STATISTICS OF MEDICINE, VOL. 15, 1575–1585 (1996)

THE USE OF THE 'BINORMAL' MODEL FOR PARAMETRIC ROC ANALYSIS OF QUANTITATIVE DIAGNOSTIC TESTS

JAMES A. HANLEY

Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montréal, Québec, H3A 1A2 Canada

SUMMARY

The binormal model is widely used for parametric receiver operating characteristic (ROC) analyses of data concerning the accuracy of medical diagnostic tests. Empirical evaluation of the performance of this model in the face of departures from binormality has been limited to interpretations of radiology-type examinations recorded on a rating scale. This paper extends the investigation to the performance of the model with biochemical and other tests recorded on an interval scale. In order to describe non-binormal pairs of distributions, a useful standardized graphical display is developed; this display also illustrates several features of ROC curves. We consider non-binormal pairs of distributions with or without a monotone likelihood ratio and show that by transformation of the underlying scale, one can make many such pairs resemble closely the binormal model. These findings justify Metz's use of the binormal model in the 'LABROC' software for ROC analyses of laboratory type data even when the raw data may 'look' decidedly non-Gaussian.

INTRODUCTION

The use of receiver operating characteristic (ROC) analyses[1] in biomedical applications has increased considerably over the past fifteen years.[2–4] The textbook by Swets and Pickett[2] marked the beginning of the widespread use of this technique outside of psychophysics; the methodology now has its own keyword classification in Index Medicus. Whereas most of the early medical applications were in radiology,[5,6] where test results are subjective and recorded on a rating scale, the methodology has seen increasing use in the evaluation of the accuracy of medical diagnostic and prognostic tests that yield numerical test results.[7–10]

Until the early 1980s, the parametric methods used to fit ROC curves and to derive summary indices of accuracy were based entirely on the binormal model.[11] This model postulates a pair of overlapping Gaussian distributions to represent the distribution of the discriminating variable (or a monotonic transformation of it) in the two populations or states to be distinguished.

Several recent developments, both in software and in methodology, have increased the number of parametric models that can be used to model ROC data. Tosteson and Begg[12] showed how ROC analyses of rating data can be performed using the ordinal regression models of McCullagh.[13] Using the concept of different 'links' (familiar to GLIM users), different underlying latent distributions (for example, probit, logit, and log links implying pairs of Gaussian, logistic and negative exponential distributions) can be used to fit ROC curves using the PLUM software package.[12] Similar model choices are available in the SIGNAL software[14] and the user's manual

CCC 0277–6715/96/141575–11

Received July 1994
Revised September 1995

devotes considerable space to the various possible distributions and their ROC curves. Egan's text[15] describes fully many of these. In a search for 'parametric ROCs on a spreadsheet', Diamond[16] proposed the use of equal-variance logistic distributions (yielding a one-parameter ROC curve), resurrecting a distribution that had been neglected since 1968.[17]

Two investigators have examined the basis for and the performance of the binormal model[18,19] to analyse *rating* data. Neither could find cases where the binormal model would seriously mislead.

However, should one use the binormal model if the diagnostic test results are recorded on a *continuous* scale? If so, how? Goddard[8] warns that if the distribution of the raw data is far from Gaussian, standard errors of accuracy indices based on *directly* fitted normal distributions can be seriously distorted. If the raw data appear to be non-Gaussian, must one then search among the various other bi-distributional forms? Or should one try to transform the raw data to bring them closer to binormality? What about the approach of Metz,[10] implemented in his LABROC program?[20] He fits ROC curves to continuous ('lab') type data by first ranking and then discretizing the data into as many categories as are possible, then fits the binormal model to these categorized data as if they were 'rating' data. Metz argues that this partially parametric approach preserves the rankings of the original data and thus minimizes the loss of information, since one is effectively replacing the raw data by Gaussian order statistics.

Alternatively, should one simply resort to non-parametric indices, such as the area under all[7,21,22] or part[9] of the ROC curve, without actually fitting a smooth ROC curve?

With so many choices, what is a user to do? Is one model 'better' than another? Does the binormal model really make strong assumptions? We undertook our investigation because of the growing number of available methods, and users' need for guidance about the choice of distribution and the impact that it will have on their results and conclusions. Our *main* question was: how flexible is the binormal model? how 'close' to binormal do other legitimate pairs of distributions look or how close can they be made to look? What shapes do various transformations of the original scale produce? A *secondary* question was whether there is one particular scale that is more 'natural' than the others for visually displaying the degree of separation achieved by a test, and whether some of the usual ROC indices of accuracy could be visually estimated from the pair of distributions displayed on this scale.

METHODS

Notation

In signal detection theory the populations or states to be distinguished are referred to as 'noise-only' and 'signal + noise'; in medical testing, they are typically referred to as non-diseased (D−) and diseased (D+). Denote the probability density functions (PDFs) of the values produced by the diagnostic test in these two populations as f and g, respectively, and the corresponding cumulative distribution functions (CDFs) as F and G.

Distributions

We studied two types of pairs of distributions, some designed to yield proper ROC curves, and some that do not necessarily do so. In order for the comparisons of these $\{f, g\}$ pairs in other scales to be meaningful, we compared only those pairs which yielded the same area under the ROC curve.

For the first type of pair, we took the scale to be (0, 1), the probability distribution f in the 'noise-only' population to be uniform, and g to be monotone increasing. Because such $\{f, g\}$ pairs

have a monotone increasing likelihood ratio g/f, they produce a proper ROC curve, that is, one which is concave downwards. We limited our investigation to g's which could be described by polynomials $g(x)$ on (0, 1) of order $k \leqslant 5$. By varying the coefficients, we sought out the $\{f, g\}$ pairs which gave the widest variation in ROC shapes, but yielded the same area under the ROC curve.

For the second type we chose $\{f, g\}$ pairs from familiar families of distributions: Gaussian, negative exponential and beta, on the $(-\infty, \infty)$, $(0, \infty)$ and (0, 1) scales, respectively.

The widely used binormal model, based on a pair of Gaussian distributions, is characterized by two parameters a and b; without loss of generality, f can be taken to be N(0, 1), so that g is N(a/b, $1/b$). The resulting ROC curve is 'proper' only if the standard deviations in the two distributions are equal, that is, if $b = 1$. Although empirical studies[20] have demonstrated that the variation in the 'signal + noise' distribution tends to be larger, that is, $b < 1$, we studied three cases: $b = 2/3$; 1; and 3/2. The area under the corresponding curve is $A = \Phi[a/\sqrt{(1 + b^2)}]$.

The pairs of negative exponential distributions are characterized by a single parameter, that is, f and g were taken to be $\exp(-x)$ and $\lambda\exp(-\lambda x)$, respectively, where $\lambda < 1$. The ROC curve has the closed 'power law' form $TP = FP^{\lambda}$ and the area under the curve is $A = 1/(1 + \lambda)$.

Beta-based ROC curves were generated with $f \sim$ beta(1, 2) and $g \sim$ beta(3, 2); distributions with any greater skewness generate higher areas close to 1·0. Although in this case the ROC area is 0·8, we were unable to find a general closed form for the equation of the ROC curve or the area under it.

Transformations

Our aim was to show how each $\{f, g\}$ pair would look in each of the other two scales. As is explained in the legends to Figures 1 and 2, we mapped those pairs where f was already U(0, 1) into the other scales using the inverse CDF transforms $s' = \Phi^{-1}(s)$ and $s' = -\ln(s)$. We mapped Gaussian, negative exponential and beta pairs (Figures 3 to 5) into the other two scales by first mapping them so that f' was U(0, 1). Φ^{-1} was deliberately chosen to force an f or f' which was uniform(0, 1) to become Gaussian, and to offer comparison with a familiar shape. The densities on the transformed scales were computed analytically when possible and using numerical methods otherwise.

These 'aliases' are only a small number of the re-expressions of each original $\{f, g\}$ pair. Several others were examined, but for space reasons they cannot be reported here. They include pairs re-mapped into (0, 1) by the transform $s' = s^{\theta}$, and from (0, 1) to $(0, \infty)$ using $s' = s/(1 - s)$.

Closeness to binormality after transformation

One can judge how close each $\{f, g\}$ pair is to binormality by using the scale on which f' was Gaussian, and judging how close the g' distribution also is to Gaussian. We restricted ourselves to a visual assessment for a number of reasons. First, in practice, sample sizes seldom exceed a few hundred and are often considerably smaller. With small n's, one would not be able to determine from one's data exactly which scale transformation produces a Gaussian distribution for even one member of the pair. Second, while one could formally test if data are within sampling variability of the binormal distribution, there are no direct methods for finding the change of scale, out of the large number of scale transformations available, which produces the closest binormal fit for any one data set. Even if there were, it is not clear that the exact same re-scaling would apply to all future data sets. Also, binormality cannot be judged using the significance level from a formal goodness-of-fit test; the assessment is necessarily subjective, based on the magnitude of the departures from binormality.

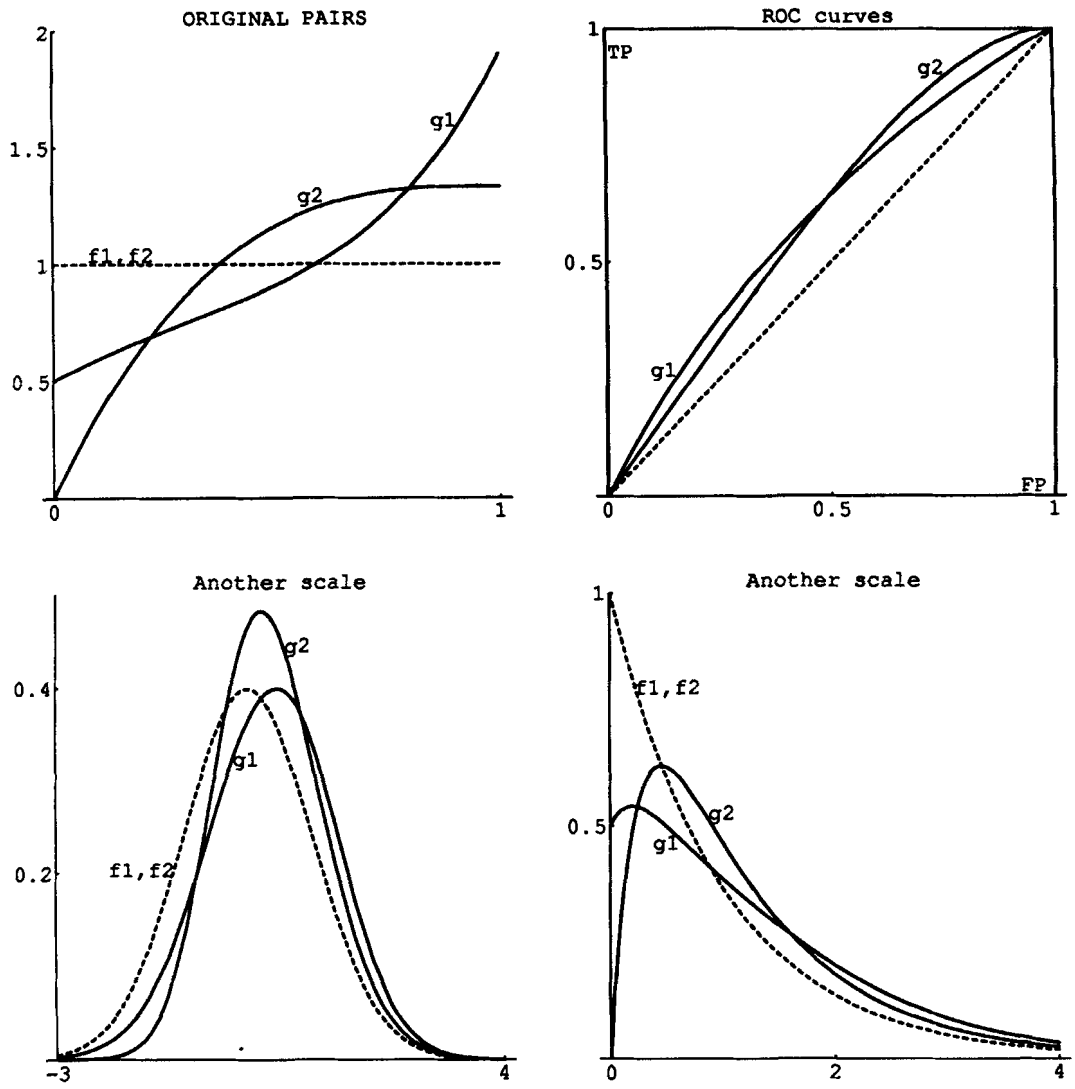

Figure 1. Two pairs of distributions, which yield the same ROC area of 0·6, displayed on different scales. Top left: *original* pairs on (0, 1) scale: the pair $\{f_1 \equiv 1$ [dashed line]; $g_1 = 4x - 4x^2 + 4x^3/3$ [solid line]}. and the pair $\{f_2 \equiv 1$ [dashed line]; $g_2 = 0{\cdot}5 + x - 3x^2/5 + x^4$ [other solid line]}. g_1 and g_2 both yield monotonic likelihood ratio g/f and the same ROC area but are as dissimilar as possible otherwise. Bottom left: the same pairs $\{f_1; g_1\}$ and $\{f_2; g_2\}$ on the $(-\infty, \infty)$ scale, where $f_1[\equiv f_2]$ is Gaussian. Bottom right: $\{f_1; g_1\}$ and $\{f_2; g_2\}$ on the $(0, \infty)$ scale, where $f_1[\equiv f_2]$ has negative exponential distribution. Top right: the ROC curves [solid lines] corresponding to the pairs $\{f_1; g_1\}$ and $\{f_2; g_2\}$

RESULTS

We first list the features of the ROC curve that can be depicted by displaying the overlapping $\{f, g\}$ distributions on a scale on which f is uniform. We then report the answers to the main question posed, namely the shapes of specific pairs of distributions across all three scales.

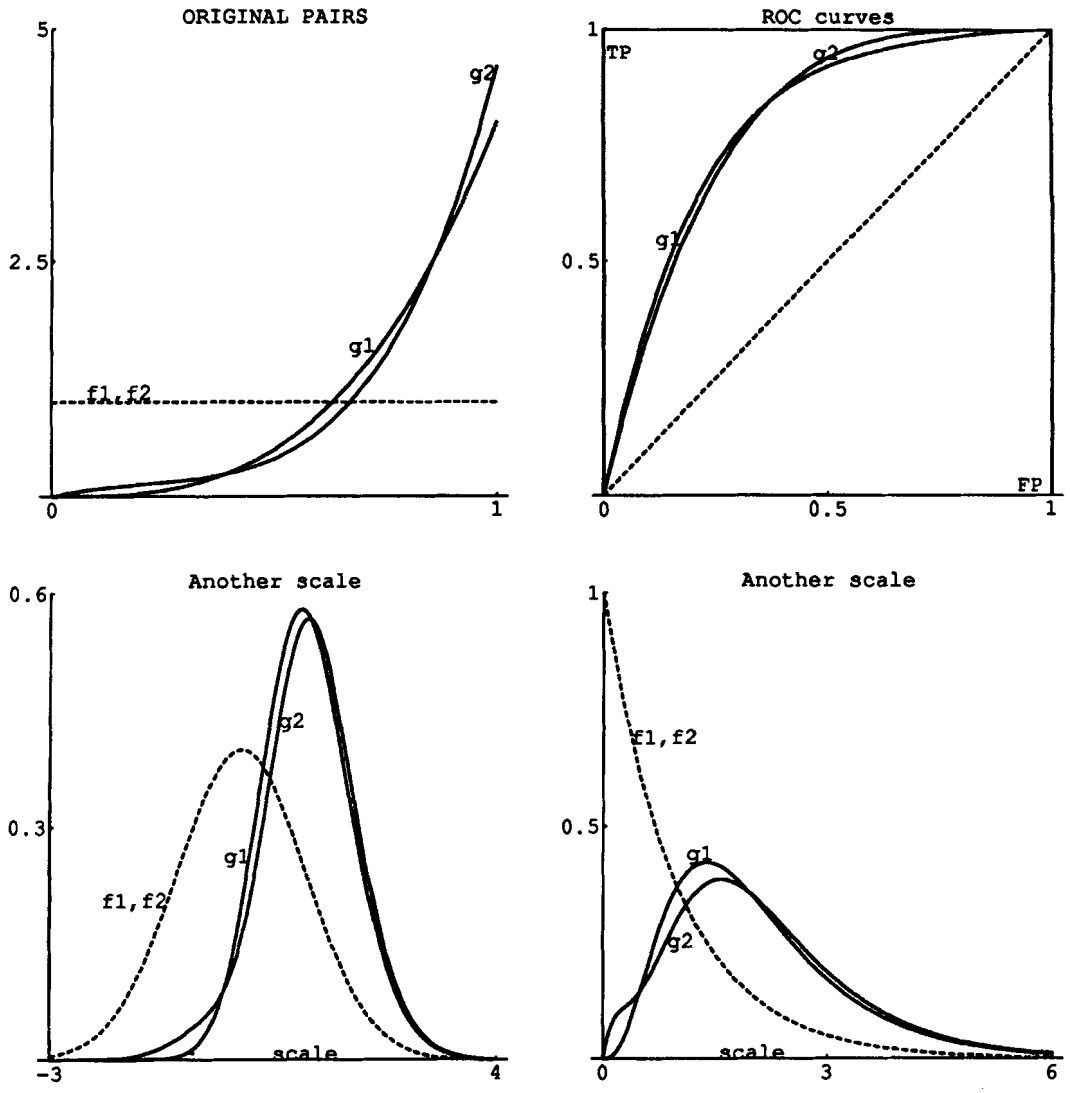

Figure 2. Two pairs of distributions which yield an ROC area of 0·8. Top left: *original* pairs of distributions, again as in Figure 1, with polynomial-based monotonic densities g_1 and g_2, on (0, 1) scale. Bottom left: The pairs $\{f_1; g_1\}$ and $\{f_2; g_2\}$ on the $(-\infty, \infty)$ scale. Bottom right: $\{f_1; g_1\}$ and $\{f_2; g_2\}$ on the $(0, \infty)$ scale. Top right: the ROC curves [solid lines] corresponding to the pairs $\{f_1; g_1\}$ and $\{f_2; g_2\}$

ROC features seen by displaying $\{f, g\}$ on (0, 1) scale on which f is uniform

Mapping the original $\{f, g\}$ pair from their original scale s, into $\{f', g'\}$ on the (0, 1) scale using the transformation $s' = F(s)$ shows several features of the associated ROC curve. To help appreciate these, one can use one of the pairs of distributions in the top left panel of any of Figures 1–5.

First, the specificity (*sp*) of the test for any cutoff s' on the (0, 1) scale is the portion of the f' distribution to the left of s'; the (0, 1) scale directly marks the percentiles of the 'non-diseased' population. The sensitivity (*se*) of the test for this given level (*sp*) of specificity is the portion of the

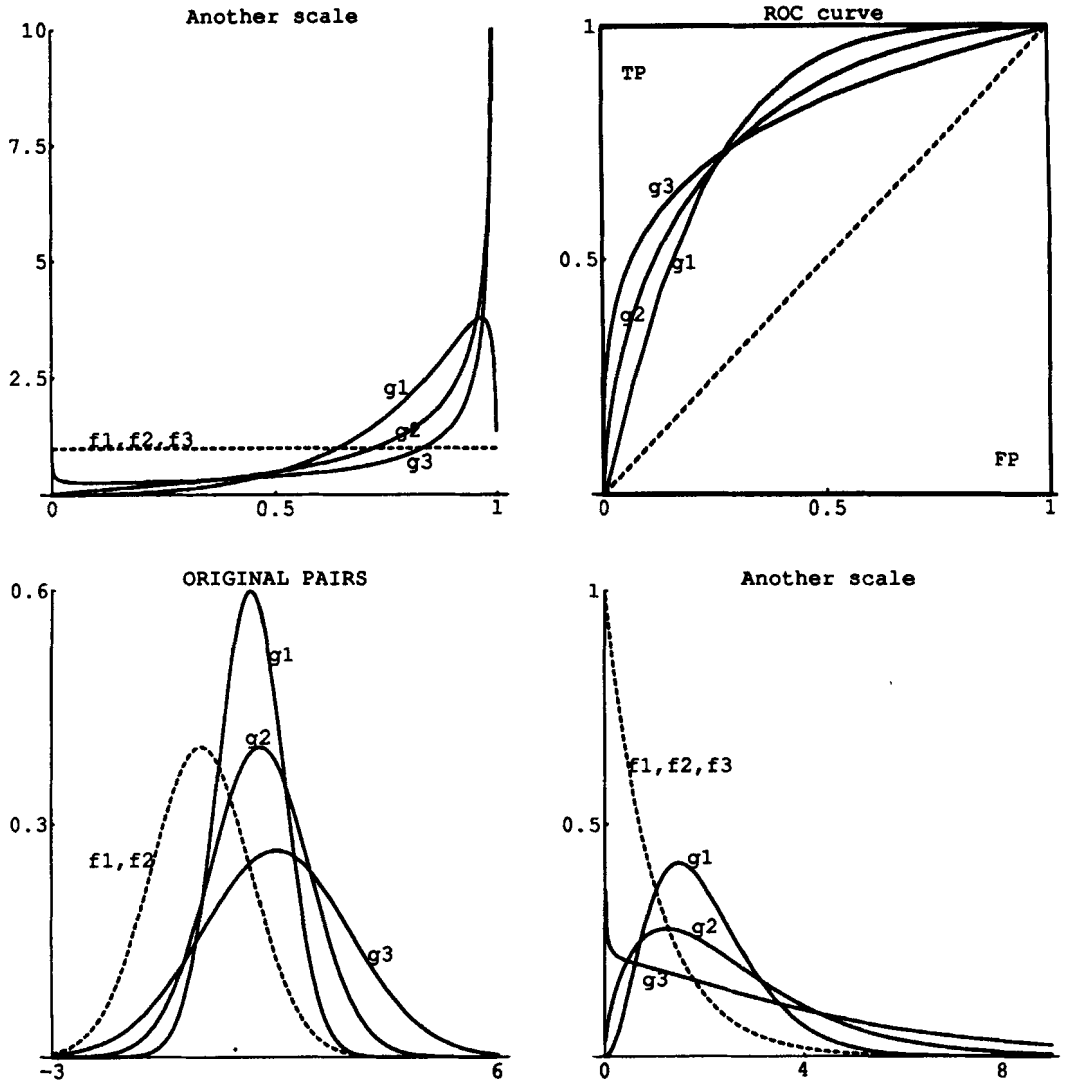

Figure 3. Three pairs of distributions which yield an ROC area of 0·8. Bottom left: *original* pairs $\{f_i; g_i\}$ of *Gaussian* densities on $(-\infty, \infty)$ scale: $f_1 = f_2 = f_3$ [dashed line]; $\sigma[g_i]/\sigma[f_i] = 2/3$, 1 and 3/2, respectively; g_i's represented by solid lines. Top left: the same pairs $\{f_i; g_i\}$ on the (0, 1) scale, with $f_1 = f_2 = f_3$ [dashed line, uniform density] and the $\{g_i\}$ shown as solid lines. Bottom right: the $\{f_i; g_i\}$ on the $(0, \infty)$ scale, where the $\{f_i\}$ have a negative exponential distribution. Top right: the ROC curves [solid lines] corresponding to the three pairs $\{f_i, g_i\}$

g' distribution that lies to the right of the point $s' = sp$ on the (0, 1) scale, that is, $se = 1 - G'(sp)$. *Second*, if f' is uniform, the area under the ROC curve is simply the mean of the g' distribution. The basis for this relationship as follows; if $X \sim f'$ and $Y \sim g'$, then the area under the ROC curve equals Prob$(Y > X)$ and so one can write it as $\int_{x=0}^{x=1} \int_{y=x}^{y=1} f'(x)g'(y)\mathrm{d}y\mathrm{d}x$. The inner integral is simply $1 - G'(x)$. If $f' = 1$, the Prob$(Y > X)$ reduces to $\int_{x=0}^{x=1} [1 - G'(x)]\mathrm{d}x$. As is well known in survival analysis,[23] this integral is the mean of a random variable with PDF g' on (0, 1).

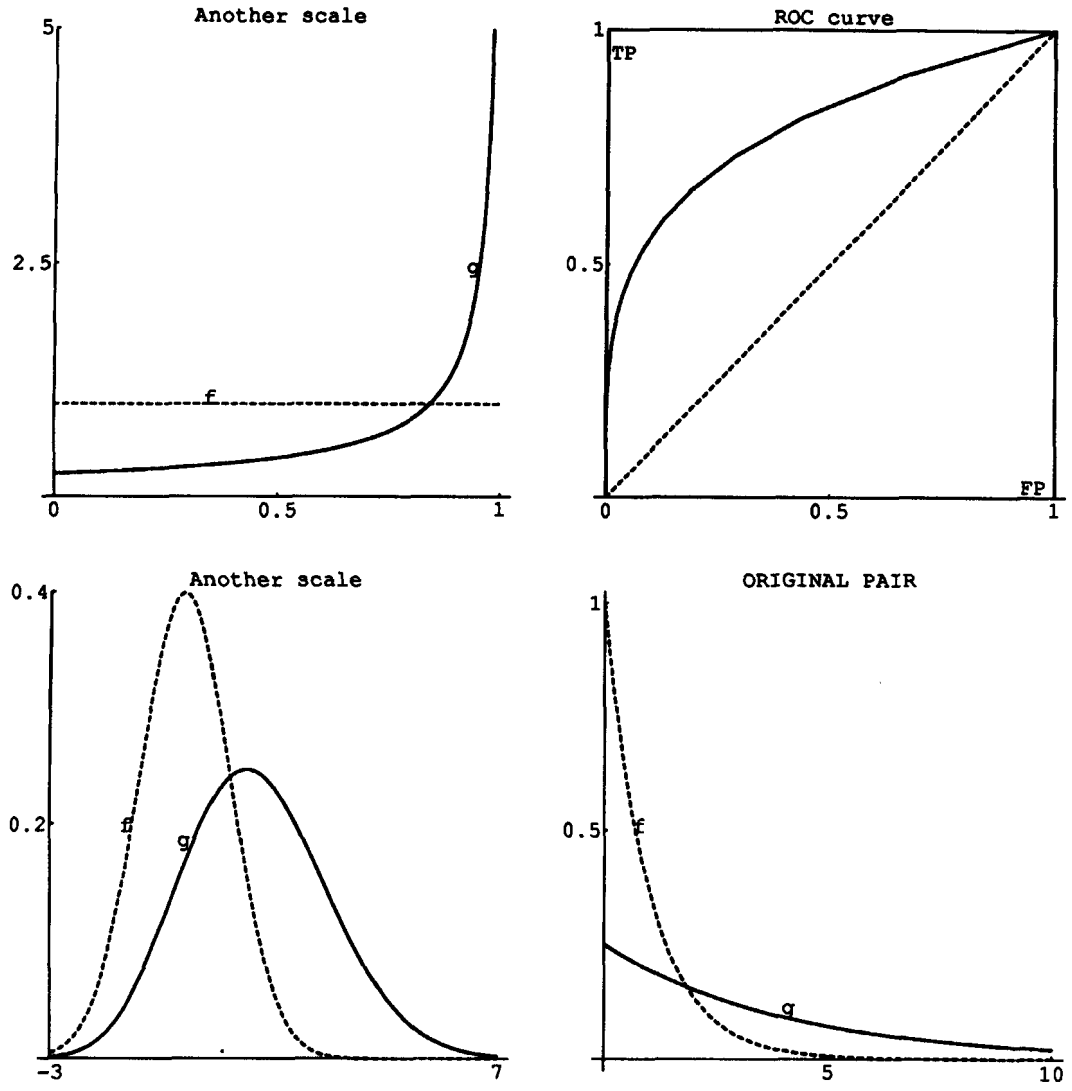

Figure 4. A pair of distributions which yields an ROC area of 0·8. Bottom right: *original* pair of *negative exponential* densities $\{f, g\}$ on $(0, \infty)$ on $(0, \infty)$, with $g[x]/f[x] = 0{\cdot}25\exp[0{\cdot}75x]$. Bottom left: the same $\{f, g\}$ on the $(-\infty, \infty)$ scale, where f is Gaussian. Top left: the pair on the (0, 1) scale, where f has uniform density. Top right: the ROC curve corresponding to $\{f, g\}$

Third, the fact that the area under the ROC curve is a mean of U-statistics distributed according to a PDF g has been used in another context by DeLong *et al.*[7] However, if the diagnostic test is informative, so that g is skewed, one may prefer a more resistant measure of central tendency. Thus, one might calculate from a data set that the *median* of g' was 0·8, say, and report this by saying that 50 per cent of the test values from the diseased population were above the 80th percentile of values in the healthy population. In addition, such a summary should be more attractive to users who object to an area index that averages all sensitivities, including those corresponding to clinically irrelevant specificities. In practice, without explicitly converting the

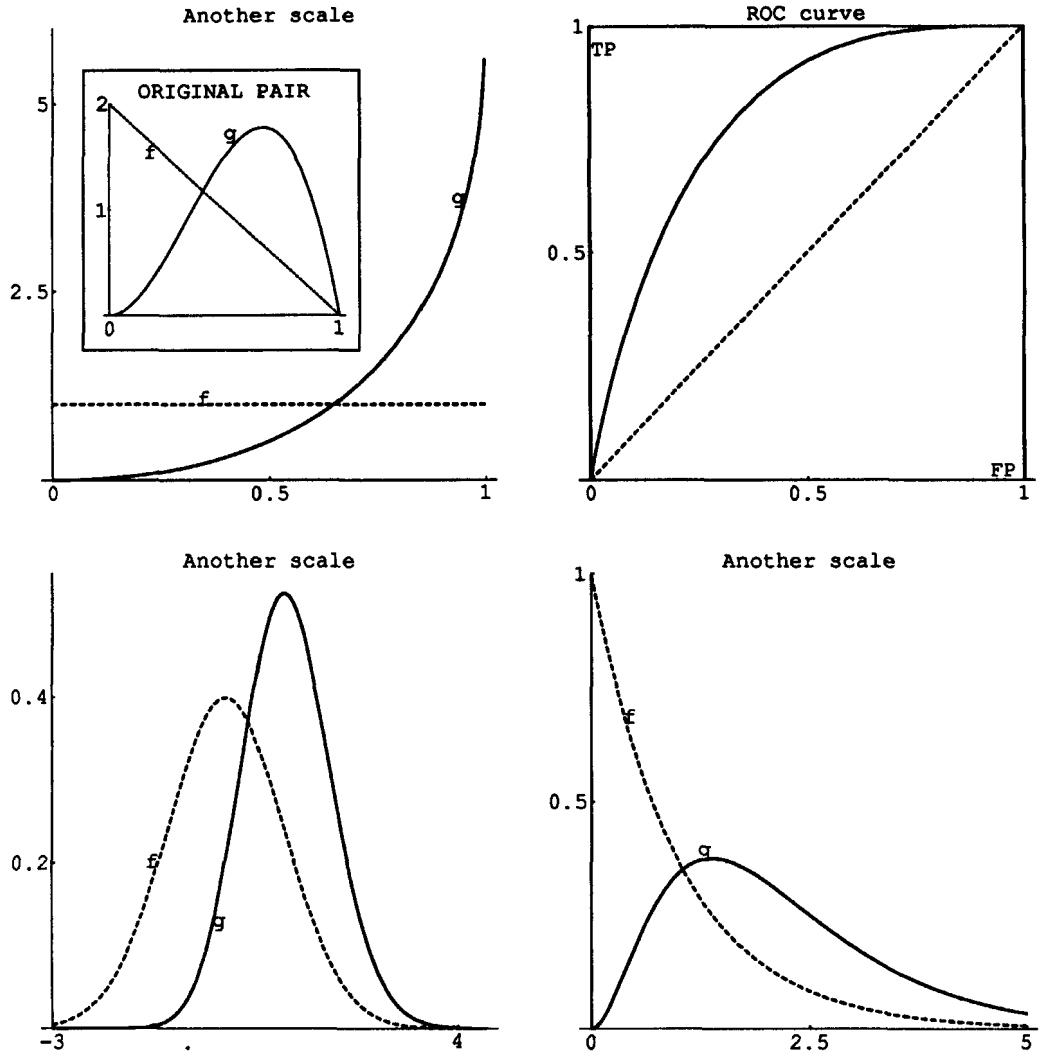

Figure 5. A pair of distributions which yield an ROC area of 0·8. Top left (inset): *original* pair of *beta* densities on (0, 1), with $f \sim$ beta [1, 2] and $g \sim$ beta [3, 2]. Top left: the same $\{f, g\}$ pair on the (0, 1) scale, but where f has uniform density. Bottom left: the same $\{f, g\}$ pair on the $(-\infty, \infty)$ scale, where f is Gaussian. Bottom right: $\{f, g\}$ on the $(0, \infty)$ scale, where f has a negative exponential distribution. Top right: the ROC curve corresponding to $\{f, g\}$

scale so that $f' = 1$, one could also measure this 'median' index of accuracy directly from the ROC curve as the specificity at a sensitivity of 0·5. The representation of ROC data using g allows the statistical variability of such a summary to be assessed.

Finally, the complement, $1 - G'$, of the cumulative distribution function associated with g' becomes the ROC curve one would obtain by plotting specificity on the horizontal axis, as do Wieand *et al.*[9] This last point emphasizes that the g' which must be paired with an f' which is U(0, 1) can be obtained directly from the ROC curve by differentiation, and that the ROC curve is only unique up to the specification of one of the two members f or g.

Pairs of distributions displayed on other scales

We studied $\{f, g\}$ pairs yielding various areas under the ROC curve, but because of lack of space, we show here only those yielding areas of 0·6 and 0·8.

Using the scale (0, 1) with $f \equiv 1$ and polynomials of the form $g = (k + 1)x^k$ to generate proper ROC curves, it was only possible to produce ROC curves with areas up to $A = 0{\cdot}83$ ($g = 5x^4$). Moreover, the higher the ROC area investigated, the less room there is for choices of g. Figure 1 shows two g's that yield the most disparate ROC curves with areas of 0·6, while Figure 2 shows two yielding an area of 0·8. Each figure also shows the representations of the $\{f, g\}$ pairs on the other scales. From these, it is evident that *even though we chose the g's without any regard for 'Gaussian-ness'*, nevertheless, when we transformed to the $(-\infty, \infty)$ scale using the inverse of the Gaussian CDF, the g member of the pair also looks remarkably Gaussian.

We also deliberately mapped f from (0, 1) into $(0, \infty)$ in such a way that it would have a negative exponential distribution (bottom right panel of Figures 1 and 2). The corresponding g's were gamma-like. There is insufficient space to show some of the many other aliases of each $\{f, g\}$ pair in the three different scales, except to say that on the non-Gaussian scales, some of the shapes are quite 'perverse'.

Figures 3–5 show ROC curves based on pairs of Gaussian, negative exponential and beta distributions and illustrate how the distributions appear when they are transformed into other scales. Again, we have deliberately mapped each one into the (0, 1) scale and from there via $y = F^{-1}(x)$ into $(-\infty, \infty)$, to show the reference distribution f in two familiar shapes. Again, it is clear that binormal distributions can have many aliases, and conversely many non-binormal pairs can be made close to binormal by a suitable change of scale.

DISCUSSION

Investigators and end-users of data are often uncomfortable with suggestions by statisticians to use data transformations, and need considerable urging before they will consider another scale. Statisticians fail to emphasize that one is only changing the *scale* and not inherently changing the observations themselves, or their relative ordering; unfortunately, even statistics books typically speak about changing *variables* rather than changing *scales*.

Our figures depict what pairs of known distributions – Gaussian, negative exponential, beta and even 'no-name' – look like on other scales. If the new scale is chosen to induce a Gaussian distribution for the 'noise-only' condition, the distribution for the 'signal + noise' condition is also close to Gaussian.

In practice, with the typical sample sizes involved, one cannot reliably estimate from one's data the scale s' on which f' and g' would be close to Gaussian. The 'rank-preserving but necessarily scale-preserving' approach used by Metz[10] in his LABROC software avoids having to explicitly perform a search for the most appropriate scale. In using the binormal model on 'rating' categories formed by ranks, it is essentially semi-parametric; by not making a commitment to any objective numerical scale, it avoids the risk of biased, but possibly more precise, estimates of diagnostic accuracy estimates.

Several recent papers dealing with quantitative test results have made statements to the effect that 'because our data were not Gaussian, we could not use the binormal model'. It is hoped that this paper dispels this misunderstanding about the use of the binormal model. It emphasizes that Metz's use of the model does not use a transformation of the actual data scale; rather it fits a latent model to categorized data formed from ranks.

There is no obvious statistical advantage to using a binormal over its close cousin the bilogistic model with these categorical 'rating' data formed by the data-dependent discretization of the continuous scale. Both models will be equally robust, provided they allow for asymmetric ROC curves. Our many re-expressions of the same pair of distributions in Figures 1–5 emphasize again that the ROC curve, and all of the indices derived from it, does not arise from one 'true' pair of distributions on a particular scale. Thus, users should be guided by practicality and availability of software and less by the choice of specific distributional forms.

Perhaps surprisingly, if one does not count the variations induced by scale changes, the choice of pairs of distributions that generate proper ROC curves is limited. Empirically, there is considerable evidence that binormal models need to take account of the usually greater variation seen in the 'signal + noise' distribution; accommodating this by the parameter b in the binormal model means that the resulting likelihood is no longer monotonic over the entire scale. It would be interesting to investigate if one could accommodate this additional variance, and at the same time preserve the monotonicity, by choosing a model other than the normal. Alternatively, since part of the increased variance in the 'signal + noise' distribution probably results from a mixture of several disease subtypes, in each of which g/f was monotone, one might model it accordingly.

ACKNOWLEDGEMENTS

This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada and Fonds de la recherche en santé du Québec. Josée Dupuis provided research assistance, and Karim Hajian-Tilaki made useful comments on the manuscript.

REFERENCES

1. Green, D. M. and Swets, J. A. *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966. Reprinted by Peninsula Publishing, Los Altos Hills, CA, 1988.
2. Swets, J. A. and Pickett, R. M. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, New York, 1982.
3. Hanley, J. A. 'Receiver operating characteristic (ROC) methodology', *Critical Reviews in Diagnostic Radiology*, **29**, 307–335 (1989).
4. Begg, C. B. 'Advances in statistical methodology for diagnostic medicine in the 1980's', *Statistics in Medicine*, **10**, 1887–1895 (1991).
5. Metz, C. E. 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, **8**, 283–298 (1978).
6. Metz, C. E. 'ROC methodology in radiological imaging', *Investigative Radiology*, **21**, 720–733 (1986).
7. DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. 'Comparing the areas under two or more correlated receiver operating characteristic curves', *Biometrics*, **44**, 837–845 (1988).
8. Goddard, M. J. and Hinberg, I. 'Receiver operator characteristic (ROC) curves and non-normal data: an empirical study', *Statistics in Medicine*, **9**, 325–337 (1990).
9. Wieand, S., Gail, M. H., James, K. L. and James, B. R. 'A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data', *Biometrika*, **76**, 585–592 (1989).
10. Metz, C. E., Shen, J-H. and Herman, B. A. 'New methods for estimating a binormal ROC curve from continuously-distributed test results', presented at 1990 annual meeting of the American Statistical Association Meeting, Anaheim, CA, 7 August 1990.
11. Dorfman, D. D. and Alf, E. 'Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data', *Journal of Mathematical Psychology*, **6**, 487–496 (1969).
12. Tosteson, A. N. A. and Begg, C. B. 'A general regression methodology for ROC curve estimation', *Medical Decision Making*, **8**, 207–215 (1988).
13. McCullagh, P, 'Regression models for ordinal data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
14. Systat Corporation. *SIGNAL software*, 1988.

15. Egan, J. P. *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
16. Diamond, G. A. 'ROC STEADY. A receiver operating characteristic curve that is invariant relative to selection bias', *Medical Decision Making*, **7**, 238–243 (1987).
17. Ogilvie, J. C. and Creelman, C. D. 'Maximum likelihood estimation on ROC curve parameters', *Journal of Mathematical Psychology*, **5**, 377–391 (1968).
18. Hanley, J. A. 'The robustness of the binormal model used to fit ROC curves', *Medical Decision Making*, **8**, 197–203 (1988).
19. Swets, J. A. 'Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance', *Psychological Bulletin*, **99**, 181–198 (1986).
20. Metz, C. E. *LABROC software*, available from Department of Radiology, University of Chicago.
21. Bamber, D. 'The area above the ordinal dominance graph and the area below the receiver operation characteristic graph', *Journal of Mathematical Psychology*, **12**, 387–415 (1975).
22. Hanley, J. A. and McNeil, B. J. 'The meaning and use of the area under an ROC curve', *Radiology*, **143**, 29–36 (1982).
23. Miller, R. G. *Survival Analysis*, Wiley, New York, 1981, p. 70.