Estimating and Comparing Diagnostic Tests' Accuracy When the Gold Standard Is Not Binary[1]

Nancy A. Obuchowski, PhD

Rationale and Objectives. Investigators often need to assess the accuracies of diagnostic tests when the gold standard is not binary-scale. The objective of this article is to describe nonparametric estimators of diagnostic test accuracy when the gold standard is continuous, ordinal, and nominal scale.

Materials and Methods. A nonparametric method of estimating and comparing the area under receiver operating characteristic (ROC) curves, proposed by DeLong et al, is extended to situations in which the gold standard is not binary. Two examples illustrate the methods.

Results. Measures of diagnostic test accuracy, their variance, and tests for comparing two diagnostic tests' accuracies in paired designs are presented for situations in which the gold standard is continuous, ordinal, and nominal scale. These summary measures of diagnostic test accuracy are analogous in form and interpretation to the area under the ROC curve.

Conclusion. Dichotomizing the outcomes of a gold standard so that traditional ROC methods can be applied can lead to bias. The methods described here are useful for assessing and comparing summary test accuracy when the gold standard is not binary scale. They have limitations similar to other summary indices.

Key Words. Diagnostic accuracy; gold standard; nonparametric statistics; receiver operating characteristic curve; ROC analysis.

Receiver operating characteristic (ROC) analysis has become the standard method for characterizing a diagnostic test's accuracy and for comparing accuracies of competing diagnostic tests (1–4). In traditional ROC analysis, there exists a gold or reference standard, independent of the diagnostic test(s), which provides a binary result as to the presence or absence of disease. Patients undergo both the gold standard procedure and the diagnostic test(s). The diagnostic test results are then compared with the results of the binary-scale gold standard to estimate accuracy (ie, sensitivity and specificity) at various cutpoints of the diagnostic test's results.

Some examples of binary-scale gold standards are surgery for determining whether or not a cancer has spread to the lymph nodes, invasive catheter angiography for determining if a significant stenosis is present or absent, and biopsy for determining if a lesion is benign or malignant. There is a vast literature on parametric, nonparametric, and semiparametric statistical methods for estimating and comparing ROC curves and their associated indices when the gold standard is binary scale (5–12).

The application of ROC curves has expanded to many fields. Even in the diagnostic radiology setting, there are many applications in which we want to characterize and compare the accuracies of diagnostic tests, but the gold standard is not binary scale. One approach might be to construct a binary scale from the nominal-, ordinal- or continuous-scale gold standard so that traditional ROC

Acad Radiol 2005; 12:1198–1204

[1] From the Department of Quantitative Health Sciences/Wb4, The Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195. Received March 3, 2005; revision received and accepted May 17, 2005. **Address correspondence to:** N.A.O. e-mail: nobuchow@bio.ri.ccf.org

doi:10.1016/j.acra.2005.05.013

Table 1
Comparison of Summary Measures of Diagnostic Test Accuracy

Gold Standard	Estimator of Accuracy	
Binary	$\hat{\theta} = \frac{1}{n_t n_s}\sum_{(i=1)}^{n_t}\sum_{(j=1)}^{n_s}\Psi(X_{it},X_{js})$	where: $\Psi = 1$if $X_{it} > X_{js}$ $\Psi = 0.5$if $X_{it} = X_{js}$ $\Psi = 0$if $X_{it} < X_{js}$.
Continuous	$\hat{\theta}' = \frac{1}{N(N-1)}\sum_{(i=1)}^{N}\sum_{(j=1)}^{N}\Psi'(X_{it},X_{js})$	where: $\Psi' = 1$if t>s and $X_{it} > X_{js}$ or s > t and $X_{js} > X_{it}$ $\Psi' = 0.5$if t = s or $X_{js} = X_{it}$ $\Psi' = 0$ otherwise.
Ordinal	$\hat{\theta}'' = 1.0 - \sum_{(t=1)}^{T}\sum_{(s>t)}^{T} w_{ts}\cdot L(t,s)\cdot(1-\hat{\theta}_{ts})$	where: is the binary-scale estimator
Nominal	SAME AS ORDINAL	where $\hat{\theta}_{ts} = \frac{1}{(n_t\cdot n_s)}\sum_{(i=1)}^{n_t}\sum_{(k=1)}^{n_s}\Psi(D_{(t-s)tj},D_{(t-s)sk})$ $\Psi(D_{(t-s)tj},D_{(t-s)sk}) = 1$ if $D_{(t-s)tj} > D_{(t-s)sk}$ $\Psi(D_{(t-s)tj},D_{(t-s)sk}) = 1/2$ if $D_{(t-s)tj} = D_{(t-s)sk}$ $\Psi(D_{(t-s)tj},D_{(t-s)sk}) = 0$ if $D_{(t-s)tj} < D_{(t-s)sk}$.

methods can be used. This approach, however, often dilutes the measure of accuracy, can conceal important relationships between the gold standard and diagnostic test, and can lead to biased estimates of accuracy (13,14).

In this article, we describe nonparametric methods for estimating and comparing the accuracies of diagnostic tests when the gold standard is not binary scale. Accuracy is defined using indices analogous to the traditional area under the ROC curve.

We illustrate the methods using two examples. In the first example magnetic resonance imaging (MRI) is used to score regions of the heart according to the amount of visible scarring. The amount of scar is compared to the results of the gold standard, positron emission tomography (PET), where each region was read as normal, hibernating, ischemic, or necrotic (ordinal-scale gold standard). In the second example, computed tomography (CT) is used to measure the size of renal tumors. Patients also underwent surgery during which the renal tumor was measured. The surgical measurement is considered the gold standard (continuous-scale gold standard).

ESTIMATING AND COMPARING DIAGNOSTIC TEST ACCURACY

We first review the nonparametric estimation of diagnostic test accuracy when the gold standard is binary. We use the methods of DeLong et al (6) to estimate the variance and covariance of the estimated accuracy because we can easily extend the method to situations where the gold standard is not binary-scale. Table 1 summarizes the various related measures of accuracy described in this article.

Binary-Scale Gold Standard

Let X_{it} denote the result of the diagnostic test for the ith patient with disease status t (as determined by the gold standard), and let X_{js} denote the result of the diagnostic test for the jth patient with disease status s. We can think of status t as being "disease present" and status s as "disease absent." X can be a rating-scale confidence score assigned by a reader according to the perceived likelihood that disease is present (eg, 1 = "definitely no disease," 2 = "probably no disease," 3 = "possibly disease," 4 = "probably disease," or 5 = "definitely disease"), a percent confidence score assigned by a reader (ie, a value between 0 and 100, where 0% means no confidence in the presence of disease and 100% is complete confidence in the presence of disease), or an objective measurement (eg, attenuation value, serum iron concentration).

A nonparametric estimator of diagnostic test accuracy is

$$\hat{\theta} = \frac{1}{n_t n_s}\sum_{(i=1)}^{n_t}\sum_{(j=1)}^{n_s}\Psi(X_{it}, X_{js}) \qquad (1)$$

where n_t and n_s are the number of patients with disease status t and s in the study sample, respectively. We assume that higher diagnostic test results are indicative of greater suspicion for disease status t; then the kernel is

defined as

$$\Psi = 1 \text{ if } X_{it} > X_{js}$$

$$\Psi = 0.5 \text{ if } X_{it} = X_{js}$$

$$\Psi = 0 \text{ if } X_{it} = X_{js}$$

$\hat{\theta}$ is an estimator of the area under the ROC curve (5,6). It is interpreted as: of two randomly chosen patients, one with disease and one without, θ is the probability that the patient with disease will have a higher result on the diagnostic test than the patient without disease (5).

DeLong et al (6) used structural components to estimate the variance of $\hat{\theta}$ and the covariance between the accuracies of two diagnostic tests. Details are given in the appendix.

Continuous-Scale Gold Standard

Suppose now that the gold standard is continuous scale. For example, tumor diameter in centimeters is measured noninvasively with CT (the diagnostic test) and compared with the diameter measured at surgery (the continuous-scale gold standard). Here, X_{it} denotes the result of the diagnostic test for the ith patient who has a continuous-scale gold-standard outcome of t. As with a binary-scale gold standard, X can be a subjective confidence score assigned by a reader or an objective measurement.

Obuchowski (14) proposed the following estimator of diagnostic test accuracy

$$\hat{\theta}' = \frac{1}{N(N-1)} \sum_{(i=1)}^{N} \sum_{(j=1)}^{N} \Psi'(X_{it}, X_{js}) \qquad (2)$$

where $i \neq j$, N is the total number of patients in the study sample and

$$\Psi' = 1 \text{ if } t > s \text{ and } X_{it} > X_{js} \text{ or } s > t \text{ and } X_{js} > X_{it}$$

$$\Psi' = 0.5 \text{ if } t = s \text{ or } X_{js} = X_{it}$$

$$\Psi' = 0 \text{ otherwise.}$$

This estimator is a linear function of Kendall's tau.

The interpretation of θ' is similar to the usual (binary-scale gold standard) measure of accuracy: of two randomly chosen patients, θ' is the probability that the patient with a higher gold standard outcome (eg, larger renal tumor) has a higher diagnostic test result than the patient with a lower gold standard outcome (eg, smaller tumor).

The estimator of the variance of $\hat{\theta}'$

$$\hat{Var}(\hat{\theta}') = \frac{1}{(N/2)((N/2-1)} \sum_{(i=1)}^{N} [V(X_{it}) - \hat{\theta}']^2 \qquad (3)$$

where the structural components are now defined as

$$V(X_{it}) = \frac{1}{(N-1)} \sum_{(j=1)}^{N} \Psi'(X_{it}, X_{js}) \left(i \neq j\right) \qquad (4)$$

Note that the variance estimator in Eq 3 is similar to the variance estimator for binary-scale gold standards (see Eq A2), but the estimator in Eq A2 has two sums of squares terms, one for diseased and one for nondiseased patients.

To compare two diagnostic tests' accuracies, θ'^1 and θ'^2, we use the test statistic given in Eq A6, where the estimator of the covariance is

$$\hat{Cov}(\hat{\theta}'^1, \hat{\theta}'^2) = \frac{1}{(N/2)((N/2)-1)} \sum_{(i=1)}^{N} [V(X_{it}^1) - \hat{\theta}'^1][V(X_{it}^2) - \hat{\theta}'^2] \quad (5)$$

Ordinal-Scale Gold Standard

With an ordinal-scale gold standard, we assume that there are T total number of categories of the gold standard outcome. For example, MRI (the diagnostic test) was used to measure the scarring in regions of the heart muscle. Patients also underwent PET, the gold standard, where the regions of the heart were rated as normal (t = 1), hibernating (t = 2), ischemic (t = 3), or necrotic (t = 4). Here, T = 4. As in the previous sections, we let X_{it} denote the result of the diagnostic test for the ith patient who has a gold-standard outcome of t, where t = 1, 2, . . . T. X can be a subjective confidence score or objective measurement.

The estimator in Eq 2 could be applied here. We note, however, that in the special case where there are only two outcomes for the gold standard, the estimator in Eq 2 is *not* equivalent to the estimator in Eq 1. The difference is that for the estimator in Eq 1, we do not compare two patients with the same gold standard outcome, yet for the estimator in Eq 2, we assign a value of 0.5 to patients with the same gold standard outcome. For a continuous-scale gold standard, we expect few pairs of patients with the same gold standard outcome. However, for an ordinal-

scale gold standard, depending on the number of categories, there may be many pairs of patients with the same gold standard outcome. Because these pairs will be assigned a value of 0.5 (ie, $\Psi' = 0.5$), the estimated accuracy will regress toward the null value of the accuracy index (ie, = 0.5) as the number of categories decreases.

An alternative estimator for ordinal-scale gold standards is given in Eq 6. This estimator is the same estimator proposed by Obuchowski et al (15) for nominal-scale gold standards.

$$\theta'' = 1.0 - \sum_{(t=1)}^{T} \sum_{(s>t)}^{T} w_{ts} \cdot L(t,s) \cdot (1 - \hat{\theta}_{ts}) \qquad (6)$$

In Eq 6, $\hat{\theta}_{ts}$ is the estimator of diagnostic test accuracy for discriminating between gold standard outcome t and gold standard outcome s. The estimator in Eq 1 is used to estimate $\hat{\theta}_{ts}$. w_{ts} is a weight and $L(t,s)$ is a penalty function; both are defined below. n_t is the number of patients with gold standard outcome t such that the total sample size is given by$N = \sum_{t=1}^{T} n_t$

The weights can be based on the relative sample sizes in the study sample as follows:

$$w_{ts} = \left[n_t n_s\right] \Big/ \left[\sum_{(i=1)}^{T} \sum_{(l>i)}^{T} n_i n_l\right] \qquad (7)$$

A population-based weighting scheme has been described as well (15).

$L(t,s)$ is a penalty function with values between zero and one. $L(t,s) = 1$ is the greatest penalty for the diagnostic test's inability to distinguish truth state t from truth state s; $L(t,s) = 0$ indicates no penalty. For example, we might assign the greatest penalty to the diagnostic test's inability to distinguish normal from necrotic heart tissue (eg, L(1,4) = 1.0). We might assign less penalty to neighboring truth states. One possible scheme would be L(t,t + 1) = 0.25, L(t,t + 2) = 0.5, and L(t,t + 3) = 1.0.

Note that when there are only two outcomes of the gold standard (ie, T = 2) and L(t,s) = 1.0, the estimator in Eq 6 *does* reduce to the estimator in Eq 1.

θ'' is interpreted as: Of two randomly chosen patients sampled from different truth states according to the weighting scheme w, θ'' is the probability that the patient with the higher truth state (eg, more damage to the heart) has a higher test score that the patient with the lower truth state, where the penalty of misclassifying patients is defined by the loss function, L.

The variance of θ'' is the sum of the estimated variances and covariances of the θ_{ts}'s weighted appropriately (15). See the Appendix for details.

Nominal-Scale Gold Standard

Obuchowski et al (15) proposed a measure of diagnostic accuracy when the gold standard is nominal scale. The estimator of diagnostic test accuracy for nominal-scale gold standards is given in Eq 6, and its variance and covariance estimators are given in Eq A7 and A8, respectively. The main difference in approach between diagnostic tests with ordinal- versus nominal-scale gold standards is how the test results (ie, confidence scores) are collected. For nominal-scale gold standards, X_{it} is not a scalar variable; rather X_{tj} is a
(1 × I) vector of confidence scores, one score for each of the T truth states (15,16). This was illustrated with a study of pediatric abdominal pain. Physicians were asked to provide a *differential diagnosis* for each patient, where the choices were normal, appendicitis, intestinal obstruction, and gastroenteritis. For each patient, physicians assigned confidence scores to these four diagnoses, where the sum of the four confidence scores equaled 100.

Define $D_{(t-s)tj}$ as the difference in confidence scores assigned to truth state t and truth state s (s > t) for the jth patient with disease status t. The diagnostic accuracy for discriminating between truth states t and s is (15)

$$\theta_{ts} = \frac{1}{(n_t \cdot n_s)} \sum_{(j=1)}^{n_t} \sum_{(k=1)}^{n_s} \Psi(D_{(t-s)tj}, D_{(t-s)sk}) \qquad (8)$$

where the kernel $\Psi(D_{(t-s)tj}, D_{(t-s)sk}) = 1$ if $D_{(t-s)tj} > D_{(t-s)sk}$, $\Psi(D_{(t-s)tj}, D_{(t-s)sk}) = 1/2$ if $D_{(t-s)tj} = D_{(t-s)sk}$, and $\Psi(D_{(t-s)tj}, D_{(t-s)sk}) = 0$ if $D_{(t-s)tj} < D_{(t-s)sk}$.

The loss function, $L(t,s)$ in the estimator of diagnostic accuracy (Eq 6) should be based on the relative errors of misclassifying patients. For example, if a patient with appendicitis is incorrectly diagnosed as normal or with simple gastroenteritis, we might assign the greatest penalty, $L(t,s) = 1.0$. We assign less penalty to the inability to distinguish gastroenteritis from normal, $L(t,s) = 0.5$.

Here, $\hat{\theta}''$ is interpreted as: of two randomly chosen patients, sampled from different truth states according to the weighting scheme w, $\hat{\theta}''$ is the probability of correctly ranking the patients, where the penalty of misclassifying patients is defined by the loss function, L.

EXAMPLES

MRI to Diagnose Heart Damage After MI

After a myocardial infarction (MI), patients underwent an MRI to diagnose the damage to the heart tissue. The MRI image was scored by a radiologist using a six-point ordinal scale describing the amount of scar present in the most damaged segment of the heart: 0 = normal, 1 = 1–24%, 2 = 25–49%, 3 = 50–74%, 4 = 75 = 99%, and 5 = 100% scarring. The patients also underwent a PET scan, which is considered the gold standard for evaluating the heart tissue; the most damaged section was scored as normal, ischemic, hibernating, or necrotic, representing increasing damage of the heart tissue.

Table 2 summarizes the PET outcomes and MRI results from 241 fictitious patients. There are 114 normal, 21 ischemic, 19 hibernating, and 87 necrotic hearts. Table 3 summarizes the six estimates of accuracy from the pairwise comparisons of the four truth categories. From Table 3, we observe that MRI has poor accuracy at discriminating between normal and ischemic and between hibernating and necrotic.

Using a loss function where $L(t,s) = 1$ for all pairs of t and s ($t \neq s$), the estimated accuracy of MRI is 0.720 with SE of 0.027 and asymptotic 95% CI of [0.667–0.773]. Thus, of two randomly chosen patients with different heart muscle damage, MRI has probability of 0.72 of revealing more scar in the patient with more tissue damage. Using a loss function where L(t,t + 1) = 0.25, L(t,t + 2) = 0.5, and L(t,t + 3) = 1.0, the estimated overall accuracy of MRI is 0.825 with SE of 0.022 and 95% asymptotic CI of [0.782, 0.868].

CT to Measure Renal Mass Size

As part of a much larger study, Herts (personal communication, March 2004) measured and reported the largest diameter of 74 papillary renal masses seen on CT images taken immediately prior to surgery (Table 4). The mean difference in size between CT and surgery was −0.15 cm with standard deviation of 1.06 and 95% CI of −0.40 to 0.10.

Using Eq 2, the estimated accuracy of CT is 0.871, with standard error of 0.021. The asymptotic 95% CI for the accuracy of CT is 0.830−0.912. Thus, of two randomly chosen renal masses, there is a 87% chance that the larger renal mass (as determined at surgery) will also have a larger measured diameter on CT than the smaller renal mass.

Also in Table 4 are the measurements made for each mass based on a fictitious competing diagnostic test. We use these data to illustrate the proposed method for comparing the accuracies of two diagnostic tests in a paired design. The mean difference in size between the second diagnostic test and surgery is 0.27 cm, with a standard deviation of 0.73 and 95% CI of 0.10−0.44. The second test significantly overestimates renal mass size. The estimated accuracy of the second test is 0.957 with standard error of 0.007. The estimated covariance between the two tests' estimated accuracies is 0.000048 (from Eq 5). The value of the test statistic for assessing the null hypothesis of no difference in accuracy between the two tests is $z = (0.871-0.957)/(\sqrt{0.021^2+0.007^2-2\times0.000048}) = -4.33$, which is statistically significant at the 0.05 level. Thus, although the second test overestimates renal mass size, relative to the size determined at surgery, it discriminates between masses of different sizes better than CT.

Table 2
MRI Results Versus PET Outcomes for 224 Patients

PET Outcome: MRI score:	Normal	Ischemic	Hibernating	Necrotic
0 = normal	40	6	0	13
1 = 1–24%	35	8	5	8
2 = 25–49%	27	4	4	20
3 = 50–74%	10	2	4	12
4 = 75–99%	2	1	5	14
5 = 100%	0	0	1	20
Total	114	21	19	87

Table 3
Estimated Accuracy of MRI for Distinguishing Four Grades of Heart Tissue Damage

	Ischemic	Hibernating	Necrotic
Normal	0.527 (0.066)	0.807 (0.050)	0.770 (0.034)
Ischemic	—	0.787 (0.069)	0.752 (0.050)
Hibernating	—	—	0.532 (0.063)

Entries are estimated accuracies and their standard errors.

DISCUSSION

This article describes a group of related methods for summarizing diagnostic test accuracy when the gold standard is not binary scale. The measures of accuracy have interpretations analogous to the area under the ROC curve, although they are not associated with a ROC-type curve. The methods are nonparametric and thus make no assumptions about the distribution of the diagnostic test results or gold standard outcomes.

Table 4
CT and Surgery Measurements of 74 Renal Masses

CT/Fi	SURG	CT/Fi	SURG	CT/Fi	SURG	CT/Fi	SURG
3.9/3.0	3.3	1.0/0.9	1.3	2.4/2.8	3.3	3.0/2.9	2.8
2.0/2.2	1.9	1.1/0.8	0.2	16.0/17.6	11.5	3.5/3.6	3.2
3.7/4.1	4.0	3.1/3.9	3.7	4.5/8.4	8.0	2.6/3.5	3.2
3.1/3.6	3.5	3.9/4.0	3.7	4.0/3.8	3.0	5.0/5.2	5.0
3.0/2.9	3.0	2.7/2.8	2.7	5.0/4.9	5.0	2.5/2.5	2.5
4.4/4.4	4.5	2.0/2.2	2.2	4.0/8.0	7.5	2.3/3.2	2.8
5.5/6.7	6.0	2.8/3.2	2.8	5.2/6.0	6.2	2.8/2.5	2.5
5.4/5.1	4.8	6.6/6.7	6.3	2.0/4.5	4.0	3.8/3.7	3.5
4.0/4.2	4.0	7.5/6.6	6.0	1.0/1.1	1.2	1.0/1.1	1.0
5.1/6.2	6.0	3.6/2.9	2.7	3.4/3.0	2.8	3.4/2.7	2.5
4.5/4.8	4.5	1.3/1.4	1.2	5.2/5.1	5.0	3.7/4.4	4.0
7.2/8.3	8.0	4.9/7.1	7.2	8.2/8.8	8.5	3.7/4.1	4.1
2.7/2.2	2.0	3.6/3.7	3.5	4.4/4.3	4.0	2.6/2.6	2.5
2.4/2.3	2.5	1.2/1.3	1.0	6.0/6.3	6.0	4.9/4.9	4.5
1.4/1.3	1.3	8.3/8.6	8.5	2.0/2.6	2.4	3.0/3.2	3.0
1.6/1.9	1.5	2.5/2.3	2.0	4.0/4.2	4.0	4.7/6.8	6.5
2.1/2.2	2.3	2.7/4.9	4.5	3.7/6.6	6.0	1.5/2.1	1.8
1.4/2.0	2.0	2.2/2.8	2.4	2.5/4.2	4.0		
3.6/3.3	.5	2.9/2.4	2.5	3.7/2.9	2.6		

CT: renal mass size as measured on computed tomography images; Fi: renal mass size as measured on a fictitious test; SURG: renal mass size as measured at surgery, considered the gold standard.

Measurements are in centimeters.

Several other measures of assessing diagnostic test accuracy when the gold standard is not binary scale have been proposed in the literature. Kijewski et al (17) were one of the first to consider the problem of a gold standard with an ordinal scale. They proposed maximum likelihood estimation of the T distinct distributions, but no summary measure of accuracy was described.

Mossman (18) described a three-dimensional ROC curve when the gold standard has exactly three nominal truth states. He used a confidence scoring scheme similar to the differential diagnosis format. Dreiseitl et al (19) developed nonparametric estimates for the volume of the three-way ROC curve, including estimates of the standard errors and covariances. The method described here for nominal-scale gold standards is similar to the ideas of Mossman (18) and Dreiseitl et al (19), but does not involve constructing an ROC hypersurface (20) and is not limited to three truth states.

Rockette (21) used a weighted average of ROC areas as a summary measure of accuracy when the gold standard is nominal scale. He compared each disease state to a control group, then took a weighted average of these pairwise estimates. The summary measure of accuracy for nominal-scale gold standards described in this article, however, allows for *all* pairwise estimates of accuracy.

An important limitation of the measures of diagnostic test accuracy described in this article is that these are *summary* indices. They may conceal important relationships between the gold standard and diagnostic test. This problem can be overcome by plotting the ROC curves of pairwise estimates of test accuracy (for ordinal- and nominal-scale gold standards) and by plotting the continuous-scale gold standard outcome versus the diagnostic test results.

Last, we note that asymptotic 95% CIs based on the variance estimators described here have coverage close to the nominal level for sample sizes of about 50 or larger (for continuous-scale gold standard) (14) or 20 patients or more per truth state (for ordinal- or nominal-scale gold standards) (15). For smaller sample sizes, bootstrap CIs may be preferred.

REFERENCES

1. Lusted LB. Signal detectability and medical decision-making. Science 1971; 171:1217–1219.
2. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.
3. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: Wiley & Sons, 2002.
4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford Statistical Science Series. Oxford, UK: Oxford University Press, 2003.
5. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
6. Hanley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843.
7. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845.
8. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory—a direct solution. Psychometrika 1968; 33: 117–124.
9. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. J Math Psychol 1969; 6:487–496.
10. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psychol 1980; 22:218–243.
11. McClish DK. Analyzing a portion of the ROC curve. Medi Decision Making 1989; 9:190–195.
12. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics 1998; 54:124–135.
13. Obuchowski NA, Lieber ML, Wians FH. ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem 2004; 50: 1118–1125.
14. Obuchowski NA. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. Statistics in Medicine. In press.
15. Obuchowski NA, Goske MJ, Applegate KE. Assessing physicians' accuracy in diagnosing pediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. Stat Med 2001; 20:3261–3278.
16. Obuchowski NA, Applegate KE, Goske MJ, et al. The "differential diagnosis" for multiple diseases: comparison to the binary-truth state experiment in two empirical studies. Acad Radiol 2001; 8:947–954.
17. Kijewski MF, Swensson PG, Judy PF. Analysis of rating data from multiple-alternative tasks. J Math Psychol 1989; 33:428–451.
18. Mossman D. Three-way ROCs. Med Decision Making 1999; 19:78–89.

19. Dreiseitl S, Ohno-Machado L, Binder M. Comparing three-class diagnostic tests by three-way ROC analysis. Med Decision Making 2000; 20:323–331.
20. Edwards DC, Metz CE, Kupinski MA. Ideal observers and optimal ROC hypersurfaces in N-class classification. IEEE Trans Med Imaging 2004; 23:891–895.
21. Rockette HE. An index of diagnostic accuracy in the multiple disease setting. Acad Radiol 1994; 1:283–286.

APPENDIX

For the binary-scale gold standard situation, DeLong et al (6) define the structural component of the ith patient with disease status t as

$$V_t(X_{it}) = \frac{1}{n_s}\sum_{(j=1)}^{n_s}\Psi(X_{it}, X_{js}) \quad \text{(A1)}$$

Similarly, for the jth patient with disease status s, DeLong et al define

$$V_s(X_{js}) = \frac{1}{n_t}\sum_{(i=1)}^{n_t}\Psi(X_{it}, X_{js})$$

Then the estimator for the variance of is $\hat{\theta}$

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\theta}\right) = \frac{1}{n_t(n_t-1)}\sum_{(i=1)}^{n_t}[V_t(X_{it}) - \hat{\theta}]^2 + \frac{1}{n_s(n_s-1)}\sum_{(j=1)}^{n_s}[V_s(X_{js}) - \hat{\theta}]^2 \quad \text{(A2)}$$

For comparing the accuracies of two diagnostic tests, DeLong et al (6) proposed the following estimator of the covariance

$$\hat{Cov}(\hat{\theta}^1, \hat{\theta}^2) = \frac{1}{n_t}S_t^{1,2} + \frac{1}{n_s}S_s^{1,2} \quad \text{(A3)}$$

where the superscripts denote the two diagnostic tests, and

$$S_t^{1,2} = \frac{1}{n_t - 1}\sum_{(i=1)}^{n_t}[V_t(X_{it}^1) - \theta^1][V_t(X_{it}^2) - \theta^2] \quad \text{(A4)}$$

and

$$S_s^{1,2} = \frac{1}{n_s - 1}\sum_{(j=1)}^{n_s}[V_s(X_{js}^1) - \theta^1][V_s(X_{js}^2) - \theta^2] \quad \text{(A5)}$$

To test the null hypothesis that $\theta^1 = \theta^2$, versus the alternative hypothesis that $\theta^1 \neq \theta^2$, the following test statistic is computed

$$z = \frac{\hat{\theta}^1 - \hat{\theta}^2}{\sqrt{\hat{Var}\left(\hat{\theta}^1\right) + \hat{Var}\left(\hat{\theta}^2\right) - 2\hat{Cov}\left(\hat{\theta}^1, \hat{\theta}^1\right)}} \quad \text{(A6)}$$

and compared with a standard normal distribution.

For the ordinal- and nominal-scale gold standard situations, the variance of $\hat{\theta}''$ is the sum of the estimated variances and covariances of the $\hat{\theta}_{ts}$ s weighted appropriately (15), as follows:

$$\hat{Var}(\hat{\theta}'') = \sum_{(t=1)}^{T}\sum_{(s>t)}^{T}\sum_{(i=1)}^{T}\sum_{(l>i)}^{T} w_{ts}w_{il}L(t,s)L(i,l)\hat{cov}(\hat{\theta}_{ts}, \hat{\theta}_{il}) \quad \text{(A7)}$$

which can be estimated using Eq A2 (see Eq A4 and A5);

$$\hat{cov}(\hat{\theta}_{ts}, \hat{\theta}_{ts}) = \hat{Var}(\hat{\theta}_{ts})$$

$$\hat{cov}(\hat{\theta}_{ts}, \hat{\theta}_{tl}) = \frac{1}{n_t}S_t^{ts,tl}$$

$$\hat{cov}(\hat{\theta}_{ts}, \hat{\theta}_{is}) = \frac{1}{n_s}S_t^{ts,is}$$

$$\hat{Cov}(\hat{\theta}_{ts}, \hat{\theta}_{it}) = \frac{1}{n_t(n_t-1)}\sum_{(j=1)}^{n_t}[V_t(X_{jt}^{ts}) - \hat{\theta}^{ts}][V_t(X_{jt}^{it}) - \hat{\theta}^{it}]$$

and $\hat{cov}(\hat{\theta}_{ts}, \hat{\theta}_{il} = 0$ when $t \neq i, l$ and $s \neq i, l$

An estimator of the covariance between two diagnostic tests' accuracies is the sum of the covariances of the $\hat{\theta}_{ts}$ s, weighted appropriately

$$\hat{Cov}(\hat{\theta}''^1, \hat{\theta}''^2) = \sum_{(t=1)}^{T}\sum_{(s>t)}^{T}\sum_{(i=1)}^{T}\sum_{(l>i)}^{T} w_{ts}w_{il}L(t,s)L(i,l)\,\hat{cov}(\hat{\theta}''^1 ts, \hat{\theta}''^2 il) \quad \text{(A8)}$$