A Goodness-of-Fit Test for a Receiver Operating Characteristic Curve from Continuous Diagnostic Test Data

By Kelly H. Zou*, Joseph L. Gastwirth†, and Barbara J. McNeil‡

Harvard Medical School and Brigham and Women's Hospital, George Washington University, and Harvard Medical School and Brigham and Women's Hospital

The receiver operating characteristic (ROC) curve is a useful way to display the performance of a medical diagnostic test for detecting whether or not a patient is diseased or healthy. The diagnostic data consist of independent random samples on continuous measurement scales from diseased and healthy populations. We propose assessing the goodness-of-fit of a model by comparing a model-based estimate with a nonparametric estimate of the area under the curve (AUC). We focus on two parametric models, so-called Bi-Normal and Bi-Weibull models, and briefly on associated semiparametric transformation models. We also consider the null hypothesis that a parametric model is valid after an unspecified monotone transformation of the measurement scales. High power of the test implies sensitivity of the AUC to model assumptions; low power implies robustness of the estimate. The test is exemplified with a data set on the diagnosis of pancreatic cancer. A simulation study of the statistical power of the test is included.

1. Introduction Diagnostic testing provides important data for medical decision making and treatment planning. The receiver operating characteristic (ROC) curve is a useful graphical and statistical tool for evaluating and comparing diagnostic tests. It is a plot of (1−specificity, sensitivity)-values at all possible two-state decision thresholds (for definitions, see [3]). Much of the ROC literature deals with ordinal rating data methods, where the values indicate the degree of certainty about the disease. For example, for cancer detection, a five-point rating scale is often employed, with 1 = definitely benign, 2 = possibly benign, 3 = probably benign, 4 = possibly malignant, and 5 = definitely malignant. Recently, diagnostic tests that yield continuous results are increasingly used. Examples of such tests are those based on tumor volume or laboratory assay such as the ELISA test for HIV infection. Note that for the ordinal rating data, it is usually assumed that there is a latent continuous variable. In this article, we confine attention to ROC curves derived from continuous tests with a moderately large number of samples of both healthy (H) and diseased (D) individuals.

There are several ways of estimating an ROC curve, along with its summary measures: First, nonparametrically, a plot of pairs of observed (1−specificity, sensitivity)-values at each possible decision threshold forms an empirical ROC curve. This is equivalent to plotting two empirical survival curves against each

*Supported in part by grants NIH RO3HS13234-01 and NCI PO1CA67165

†Supported in part by grant NSF SBR9507926

‡Supported in part by grant NIH UO1CA62462

other. Second, parametric bi-distributional models can be assumed. The frequently used Bi-Normal model assumes two independent normal distributions with different population means and variances [4]. Other models considered include Bi-Logistic [20], Bi-Exponential [6], and Bi-Gamma [5]. Finally, semiparametrically, a Bi-Normal model is assumed to hold after both measurement scales are subjected to a common unspecified monotone transformation [10, 14, 17].

With such a variety of choices of distributional assumptions and estimation methods, it is necessary to assess whether the resulting ROC curve can be relied on, i.e., whether the fitted curve is consistent with the assumptions. The goodness-of-fit (GOF) issues have been examined only for rating data (e.g., [18, 22, 20]). For continuous data, this issue has not been investigated, although tests of marginal normality, e.g., Shapiro-Wilk [13] and the z-test [12], can be applied separately to the H and D samples by assuming independence between them.

In this article we construct a large-sample GOF test based on the transformed area under the curve (AUC) to ascertain how sensitive the area is to modeling assumptions. The AUC is a popular summary measure of the overall diagnostic accuracy, ranging from 0.5 to 1, representing "chance" and "truth." The full AUC corresponds to the probability of a pair of H and D measurement values being in the correct order, and is the Mann-Whitney U statistic when estimated nonparametrically [1, 8]. Parametrically the area is an explicit function of the ROC parameters. If the modeling assumptions are valid, then parametric modeling leads to a more efficient AUC estimate. Comparing AUC estimates from both nonparametric and model-dependent methods provides a basis for assessing the fit of the parametric model. Besides AUC, popular summary measures include sensitivity at a fixed specificity [24], point of intersection [23], and maximal improvement of sensitivity [16].

An ROC curve is invariant to any monotone transformation of the H and D measurement scales. Consequently, the AUC will not distinguish between an original parametric model and the corresponding monotone transformation models. When the AUC is the main summary of overall diagnostic accuracy, taking a monotone transformation should not affect the utility of the goodness-of-fit hypothesis test in the proposed procedures.

In Section 2, we summarize background information about ROC curves. In Section 3, we propose the GOF test based on comparing a nonparametric estimate with an efficient parametric or semiparametric estimate of the AUC. Section 4 gives notations and assumptions for the empirical, the parametric Bi-Normal and Bi-Weibull, and briefly their associated semiparametric models. In Section 5, we illustrate the proposed test with a clinical example on the diagnosis of pancreatic cancer. A simulation study is presented in Section 6, providing empirical power of the proposed test at alternative hypotheses. Summary and discussions are presented in Section 7.

2. Notation and Assumptions We now formally define an ROC curve. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, be independently and identically distributed samples from H and D populations, with underlying absolutely continuous cumulative distributions functions F and G, respectively. The corresponding empirical cumulative distribution functions are denoted by $\widehat{F}_m$ and $\widehat{G}_n$. Let $N = m + n$. Throughout this article we will express the survival function as $\overline{H}(t) = 1 - H(t) = P(T \geq t)$, for any arbitrary cumulative distribution function $H(t)$.

At any pre-specified decision threshold t, the underlying ROC curve is a plot of the "true positive rate" (TPR or sensitivity), $q(t) = \overline{G}(t)$, against the "false positive rate" (FPR or 1−specificity), $p(t) = \overline{F}(t)$, for $t \in (-\infty, \infty)$. Alternatively, one may express q as a function of p such that $q(p) = \overline{G}\{\overline{F}^{-1}(p)\}$, for $p \in (0,1)$. The empirical ROC curve is defined similarly using $\widehat{F}_m$ and $\widehat{G}_n$. A popular overall summary of diagnostic accuracy is the area under the curve (AUC):

$$A = P(X < Y) = \int F(y)\, dG(y) = \int \overline{G}(x)\, dF(x). \tag{2.1}$$

Nonparametrically, the empirical area is equivalent to the Mann-Whitney Wilcoxon U-statistic:

$$\widehat{A}_N = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} 1\{X_i < Y_j\}. \tag{2.2}$$

For later reference, we note that the curve is invariant to the same monotone transformation of both H and D measurement scales. That is, let ψ be an absolutely continuous and strictly increasing function, so that $X' = \psi(X)$, $Y' = \psi(Y)$. Then $A = P(X < Y) = P(X' < Y')$ [10].

3. A Goodness-of-Fit Test We compare a nonparametric estimate $\widehat{A}_N$ with an efficient parametric estimate $\widehat{A}_P$ of the AUC. Because AUC is confined to $(0,1)$, in order to improve the large-sample approximation, a probit transformation, $W = \Phi^{-1}(A)$, of the area is recommended. We use the probit transformation because the transformed parametric Bi-Normal AUC is a simple function of the two curve parameters (see Section 4.2).

Let $\widehat{\Delta}$ denote the difference between the estimates $\widehat{W_N} = \Phi^{-1}(\widehat{A}_N)$ and $\widehat{W_P} = \Phi^{-1}(\widehat{A}_P)$. We need an estimate of its standard error. The ratio $\mathrm{Var}(\widehat{W_P})/\mathrm{Var}(\widehat{W_N})$ of the large-sample variances of these two area estimates is the asymptotic relative efficiency (ARE) of $\widehat{W_N}$ relative to $\widehat{W_P}$, assuming the parametric model is correct. From [19] and [7], this ARE can also be represented as the squared correlation coefficient ρ^2 between the two area estimates. Therefore, $\mathrm{Cov}(\widehat{W_N}, \widehat{W_P}) = \rho \cdot \sqrt{\mathrm{Var}(\widehat{W_N}) \cdot \mathrm{Var}(\widehat{W_P})} = \mathrm{Var}(\widehat{W_P})$, and so

$$\mathrm{Var}(\widehat{\Delta}) = \mathrm{Var}(\widehat{W_N}) + \mathrm{Var}(\widehat{W_P}) - 2\mathrm{Cov}(\widehat{W_N}, \widehat{W_P}) = \mathrm{Var}(\widehat{W_N}) - \mathrm{Var}(\widehat{W_P}). \tag{3.3}$$

The proposed GOF test statistic is

$$\widehat{D} = |\widehat{\Delta}|/\sqrt{\mathrm{Var}(\widehat{\Delta})} = |\widehat{\Delta}|/\sqrt{\mathrm{Var}(\widehat{W_N}) - \mathrm{Var}(\widehat{W_P})},$$

which, if the parametric model is correct, asymptotically is the absolute value of a $N(0,1)$ random variable.

4. Bi-Distributional Assumptions

4.1. *The Empirical Distributions* Without ties present in the combined data from the H and D samples, the empirical AUC is equivalent to the expression for the U-statistic given in (2.2). The variance of $\widehat{A}_N$ is well known (see [9], for example):

$$\mathrm{Var}(\widehat{A}_N) = \{p_1(1-p_1) + (n-1)(p_2 - p_1^2) + (m-1)(p_3 - p_1^2)\}/(mn),$$

where

$$p_1 = P(X_1 < Y_1) = \int \overline{G}(x)\,dF(x), p_2 = P(X_1 < Y_1, X_1 < Y_2) = \int \{\overline{G}(x)\}^2\,dF(x),$$

$$p_3 = P(X_1 < Y_1, X_2 < Y_1) = \int \{F(y)\}^2\,dG(y).$$

For any F and G, the p_i's can be compared by numerical integration, with F and G estimated empirically.

The variance of the probit transformed area estimate $\widehat{W_N} = \Phi^{-1}(\widehat{A}_N)$ is obtained by the delta method and equals:

$$\mathrm{Var}(\widehat{W_N}) = \mathrm{Var}(\widehat{A}_N)/\{\phi(W)\}^2,$$

where ϕ is the probability density function of the standard normal distribution, estimated at $W = W(A)$ with A being the underlying true AUC. In practice, we substitute $\widehat{W_P}$ for W when a particular parametric model is assumed under the null hypothesis. Similarly, we substitute $\widehat{W_S}$ for W when a semiparametric model is assumed.

4.2. *The Bi-Normal Model* Let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \tau^2)$, two normal distributions with different means and variances. Consider the common transformation of the H and D measurement scales using $\psi(t) = (t-\mu)/\sigma$. Then X' and y' still have two normal distributions: $X' = \psi(X) \sim N(0,1)$ and $Y' = \psi(Y) \sim N(\alpha, \beta^2)$, with the Bi-Normal ROC curve parameters $\alpha = (\nu - \mu)/\sigma$ and $\beta = \tau/\sigma$.

From (2.1), the Bi-Normal area is an explicit function of these curve parameters [18]:

$$A_P = \Phi\left(\alpha/\sqrt{1+\beta^2}\right)$$

with transformed area $W_P = \alpha/\sqrt{1+\beta^2}$. The parameters, (α, β), are estimated by maximizing their likelihood functions, yielding:

$$\widehat{\alpha} = (\overline{y} - \overline{x})/s_x \text{ and } \widehat{\beta} = s_y/s_x,$$

where $\overline{x} = \frac{1}{m}\sum x_i$ and $s_x^2 = \frac{1}{m}\sum(x_i - \overline{x})^2$, the sample mean and variances of the H sample, and similarly $\overline{y}$ and s_y^2 of the D sample.

The large-sample variance matrix of these estimates is the following:

$$\mathrm{Var}(\overline{x}, s_x, \overline{y}, s_y) = \mathrm{Diag}\left(\sigma^2/m, \sigma^2/(2m), \tau^2/n, \tau^2/(2n)\right).$$

From the delta method, it follows that the resulting variance matrix of $(\widehat{\alpha}, \widehat{\beta})$ is

$$\mathrm{Var}(\widehat{\alpha}) = \frac{n(\alpha^2+2) + 2m\beta^2}{2mn}, \quad \mathrm{Var}(\widehat{\beta}) = \frac{m+n}{2mn}\beta^2, \quad \mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = \frac{\alpha\beta}{2m}.$$

Finally, the large-sample variance of the estimated transformed area, again by the delta method, is

$$\text{Var}(\widehat{W_P}) = \frac{1}{1+\beta^2}\text{Var}(\widehat{\alpha}) - \frac{2\alpha\beta}{(1+\beta^2)^2}\text{Cov}(\widehat{\alpha},\widehat{\beta}) + \frac{\alpha^2\beta^2}{(1+\beta^2)^3}\text{Var}(\widehat{\beta}).$$

It is often appropriate to assume an easy-to-use transformation, ψ, such as the log transformation, to make data appear nearly Bi-Normal. One may also adopt a data-driven transformation such as a Box-Cox [2] power transformation.

4.3. *The Bi-Weibull Model* Let $X \sim \textit{Weibull}(\alpha,\kappa)$, with survival function $\overline{F}(x) = \exp(-\alpha x^\kappa)$. Likewise, let $Y \sim \textit{Weibull}(\beta,\kappa)$, with $\overline{G}(y) = \exp(-\beta y^\kappa)$. This is a Bi-Weibull model with a common shape κ but different scale parameters (α,β). Take $\psi(t) = \alpha t^\kappa$. Then X' and Y' have two exponential distributions: $X' = \psi(X) \sim \textit{Exponential}\ (1)$ and $Y' = \psi(Y) \sim \textit{Exponential}\ (1/\nu)$, with the Bi-Exponential ROC curve parameter $\nu = \alpha/\beta$.

From (2.1), the Bi-Weibull AUC is

$$A_P = P(X' < Y') = \frac{\nu}{1+\nu}. \tag{4.4}$$

Furthermore, a logit transformation, instead of a probit transformation, can be applied to the area to improve the large-sample approximation. The transformed AUC is then $W_P = \log(\nu)$. Following [11], straightforward computations lead to

$$\text{Var}(\widehat{W_P}) = N/(mn) + W_P^2/(1.6449N).$$

5. A Clinical Example We illustrate our GOF test using data on the cancer antigen assay CA125 for pancreatic cancer analyzed by [21]. There were $m = 51$ controls with pancreatitis (healthy sample) and $n = 90$ diseased cases with pancreatic cancer (diseased sample), from whom sera was collected and analyzed. Because the data were highly positively skewed, we display the two samples separately, but on a common log scale. See Figure 1 for the histograms, with dashed bars representing the healthy sample and solid bars representing the diseased sample, respectively. We first analyzed the data nonparametrically. We then assumed a parametric Bi-Normal model after a log and after a Box-Cox transformation with a power coefficient $\lambda = -0.5$. Finally, a semiparametric Bi-Normal model was fitted, with estimates calculated using the standard software program "LABROC4" [15]. The resulting ROC curves are plotted in Figure 2.

The numerical results are summarized in Table 1. The first column gives four methods for estimating the AUC. In the second and third columns, we report the estimated (α,β)-parameters for the parametric and semiparametric models. The estimated AUC's, along with their probit transformed versions, are given in the fourth and fifth columns. In the next two columns, the estimated variances of both nonparametric and parametric (or semiparametric) area estimates are presented, under the assumption that the corresponding parametric (or semiparametric) model is valid. The last two columns give the GOF test statistic $\widehat{D}$ and the corresponding two-sided p-value. The results indicated that both Box-Cox and semiparametric

Figure 1: Histograms of CA125 data on a log scale: dashed bars represent the healthy (pancreatitis) sample; solid bars represent the diseased (pancreatic cancer) sample

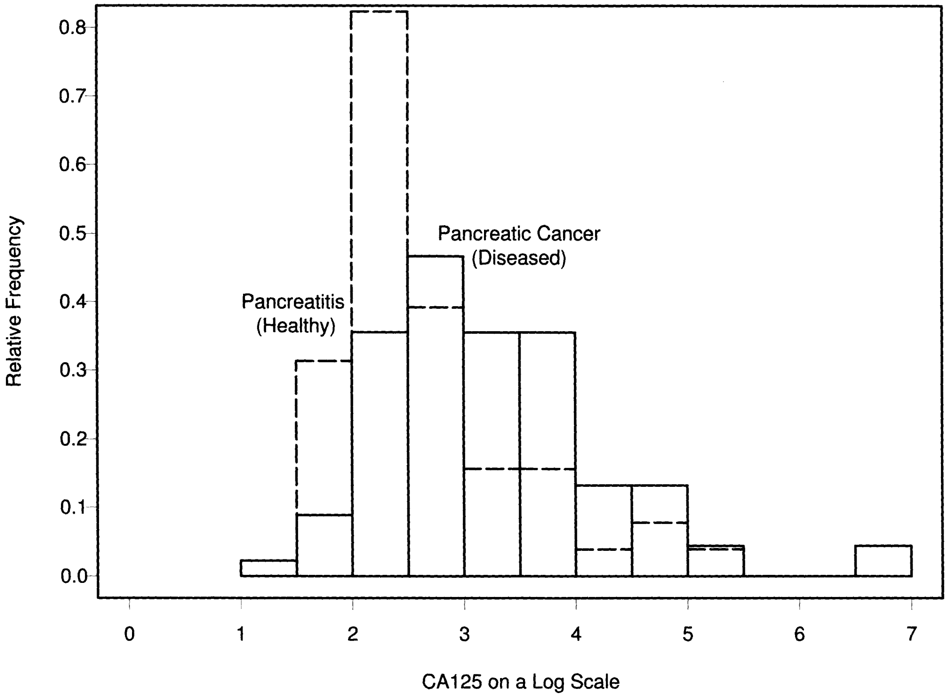


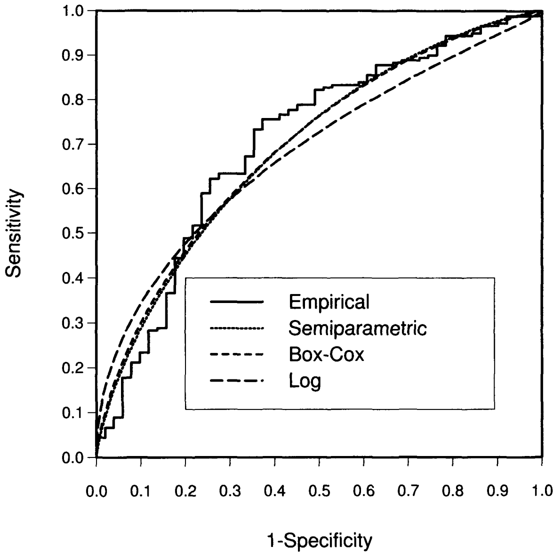
Table 1: Estimated Bi-Normal ROC curve parameters $(\hat{\alpha}, \hat{\beta})$, areas $\hat{A}$ and $\widehat{W}$, along with estimated variances, goodness-of-fit test statistic $\hat{D}$, and p-value for example data

Method	$\hat{\alpha}$	$\hat{\beta}$	$\hat{A}$	$\widehat{W}$	$\widehat{\text{Var}}(\widehat{W}_N)$	$\widehat{\text{Var}}(\widehat{W}_P)$	$\hat{D}$	p
Nonparametric	–	–	.706	.540	.0101	–	–	–
Log	.768	1.269	.683	.476	.0159	.0152	2.41	.02
Box-Cox ($\lambda = -.5$)	.739	1.032	.697	.514	.0169	.0162	1.00	.32
Semiparametric	.719	1.001	.706	.540	.0170	.0164	1.26	.21

models give satisfactory fits, with p-values of 0.32 and 0.21, respectively. However, the p-value from the log Bi-Normal model is only 0.02, indicating lack-of-fit.

These results are consistent with the direct tests of normality using the z-test. The two-sided p-values from the log Bi-Normal model are only $p_H = 9.7 \times 10^{-5}$ and $p_D = 3.5 \times 10^{-3}$. In contrast, under the Box-Cox Bi-Normal model, $p_H = 0.07$ and $p_D = 0.12$. We also apply the Bonferroni correction, using the significance level of 0.0253 for the individual tests of normality of the H and D distributions, thereby attaining an overall level of 5%. Thus, log Bi-Normal model is rejected, but not the

Figure 2: ROC curves for CA125 data by four estimation methods



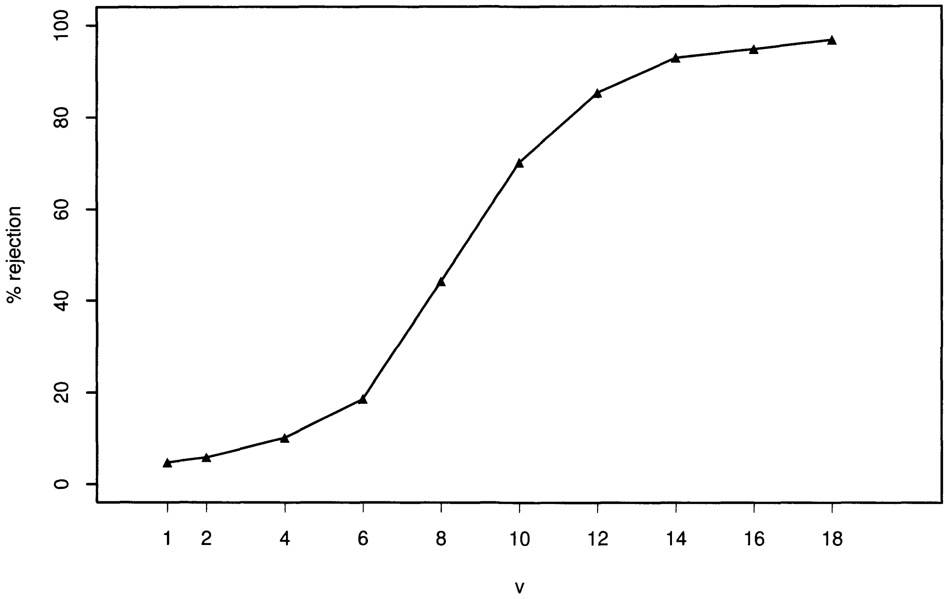
Box-Cox model. Direct testing of normality is not possible for the semiparametric model.

6. A Simulation Study For evaluating the empirical power of the GOF test, we conducted the following Monte Carlo simulation study. For typical non-normal distributions, we generated two random samples, of size 100 each, from two exponential distributions, *Exponential* (1) and *Exponential* $(1/\nu)$, respectively. This is a special case of the Bi-Weibull model. From (4.4), the underlying AUC was $\nu/(1+\nu)$. Data was analyzed assuming a Bi-Normal model. To determine the power of the GOF test for each ν, from 1 to 18, we generated 500 Monte Carlo simulations. The means and variances of the Bi-Normal distributions were estimated based on the sample values. The rejection profiles of the GOF test as a function of ν is presented in Figure 3.

7. Discussion In this article we have explored whether and when a simple procedure based on the estimated AUC can be used to assess model adequacy. We constructed a goodness-of-fit test for assessing ROC curves using probit transformed area. This test is relatively simple to apply, especially under the Bi-Normal or Bi-Weibull assumptions. We would not expect the proposed test to be as powerful as a formal GOF test based on checking all of the parametric assumptions. However, our test can provide a useful check on semi-parametric procedures for which there do not exist formal GOF tests.

The clinical example and the power study implied that the estimates of AUC are quite robust to modest departures from model assumptions. Our simulation study

Figure 3: Simulated empirical power of the Bi-Normal GOF test as a function of ν, using two samples of size $m = n = 100$ from Exponential(1) and Exponential$(1/\nu)$



indicated that the test has little power to detect moderate departures from model assumptions.

In this test the difference in the variances between $\widehat{W_N}$ and $\widehat{W_P}$ given in (3.3) can be small if nonparametric estimation method is highly efficient. In Table 2, we present the true large-sample variances of these two estimates if estimation was done by maximum likelihood parametrically. We assume various underlying areas ranging from 0.55 to 0.95, at β fixed to be 1.5 for unequal variances. Generally, the nonparametric method is relatively efficient and can even be as efficient as 0.952 when sample sizes are large (i.e., $m = n = 200$) and the ROC curve area is moderately high (0.85), i.e., neither too low (0.5) nor too high (1.0). In practice, estimating such small difference can present problems and the resulting test statistic may not exist. Aside from low power, the denominator of the proposed GOF test statistic may be close to 0. An alternative re-sample solution such as the bootstrap or jackknife may be provided instead.

In summary, procedures checking all of the parametric assumptions are preferable to the global check we have investigated. Therefore, we recommend testing parametric assumptions such as normality, if possible, before fitting a parametric model. The proposed method can serve as a diagnostic tool or check of a semiparametric model until a more formal GOF test for them is developed.

8. Acknowledgment We would like to thank Dr. W. J. Hall and the Department of Biostatistics of the University of Rochester. We also thank Dr. H.S. Wieand

Table 2: Large-sample variances of nonparametric and parametric estimates of the transformed area at the underlying parameter values, with unequal variances

A	α	$m = n$	W	Var($\widehat{W_N}$)	Var($\widehat{W_P}$)	ARE
.55	.227	25	.126	.0440	.0402	.912
		50		.0218	.0200	.921
		100		.0109	.0100	.924
		200		.0054	.0050	.927
.70	.945	25	.524	.0467	.0432	.924
		50		.0231	.0216	.934
		100		.0115	.0107	.939
		200		.0057	.0054	.941
.85	1.868	25	1.036	.0563	.0523	.930
		50		.0278	.0262	.943
		100		.0138	.0131	.950
		200		.0069	.0065	.952
.90	2.310	25	1.282	.0647	.0589	.910
		50		.0318	.0294	.926
		100		.0158	.0147	.933
		200		.0078	.0074	.938
.95	2.965	25	1.644	.0849	.0711	.837
		50		.0415	.0355	.856
		100		.0205	.0178	.866
		200		.0102	.0089	.872

of the University of Pittsburgh for providing the example data. We appreciate the comments from Dr. John E. Kolassa, Rutgers University, and from two anonymous referees.

REFERENCES

[1] D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[2] G.E.P. Box and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Ser. B*, 42:71–78, 1964.

[3] G. Campbell. General Methodology i: advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13:499–508, 1994.

[4] D.D. Dorfman and E. Alf. 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.

[5] D.D. Dorfman, K.S. Berbaum, , and C.E. Metz. Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology*, 4:138–49, 1997.

[6] W.L. England. An exponential model used for optimal threshold selection on ROC curves. *Medical Decision Making, 8:120–131, 1988.*

[7] W.J. Hall and D.J. Mathiason. On large-sample estimation and testing in parametric models. *International Statistical Review*, 58:77–97, 1990.

[8] J. Hanley and B.J. McNeil. The meaning and use of the area under a ROC curve. *Radiology*, 143:27–36, 1982.

[9] T.P. Hettmansperger. *Statistical Inference Based on Ranks.* Krieger Publishing Co., Malabar, FL, 1991.

[10] F. Hsieh and B.W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24:25–40, 1996.

[11] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*, volume 1. Wiley-Interscience, New York, 2nd edition, 1994.

[12] C.-C. Lin and G.S. Mudholkar. A simple test for normality against asymmetric alternatives. *Biometrika*, 67:455–461, 1980.

[13] A. Madansky. *Prescription for Working Statistician.* Springer-Verlag, New York, 1988.

[14] C.E. Metz, B.A. Herman, and J.H. Shen. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, 17:1033–53, 1998.

[15] C.E. Metz, J.-H. Shen, P.-L. Wang, and H.B. Kronman. *Fortran program LABROC4*, 1997.

[16] A.J. O'Malley, K.H. Zou, J.R. Fielding, and C.M.C. Tempany. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: Prostate biopsy and spiral ct of ureteral stones. *Academic Radiology*, 8:713–725, 2001.

[17] M.S. Pepe. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54:124–35, 1998.

[18] J.A. Swets and R.M. Pickett. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press, New York, 1982.

[19] C. van Eeden. The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *Annals of Mathematical Statistics*, 34:1442–1451, 1963.

[20] S.J. Walsh. Goodness-of-fit issues in ROC curve estimation. *Medical Decision Making*, 19:193–201, 1999.

[21] S. Wieand, M.H. Gail, B.R. James, and K.L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76:585–592, 1989.

[22] X.H. Zhou. Testing an underlying assumption on a ROC curve based on rating data. *Medical Decision Making*, 8:197–203, 1998.

[23] K.H. Zou and W.J. Hall. Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics*, 27:621–631, 2000.

[24] K.H. Zou, W.J. Hall, and Shapiro D.E. Smooth nonparametric receiver operating characteristic curves for continuous diagnostic test data. *Statistics in Medicine*, 16:2143–2156, 1997.

Department of Health
Care Policy
Harvard Medical School
180 Longwood Ave.
Boston, MA 02115 USA

Department of Radiology
Brigham and Women's
Hospital
Harvard Medical School
75 Francis Street
Boston, MA 02115 USA
zou@bwh.harvard.edu

Department of Statistics
George Washington
University
Washington, DC 20052 USA
gastwirj@mail.nih.gov

Department of Health
Care Policy
Harvard Medical School
180 Longwood Ave.
Boston, MA 02115, US.A.

Department of Radiology
Brigham and Women's
Hospital
Harvard Medical School
75 Francis Street
Boston, MA 02115
mcneil@hcp.med.harvard.edu