Response to ROC Steady

JAMES A. HANLEY, PhD, COLIN B. BEGG, PhD

In this issue of the journal, Diamond proposes a new method for constructing ROC curves.[2] He believes that the method has three principal advantages over traditional ROC techniques, namely "1) the area under the curve is defined by a single parameter" (resulting in considerable simplicity); "2) the associated standard deviation is analytically defined and readily computed"; and "3) the curve is completely invariant with respect to selection bias." To achieve these goals he has employed a model based on a log odds (logistic) relationship between the test result and true disease status. In this model the single parameter, m, is the common odds ratio between test result and true disease status, common in the sense that it is assumed to be constant regardless of which "cut-off" value is selected for the test result. He proposes estimating m using a weighted average of the odds ratio estimates at each of the cut-off values for which data are available. The smooth ROC curve can then be plotted without recourse to the much more complex iterative calculations that are usually employed for the conventional Dorfman–Alf approach.[3]

Diamond has shed light on some interesting aspects of ROC curves, and we are pleased to have this opportunity to further the discussion, with a view to continuing the development of ROC methodology, an important tool for assessing the value of diagnostic tests. However, we do have some disagreements with Diamond with regard to the advantages cited above, and we deal with them in turn:

1. Simplicity "1) the area under the curve is defined by a single parameter." It would certainly be desirable if the essential features of ROC curves could typically be summarized by a single parameter, and indeed in the examples used by Diamond the fit does seem to be reasonably good. However, this issue has been discussed extensively in the literature,[8] and the consensus seems to be that "scale" terms are usually necessary in addition to "location" terms such as m. Scale terms accommodate differences in "spread" of the test responses, and the empirical evidence seems to suggest

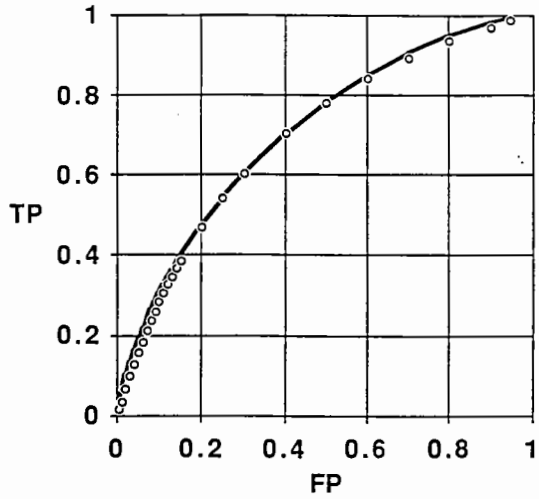

FIGURE 1. One-parameter logistic (*solid line*) and binormal (*open circles*) ROC curves fitted by the method of maximum likelihood to Diamond's ECG response data. The two curves are virtually indistinguishable, and further question the claim that the one-parameter logistic model can, of itself, eliminate verification bias.

that there is, for example, consistently more spread (variation) in test interpretations in diseased subjects than normal subjects in radiologic applications, leading to slopes that are less than unity (Greenes, personal communication). Swets[8] points to some examples where slopes are definitely above unity (when examining abnormal tissue cells) and others where they are below unity (brain tumors).

What other attributes does the l-parameter (logistic) curve have? The first is that it is virtually indistinguishable from the l-parameter curve based on underlying normal distributions. Both are symmetric about the negative diagonal, differing only in their extremes, where there are not sufficient data to really choose one over the other. Figure 1 shows the virtual superimposition of the fit of the two models to Diamond's data from the exercise ECG.* Indeed, this figure in itself would suggest that *if* the logistic model were invariant to verification bias (see 3 below), then for all practical purposes, the binormal model would be likewise.

The historical development of these two models is instructive. Their similarity has long been known.[8] Formal methods for fitting the logistic ROC were published by Ogilvie and Creelman[7] in 1968, one year before those for the binormal model (Dorfman and Alf, 1969).

Received from the Department of Epidemiology and Biostatistics, McGill University, Montréal, PQ, Canada H3A 1A2 (JAH), and the Department of Biostatistics, Harvard University, and Department of Biostatistics, Dana-Farber Cancer Institute, Boston, Massachusetts (CBB). Supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada (JAH) and the U.S. National Cancer Institute, Award CA-31247 (CBB).

Address correspondence and reprint requests to Dr. Hanley.

*For both curves we estimated the parameters using maximum likelihood.

Interestingly, Ogilvie and Creelman chose the logistic model for "practical rather than theoretical reasons" since "the logistic distribution is easier to handle mathematically and computationally." They stressed that "the two distributions are quite similar in shape within the typical range of empirical proportions." Dorfman and Alf argued that "a procedure assuming underlying normal distributions was preferred" because "a stable relation could not be found between the sigma ratio of signal detection theory and the analogous parameter of the logistic model." In reality, such a relation does exist for the logistic model. In any case, the binormal model has become the model of choice.

2. Statistical Properties "2) the associated standard deviation is analytically defined and readily computed." What is more important than the small differences in these two models (i.e., the logistic and the normal) is the common approach authors have taken to fit them. In using the method of maximum likelihood, both Dorfman and Alf and Ogilvie and Creelman formally recognized the multinominal structure of the rating data and the fact that successive ROC operating points are inescapably correlated; these correlations are automatically taken into account in arriving at parameter estimates and standard errors, i.e., their approaches are as important for their maximum likelihood approach to estimation as for the particular distributions chosen.

Diamond's approach to the problem involves computing a weighted average of the successive estimates of m, the weights being the inverse variances of the individual estimates. Unfortunately, the standard error produced by this method is too small, since the standard error formula for m omits the fact that the successive estimates of m are positively correlated. A simple way of recognizing why this is a problem is to imagine constructing a curve from continuous data. In such a case we could create as many estimates of m as there are data points (less one) by successively changing the cut-off value to include the next point. Successive estimates of m which are either contiguous or close to each other will be virtually identical, yet Diamond's formula would treat them as independent observations. Clearly, the more we divide the scale (by introducing more cut-off points) the more (spuriously) precise our summary estimate of m will be.

In fact, a general methodology is available for estimating parameters and computing the appropriate standard errors of the ordinal-type data typically used for constructing ROC curves.[6] This methodology includes both the logistic and normal models discussed here, as well as a variety of other models, and also permits adjustment for covariates. The application of these techniques to ROC curves has been developed by one of us (CBB).[9]

3. Verification Bias "3) the curve is completely invariant with respect to selection bias." In asserting a correction for verification bias, Diamond uses as justification the argument that selection for verification does not directly depend on true disease status, as originally suggested by one of us (CBB).[1] He then demonstrates that if the selection depends *only* on the test result, then the odds ratio, i.e., m, is invariant to the bias. He then suggests that a model based only on the odds ratio will, *ipso facto*, be invariant to bias.

Table 1 • Diagnostic Performances in All Patients and in Verified-only Patients (D = True Disease Status; R = Result of Diagnostic Test)

	R				Performance Using (as Positive)	
	− −	+ −	+ +		+ +	+ − or + +
A. If D was verified in *all* patients						
D −	600	300	100	FP:	0.100	0.400
				TN:	0.900	0.600
D +	273	419	308	TP:	0.308	0.727
				FN:	0.692	0.273
				τ+:	0.204	0.564
				τ−:	0.796	0.436
			SLOPE (m):			
					0.250	0.250
B. If D was verified in only a fraction of patients (in ¾ of R + + patients; ½ of R + − pts; ¼ of R − − pts)						
D −	150	150	75	FP:	0.200	0.600
				TN:	0.800	0.400
D +	68	210	231	TP:	0.454	0.866
				FN:	0.546	0.134
			SLOPE (m):			
					0.301	0.232

There are two problems with this approach. First, in general, selection for verification will depend on other relevant factors in addition to the test result, such as presenting signs and symptoms. Therefore, to correct for bias one must be prepared to adjust for these factors, as outlined in Begg and Greenes' general formula.[1]† Diamond's model is restricted to the assumption that selection depends only on the test result. We speculate that this assumption will be inadequate, and therefore at best a crude approximation, for most clinical applications.

The second problem is that even if the assumption that selection depends *only* on the test result is correct, Diamond's model still does not properly correct the bias. To illustrate this, we have constructed some hypothetical data for which the constant odds ratio model is true, but in which verification bias nonetheless skews our estimates.

†Our focus in our original paper[1] on the simple formula without covariates was purely for expository purposes. Similarly, in our later paper (Gray et al.[4]) we used the simple formulas in our exposition because data on relevant covariates were unavailable, as is frequently the case with retrospective data of this nature.—*CBB*

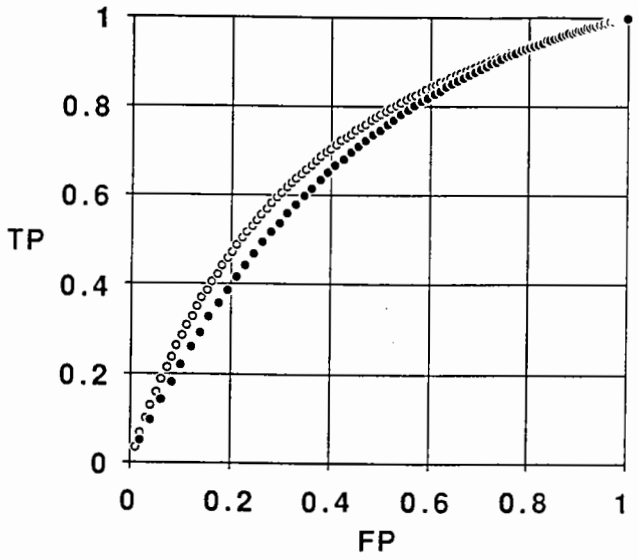

FIGURE 2. Persistence of verification bias in logistic ROC curve. The open symbols represent a one-parameter logistic ROC when all tested subjects are verified. The solid symbols represent the ROC curve that is generated if various percentages (1%–100%) of those in the different test result categories are verified.

Table 1 shows "total-population" data that generate two ROC points obeying the "constant slope" property of the Diamond model, as well as the data one would observe if only three fourths of those rated + +, half of those rated + − and one fourth of those rated − − went on to be verified. It is evident that the two slopes calculated from the verified sample differ both from one another and from the common slope of the population curve.

To appreciate that this "unsteadiness" is not simply an artifact of the small number of rating categories or of rounding, we computed an analogous curve based on 100 rating categories (fig. 2). Clearly the ROC curve in the verified sample is different from the "constant slope" curve in the overall group, even though the difference is not especially substantial.

The reason for the discrepancy is that with Diamond's method each estimate of m involves an aggregation of categories of test results for which the verification probabilities differ. For example, the first slope estimate (0.301) in Table 1*B* is obtained by aggregating the categories − − and + −. However, these categories have different verification rates, 0.25 and 0.50 respectively. Likewise, the second slope estimate involves aggregation of + − and + +. In either case this leads to bias, although the errors do tend to cancel each other out when the average slope is calculated, as Diamond has done. However, an empirical curve computed naively from a verified sample will not possess the symmetric shape mandated by Diamond's model, and any estimates of the popular area index[5] are likely to be distorted also.

Summary and Recommendations

The goals of Diamond's approach are laudable, namely to provide a simple, robust methodology that is easily calculated by hand or on a spreadsheet, and to circumvent the important, common problem of verification bias. However, we believe his model is overly restrictive. As a consequence, the method may provide poor fit due to the inability to accommodate scale terms, and may not adequately compensate verification bias by ignoring relevant covariates. Therefore, there is a danger of oversimplification and superficial analysis.

References

1. Begg CB, Greenes RA: Assessment of diagnostic tests when disease verification is subject to verification bias. Biometrics 39:207–215, 1983
2. Diamond GA: ROC Steady: a receiver operating characteristic curve that is invariant relative to selection bias. Med Decis Making 7:xxx–xxx, 1987
3. 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
4. Gray R, Begg CB, Greenes RA: Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 4:151–164, 1984
5. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36, 1982
6. McCullagh P: Regression models for ordinal data (with discussion). J R Statistical Soc B 42:109–142, 1980
7. Ogilvie JC, Creelman CD: Maximum-likelihood estimation of receiver operating characteristic curve parameters. J Math Psychol 5:377–391, 1968
8. Swets JA: Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. Psychol Bull 99:181–198, 1986
9. Tosteson AN, Begg CB: A General Regression Methodology for ROC Curve Estimation. Technical Report No. 525z. Division of Biostatistics and Epidemiology, Dana-Farber Cancer Institute, 44 Binney St., Boston, MA [unpublished]