximum error" reported in table 3, the differences are still very small.

Although the empirical evidence provided by Swets and by this investigation for the robustness of the binormal form is compelling, one could still suspect that this robustness is not an intrinsic characteristic of ROC curves. Why, then, is it so in all the cases, both empirical and simulated, that have been examined?

The most important reason is the small number of rating categories. A recent study[6] of the accuracy of a prognostic scale to predict cases of postsurgical complications illustrates this. Some 2,500 cases and almost 60,000 non-cases were studied; in spite of the unusually high statistical power to separate sampling fluctuations from true lack of fit, the fit to the binormal model was virtually perfect ($\chi^2 = 2.8$ on 2 df). Figure 4 shows why the binormal form fits these data, or almost any rating data, so well. As emphasized at the outset, the binormal form does not restrict one to distributions that are explicitly Gaussian; a pair of asymmetric distributions can often be made approximately Gaussian by shrinking or expanding the underlying scale. In this example, the cases already had a Gaussian-like distribution on the risk index; more importantly, almost half the non-cases were in the "lowest-risk" category, making the task of fitting a Gaussian curve to the non-cases especially easy. If the distribution of cases had had a long left tail, with a large portion concentrated in the "highest-risk" category at the right, the task would have been even easier, since one could likewise fit a "half-Gaussian" distribution to them and disturb even less the fit to the non-cases. One gets the clear impression that many other distributions would fit the 2×5 data table equally well.

With the limited resolution inherent in rating data, it is not possible accurately to distinguish one distributional model from another. A true test of the binormal model is possible only in signal-detection systems yielding a fine-grain numerical quantity, such as a tissue density or probability calculated from a prediction equation.

Stated in equivalent, but more mathematical, terms, the binormal form is highly parameterized relative to the grain of rating method data. No matter how large

the numbers of cases and non-cases, an experiment that has five rating categories still produces only eight independent pieces of data. The method of estimation requires that six of these be used in the fitting, an exceptionally large ratio of parameters to data points. When so many of the data items are used to fit the model, one should expect a good fit.

The above observations are not meant to disparage the binormal model, but rather to elucidate *why it fits so well*. Some other models, such as a pair of logistic distributions with unequal variances, would probably work just as well, but would require at least as many parameters, and the computational effort required to estimate their parameters would be the same. For all of these reasons, the binormal model can continue to be used.

This work was carried out with the technical assistance of Carl Brewer.

References

1. Birdsall TG: The theory of signal detectability: ROC curves and their character. Dissertation Abstracts International, 28 1B
2. Dorfman DD, Alf E: Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. J Math Psychol 6:487–496, 1969
3. Efron B, Gong G: A leisurely look at the bootstrap, the jackknife and crossvalidation. Am Statistician 37:36–48, 1983
4. Egan JP: Signal Detection theory and ROC Analysis. New York, Academic Press, 1975
5. Green DM, Swets JA: Signal Detection Theory and Psychophysics. New York, Wiley and Sons, 1966
6. Haley RW, Culver DH, Morgan WM, et al: Identifying patients at high risk of surgical wound infection. Am J Epidemiol 121:206–211, 1985
7. McNeil BJ, Hanley JA: Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis Making 4:137–150, 1984
8. Metz CE: ROC methodology in radiological imaging. Invest Radiol 21:720–733, 1986
9. Metz CE, Wang P-L, Kronman HB: A new approach for testing the significance of differences between ROC curves from correlated data. *In:* Deconink F (ed): Information Processing in Medical Imaging. The Hague, Nijhoff, 1984, pp 432–445
10. Metz CE: FORTRAN programs ROCFIT, CORROC and ROCPWR. Available from C Metz, Department of Radiology, University of Chicago, Chicago, IL
11. Ogilvie JC, Creelman CD: Maximum likelihood estimation of ROC curve parameters. J Math Psychol 5:377–391, 1968
12. Swets JA, Tanner WP, Birdsall TG: Decision processes in perception. Psychol Rev 68:301–340, 1961
13. Swets JA: Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 99:100–117, 1986
14. Swets JA: Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 99:181–198, 1986
15. Swets JA, Pickett RM: Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York, Academic Press, 1982