Multireader, Multicase Receiver Operating Characteristic Analysis:

An Empirical Comparison of Five Methods[1]

Nancy A. Obuchowski, PhD, Sergey V. Beiden, PhD, Kevin S. Berbaum, PhD, Stephen L. Hillis, PhD, Hemant Ishwaran, PhD, Hae Hiang Song, PhD, Robert F. Wagner, PhD

Rationale and Objectives. Several statistical methods have been developed for analyzing multireader, multicase (MRMC) receiver operating characteristic (ROC) studies. The objective of this article is to increase awareness of these methods and determine if their results are concordant for published datasets.

Materials and Methods. Data from three previously published studies were reanalyzed using five MRMC methods. For each method the 95% confidence intervals (CIs) for the mean of the readers' ROC areas for each diagnostic test, the *P* value for the comparison of the diagnostic tests' mean accuracies, and the 95% CIs for the mean difference in ROC areas of the diagnostic tests were reported.

Results. Important differences in *P* values and CIs were seen when using parametric versus nonparametric estimates of accuracy, and there were the expected differences for random-reader versus fixed-reader models. Controlling for these differences, the Dorfman-Berbaum-Metz (DBM), Obuchowski-Rockette, Beiden-Wagner-Campbell, and Song's multivariate Wilcoxon-Mann-Whitney (WMW) methods gave almost identical results for the fixed-reader model. For the random-reader model, the DBM, Obuchowski-Rockette, and Beiden-Wagner-Campbell methods yielded approximately the same inferences, but the CIs for the Beiden-Wagner-Campbell method tend to be broader. Ishwaran's hierarchical ROC sometimes yielded significance not found with other methods. Song's modification of DBM's jack-knifing algorithm sometimes led to different conclusions than the original DBM algorithm.

Conclusion. In choosing and applying MRMC methods, it is important to recognize: (1) the distinction between random-reader and fixed-reader models, the uncertainties accounted for by each, and thus the level of generalizeability expected from each; (2) assumptions made by the various MRMC methods; and (3) limitations of a five- or six-reader study when the reader variability is great.

Key Words. Receiver operating characteristic (ROC) curve; ROC analysis; multireader study; multireader multicase (MRMC) study; diagnostic accuracy.

Receiver operating characteristic (ROC) curves have now been widely accepted as the standard method for describing the accuracy of a diagnostic test (1). There is a large body of literature on methods for estimating and comparing ROC curves and the various indices that can be derived from them (2–8).

Acad Radiol 2004; 11:980–995

[1] From the Departments of Biostatistics and Epidemiology, and Radiology, Cleveland Clinic Foundation, Cleveland, OH; the Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Washington, DC; the Department of Radiology, University of Iowa, Iowa City, IA; the Program for Interdisciplinary Research in Health Care Organization, Iowa City VA Medical Center, Iowa City, IA; and the Department of Biostatistics, Catholic University Medical College, Seoul, Korea. Received December 10, 2003; revision requested February 25, 2004; revision received April 19; revision accepted April 26. **Address correspondence to** N.A.O., Department of Biostatistics and Epidemiology, Desk Wb4, The Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195. e-mail: nobuchow@bio.ri.ccf.org

doi:10.1016/j.acra.2004.04.014

One important application of ROC analysis is the comparison of diagnostic tests that rely on a trained reader for subjective interpretation, eg, digital mammography compared with film mammography. Because of the known inherent variability in readers' accuracies (9), studies of such diagnostic tests usually involve several trained readers, commonly 4–15. These studies are usually designed as factorial experiments, in that (A) the same patients undergo all of the diagnostic tests under study and (B) the same readers interpret the results from all of the diagnostic tests. This factorial design is efficient in terms of reducing the number of patients and readers needed (10); however, the complicated correlation structure is challenging to analyze appropriately. Furthermore, the goal of these multireader studies is not to report the accuracy of the diagnostic tests for the patients in the study, but rather to report how the tests perform, on average, for similar patients. Likewise, it is often important to report how the tests will perform for readers similar to the ones in the study (11). For the analyst, this means that there are several sources of variability to account for.

In this study we examined five different methods for analyzing multireader ROC studies. The first of these (12) was described in 1992 and the most recent (13) in 2000. The five methods are diverse, including various resampling techniques, modified analysis of variance (ANOVA), and ordinal regression models. We review the main features of each method. We then apply each of the five methods to three previously published datasets and compare the results. Because these are not simulated dataets, we do not know the true population parameter values or distributions of the datasets; our intent, therefore, is not to identify problems with any method, but rather to assess the similarities and differences of their empirical results.

REVIEW OF MULTIREADER ROC METHODS

Multiple-reader, multiple-case (often referred to as MRMC) ROC studies are commonly used in phases II and III* of the assessment of the accuracy of diagnostic tests. The goals of these studies are to estimate and compare accuracies of diagnostic tests. Typically, a multireader ROC study involves a sample of c patients (including some with and some without the disease of interest) who have undergone two or more diagnostic tests (or their images have been displayed in two or more different display modes or analyzed with two or more different computer algorithms, etc). The images generated by the tests (display modes, algorithms, etc) are interpreted by a sample of r readers (often radiologists by training) who are blinded to the true disease status of the patients and to the findings of other competing tests and other readers. Table 1 illustrates the layout of this factorial design, which has been referred to as a "paired-patient, paired-reader" design (10). Other designs are also possible, including "unpaired-patient paired-reader," "paired-patient unpaired-reader," and "unpaired-patient unpaired-reader."

Table 1
Layout for Multi-Reader Factorial Design

	Diagnostic Test 1			Diagnostic Test 2		
	Reader 1	Reader j	Reader r	Reader 1	Reader j	Reader r
Patient 1	X_{111}	X_{1j1}	X_{1r1}	X_{211}	X_{2j1}	X_{2r1}
Patient k	X_{11k}	X_{1jk}	X_{1rk}	X_{21k}	X_{2jk}	X_{2rk}
Patient c	X_{11c}	X_{1jc}	X_{1rc}	X_{21c}	X_{2jc}	X_{2rc}

Where X_{ijk} is the confidence score assigned by reader j to the k-th patient on the i-th diagnostic test.

The factorial, or paired-patient paired-reader, design has several main sources of noise, or variability, in the measurement of diagnostic test accuracy, the most obvious being the variability between patients and the variability among readers. For example, there is a range of patient difficulty (because of the differences in tissue densities or disease characteristics) and a range of reader skill (because of training, experience, and natural aptitude). Moreover, because there are several ways in which the readers' interpretations may be correlated, there will also be several ways in which the estimates of system performance (ie, accuracy) can be correlated: correlation between estimates of accuracies across tests because the same patients undergo each of the tests; correlation between these estimates across tests because the same readers interpret the results from each of the tests; and correlation among estimates of readers' accuracies for the same test because all of the readers are interpreting the same images. Some models incorporate these correlations explicitly, while others incorporate them implicitly through so-called interaction terms. The MRMC methods we examine in this article handle variances, correlations,

*There are usually three phases to the clinical assessment of a diagnostic test's accuracy (8). Phase I is the exploratory phase; its role is to determine if the diagnostic test has *any* ability to discriminate diseased from nondiseased patients. Phase II is the challenge phase, where the accuracy of one or more tests are estimated and compared on difficult cases; the goal is to identify weaknesses of the tests. In phase III, mature tests are applied to well-defined and generalizeable clinical populations to estimate, compare, and report clinically useful measures of accuracy and predictive ability.

Table 2
Key Features of Five Methods

Key Features	DBM	OR	multiWMW	BWC	HROC
Unit for model/analysis	pseudovalue for each pt	summary measure of accuracy	confidence score for each pt	summary measure of accuracy	latent variable for each pt
Reader differences	random or fixed	random or fixed	fixed	random or fixed	random or fixed
Patient or reader covariates in model?	ANOVA with pseudovalues possible but not currently supported	no	no	no	yes
Measure of accuracy	any ROC index	any ROC index	ROC area	any ROC index	any ROC index
Basis for comparing tests	means of accuracies	means of accuracies	means of accuracies	means and variances of accuracies	means, variances, and reader cutpoint
Software available to public?	yes (15,16)	yes (20)	no	no	no

NOTE. pt = patient; DBM = Dorfman-Berbaum-Metz method; OR = Obuchowski-Rockette method; WMW = Wilcoxon-Mann-Whitney method; BWC = Beiden-Wagner Campbell method; HROC = heirarchical ordinal regression for ROC curves; ANOVA = analysis of variance.

and interactions in different ways with different assumptions.

Another major difference between the methods is how they describe, or model, test accuracy. There are several different possible levels at which to model test accuracy: one could specify a model for the confidence scores assigned by the readers to the images; one could define a model for a transformation of the observed confidence scores (eg, pseudovalues); one could directly model the summary measure of accuracy (eg, the ROC area); and one could model both the confidence scores and summary measures of accuracy. The MRMC methods compared in this article use all of these approaches.

Another important difference is how the methods handle variation in reader performance. In phase II studies, the readers are often selected from available readers at the institution where the study is performed. This sample of readers is often not generalizeable to a broad population of radiologists. For these studies, the conclusions of the study should pertain to these particular readers only (so called "fixed effects"). In phase III studies, however, the readers should represent a well-defined population of readers so that the study's estimates of accuracy are generalizeable to patients, as well as readers, at other institutions. Here, the variation in readers' performance is treated as a source of variability (so called "random effects"). Some of the multireader methods described here can be applied to either fixed effects or random effects situations; others are suitable to one of these situations.

Finally, differences in accuracy between two tests can occur in many ways. The difference most commonly analyzed is the difference in the means of the readers' accuracies (eg, ROC areas) for the two tests. However, the distribution of readers' accuracies can have the same location (ie, mean) for two tests but differ in spread (variance). There could also be differences in the way readers use the confidence scale for two tests. If the confidence scale is used to make decisions about patient management, this type of difference can be important, even when the summary measures of accuracy are the same. The methods reviewed in this article all address the first type of difference (ie, difference in the means of the readers' accuracies), and some address these other types of differences.

In Table 2 we compare the five multireader methods on these key features. We now review the methods in chronological order of their appearance in peer-reviewed journals.

ANOVA of Pseudovalues or "Dorfman-Berbaum-Metz Method" (DBM)

Dorfman et al (12,14) proposed an ANOVA on pseudovalues to analyze multireader ROC data. Their basic idea is to compute jack-knife pseudovalues. The jack-knife pseudovalue of the k-th patient is simply the weighted difference in the accuracy, estimated from all patients, minus the accuracy estimated without the k-th patient; these pseudovalues serve as transformations of the original data. A mixed-effects ANOVA is performed on the pseudovalues to test the null hypothesis that the mean accuracy of readers is the same for all of the diagnostic tests studied. Accuracy can be characterized using

Table 3
Statistical Models of Five Methods

Method (Reference)	Statistical Model
DBM (12,14)	$Y_{ijk} = \mu + \alpha_i + B_j + C_k + (\alpha B)_{ij} + (\alpha C)_{ik} + (BC)_{jk} + (\alpha BC)_{ijk} + Z_{ijk}$
OR (18,19)	$\theta_{ijq} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + E_{ijq}$
multivariate WMW (17)	no model specified - fully nonparametric
BWC (21,23,24)*	$\theta_{ijk} = \alpha_i + B_j + C_k + (\alpha B)_{ij} + (\alpha C)_{ik} + (BC)_{jk} + (\alpha BC)_{ijk} + Z_{ijk}$
Hierarchical ordinal regression† (13)	$M_{k,j} = B^T x_{k,j} + (Z_{k,j} + \delta_{k,j}) exp(\lambda^T x_{ikj})$ where $W_{k,j} = r_s$ if $CP_{s-1} < M_{k,j} \leq CP_s$

Left-side: Y_{ijk} is the pseudovalue for the i-th diagnostic test (i = 1, . . . , t), j-th reader (j = 1, . . . , r), and k-th patient (k = 1, . . . , c). θ_{ijq} is the diagnostic accuracy of the i-th test, j-th reader, and q-th reading occasion of the same reader using the same diagnostic test (q = 1, . . . , Q) (in our examples Q = 1). θ_{ijk} is the diagnostic accuracy of the i-th test, j-th reader, and k-th patient. M_{kj} is the latent variable for the k-th patient and j-th reader; W_{kj} is the observed ordinal confidence score; r_s is some ordered scale; and CP_s is an ordered cutpoint.

Right-side: μ is the overall population mean. α_i is the fixed effect of diagnostic test i. B_j is the random effect of reader j which is distributed with zero mean and variance σ_B^2; β_j is the fixed effect of reader j; and B^T is a rx1 vector of reader accuracy location parameters. C_k is the random effect of patient k which is distributed with zero mean and variance σ_C^2. $(\alpha B)_{ij}$ is the random effect of the interaction between diagnostic tests and readers which is distributed with zero mean and variance $\sigma_{\alpha B}^2$. $(\alpha C)_{ik}$ is the random effect of the interaction between diagnostic tests and patients which is distributed with zero mean and variance $\sigma_{\alpha C}^2$. $(BC)_{jk}$ is the random effect of the interaction between readers and patients which is distributed with zero mean and variance σ_{BC}^2. $(\alpha BC)_{ijk}$ is the random effect of the interaction between diagnostic tests, readers, and patients which is distributed with zero mean and variance $\sigma_{\alpha BC}^2$. x_{kj} are r-dimensional indicator vectors which specify the specific reader; they are multiplied by either 1/2 or −1/2 depending upon whether the k-th patient was diseased or not, respectively. $\delta_{k,j}$ are discrete variables that define the link function. λ^T is a rx1 vector of reader accuracy scale parameters. The error terms are as follows: Z_{ijk} is the random error for the i-th test, j-th reader, k-th patient, which is independently distributed with zero mean and variance σ_Z^2; E_{ijq} is the random error for the i-th test, j-th reader, q-th measure of accuracy, which follows a multivariate distribution involving the three correlations, r_1, r_2, and r_3 (see text for definitions), and the variances due to samples of patients, σ_P^2, and the variance due to within-reader variability, σ_W^2; and $Z_{k,j}$ is the random error for the k-th patient and j-th reader.

*BWC have proposed a generalization of this model for when the diagnostic tests have unequal variances (23).

†This is one commonly used form of the HROC model; the model specifies different links for each of the r readers and allows for a location and scale difference across diagnostic tests. The model can also accommodate covariates due to patient and/or reader characteristics.

any summary measure (eg, sensitivity, specificity, the area under the ROC curve, partial area under the ROC curve, sensitivity at a fixed false-positive rate, etc). Furthermore, these measures of accuracy can be estimated parametrically or nonparametrically.

The statistical model is given in the first row of Table 3. We use Dorfman et al's notation throughout the article (12,14), where main fixed effects are denoted with Greek letters and main random effects with capital English letters. The jack-knife pseudovalue, Y_{ijk}, for the *i*-th diagnostic test, *j*-th reader, and *k*-th patient is written as a linear function of an overall population mean, a fixed effect of diagnostic test *i*, a random effect because of the *j*-th reader (note that this reader effect can also be treated as fixed), a random effect because of the *k*-th patient, four interactions of these effects, and a random error. The null hypothesis is that the fixed treatment effects are equal (ie, $\alpha_1 = \alpha_2 = \ldots = \alpha_1$, where *t* is the total number of diagnostic tests studied. Dorfman et al assume that the random effects and error term in the model are normally and independently distributed.

Charles Metz at the University of Chicago and the late Donald Dorfman and his colleagues at the University of Iowa have both provided computer programs that perform the Dorfman et al (12) MRMC analysis. These programs, LABMRMC (15) and MRMC2.0 (16), are both FORTRAN programs that run under Windows (Microsoft; Seattle, WA) and share a number of components. Recently, a method of sample size estimation based on the Dorfman-Berbaum-Metz (DBM) method has been developed and is being tested (see Appendix).

Song (17) suggested and tested several modifications of the ANOVA of pseudovalues method. When generating the pseudovalues, Dorfman et al (12) delete one patient at a time, regardless of whether the patient does or does not have the disease. One of Song's modifications is to generate the pseudovalues by deleting one patient from the sample of patients with disease and one patient from the sample of patients without disease. The method is applicable in studies with equal numbers of patients with and without disease (balanced design). It is also important to point out that the results of Song's method depend on the

randomization procedure used and hence are not unique. This variation of the ANOVA of pseudovalues method will be illustrated in example 2 of this article. Note that Song treats readers as fixed, but the modifications can be used for a random-readers model, as well.

ANOVA with Corrected F-test or "Obuchowski-Rockette Method" (OR)

Instead of modeling the jack-knife pseudovalues, Obuchowski and Rockette (18,19) modeled the accuracy (eg, the ROC curve area) of the *j*-th reader using the *i*-th diagnostic test on the *l*-th reading occasion (see Table 3). As with the ANOVA of pseudovalues method, accuracy can be characterized using any index of accuracy, and either parametric or nonparametric estimates of accuracy can be used.

Obuchowski and Rockette assume that the accuracy indices follow a treatment-by-reader ANOVA model with correlated errors, where the correlation structure is characterized by three correlations: r_1 is the correlation in the error terms of the same reader using different diagnostic tests, r_2 is the correlation in the error terms of different readers using the same diagnostic test, and r_3 is the correlation in the error terms of different readers using different diagnostic tests. These three correlations have been mapped to functions of the interaction terms in the DBM model (10). Specifically,

$$r_1 = \frac{\sigma_C^2 + \sigma_{BC}^2}{\sigma_C^2 + \sigma_{BC}^2 + \sigma_{\alpha C}^2 + \sigma_{\alpha BC}^2},$$

$$r_2 = \frac{\sigma_C^2 + \sigma_{\alpha C}^2}{\sigma_C^2 + \sigma_{BC}^2 + \sigma_{\alpha C}^2 + \sigma_{\alpha BC}^2}, \text{ and}$$

$$r_3 = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{BC}^2 + \sigma_{\alpha C}^2 + \sigma_{\alpha BC}^2}. \quad (1)$$

The correlations are important in determining sample size for a future MRMC study (19) (see Appendix).

Obuchowski and Rockette perform a mixed-effects ANOVA on the accuracy indices (OR method), treating both patients and readers as random effects (note that the analysis can also be performed treating readers as a fixed effect) (18). They modify the usual ANOVA *F*-tests to correct for the correlations between and within readers. The correlations for the modified *F*-tests must be estimated from the data. A FORTRAN program to perform these analyses is available to the public (20).

Multivariate WMW Statistic

Song (17) proposed an extention of the nonparametric approach of DeLong et al (6) to the multireader scenario. DeLong et al used *U*-statistics to compare the ROC areas of several diagnostic tests when only one reader interprets each test; Song extended this to the multivariate situation to handle data from multiple readers. Song developed a Wald statistic for testing whether the accuracies of the diagnostic tests are equal. No model is specified because the approach is fully nonparametric; however, the formulation of the test statistic treats readers as a fixed effect. The method is applicable to accuracy as measured by the area under the ROC curve.

Bootstrap of Components-of-Variance or "BWC Method"

Beiden, Wagner, Campbell (21) (BWC method) approached the MRMC ROC problem from the point of view of the general components-of-variance model previously analyzed by Roe and Metz (22). The BWC method uses the same underlying components-of-variance model as that used by DBM (12) (see Table 3), and thus, the BWC model can also be linked to that of OR through equation 1. Roe and Metz (22) laid out a very general framework, showing how that model was relevant to an entire family of different experiments on a specified population of readers and cases. BWC then showed how a subset of that large family of experiments could be carried out on the population to obtain a particular set of observable variances, and then a system of equations solved to estimate the strengths of the underlying unobservable model variance components. In the real world of a finite data set, corresponding bootstrap experiments replace the population experiments and finite-sample estimates of the variance components are obtained from the same system of equations. The second-order method of the jack-knife-after-bootstrap is then used to estimate the uncertainties in those estimates (23,24). These estimates can be used to size future studies (see Appendix). Confidence intervals on the difference of accuracy estimates across modalities averaged over readers are obtained using the bias-corrected and accelerated bootstrap, a higher-order approach that can be shown to be accurate to second order (25).

Thus, the BWC method is completely nonparametric in structure; the particular accuracy measure that is used (ROC area, for example) can be obtained either parametrically or nonparametrically. Although the BWC approach shares with ANOVA the use of components-of-variance

models, it is otherwise different from ANOVA; the models and computer implementation are distribution-free. (For the present study, normality was assumed only for calculating the *P* values, as well as for the power calculations in the Appendix. In the next generation of the BWC software, *P* values will be derived directly from the bootstrap.)

Hierarchical Ordinal Regression for ROC Curves

Ishwaran and Gatsonis (13) developed Bayesian hierarchical ordinal regression models (HROC models) to deal with multireader ROC data and other types of multilevel clustered data encountered in radiology studies. The HROC models are based on a latent variable construction in which the underlying latent scale for the continuous latent variable is divided into a set of contiguous intervals, formed by unknown cutpoint parameters, with each interval corresponding to a specific value for the ordinal response (the number of cutpoint parameters equals the number of ordinal categories minus one). Ordinal response is determined by which interval the latent variable lies in. For repeated ordinal responses there is a multivariate latent variable, and the ordinal response values are determined by which intervals the multivariate latent variables fall into. For repeated measurements the HROC models assume a conditional multivariate normal latent variable distribution. The parameters making up the mean and variance of this distribution are assumed to have a Bayesian hierarchical prior structure. These parameters are selected to model possible effects from patients, individual readers, different hospitals, or other effects in the data. The HROC models can also be extended to allow for more than one set of cutpoint parameters. For example, it allows for different cutpoints for readers, making it possible to analyze differences in the way readers use the ordinal scale in interpreting a diagnostic scan (26). The hierarchical prior structure allows for a wide range of marginal distributions for the unobserved latent variable; for example, the well known binormal distribution is a special case of the HROC models; extensions to normal location mixture distributions are also possible. The computation of the parameter estimates, their 95% credible intervals (the Bayesian analogue of a 95% confidence interval), and Bayesian *P* values for comparing diagnostic tests, can be complex. Therefore, Gibbs sampling algorithms are used for estimating the model parameters.

Ordinal Regression using Generalized Estimating Equations

One published approach not considered in this article is the ordinal regression approach of Toledano and Gatsonis (27). Their method is appropriate only for ordinal confidence scales, however, and two of the examples considered here used probability estimates collected on a continuous scale.

METHODS

We analyzed three previously published multireader ROC datasets. Permission for use of these datasets was granted by the principal investigators of each study. In each dataset, two or more diagnostic tests were compared using a factorial design involving five or six readers. The readers' confidence scores were reported either on a five-point ordinal rating scale, or a quasi-continuous 0% to 100% confidence score.

For each dataset, using each MRMC method, we reported the following results:

1. 95% confidence interval (CI) for the mean ROC area of each diagnostic test (note that the HROC method gives a 95% credible interval instead);
2. *P* value of the test of the hypothesis that the mean accuracies of the diagnostic tests are equal; and
3. 95% CI or credible interval for the mean difference in ROC areas of the diagnostic tests.

DATASETS

The first example comes from a study comparing the relative performance of spin-echo magnetic resonance imaging (SE-MRI) to cinematic presentation of MRI (CINE-MRI) for the detection of thoracic aortic dissection (28). Forty-five patients with an aortic dissection and 69 patients without a dissection were imaged with both SE-MRI and CINE-MRI. Five radiologists independently interpreted all of the images using a five-point ordinal scale: 1 = definitely no aortic dissection, 2 = probably no aortic dissection, 3 = unsure about aortic dissection, 4 = probably aortic dissection, and 5 = definitely aortic dissection.

In the second example the performance of film-screen mammography is compared with digitized mammography (29). Thirty patients with breast cancer and 30 patients without breast cancer were imaged with both film-screen

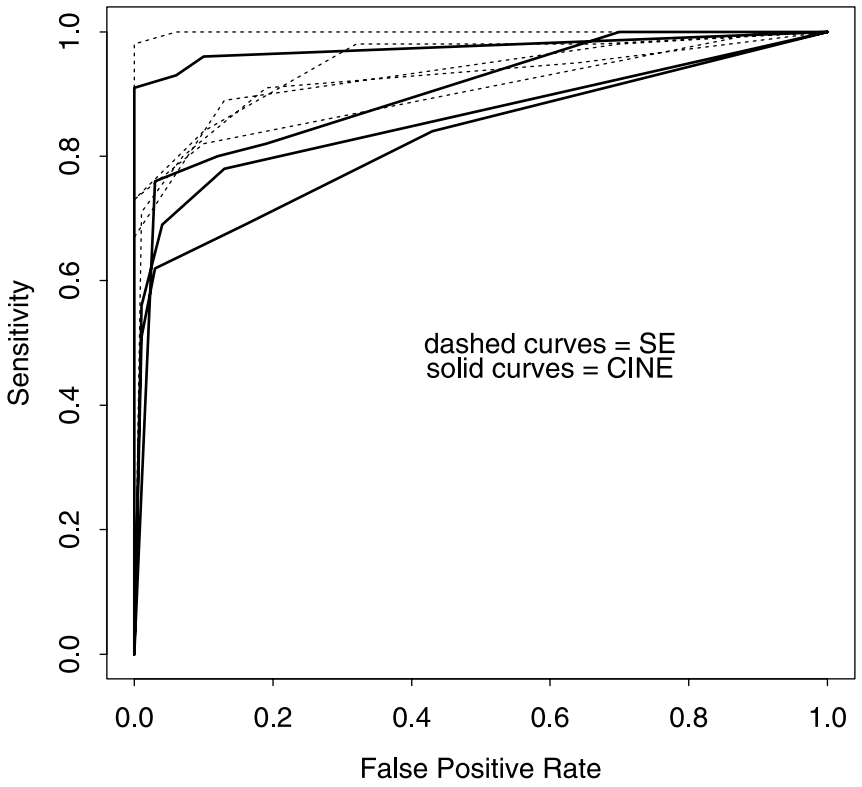

Figure 1. Empirical ROC curves of the five readers (example 1) for SE-MRI (dashed) and CINE (solid) in the detection of aortic dissection. The average areas under the empirical curves are 0.941 (SE) and 0.897 (CINE).

and digitized mammography. Six radiologists independently interpreted all of the images using a quasi-continuous 0% to 100% confidence scale.

The third example comes from a study comparing the accuracy of high, medium, and low resolution chest images for the detection of interstitial disease. The analyzed dataset is a subset from Herron et al (30) and consisted of 40 patients with and 110 without interstitial disease. Five radiologists independently interpreted all of the images using a quasi-continuous 0% to 100% confidence scale.

RESULTS

SE Versus CINE-MRI for Detection of Aortic Dissection

The empirical ROC curves of the five readers are illustrated in Figure 1. The ROC curves for SE-MRI tend to be above those of CINE; the arithmetic averages of the readers' areas under the empirical curves were 0.941 and 0.897, respectively. Several of the empirical ROC curves have long horizontal or vertical steps because the reader did not use all five categories to describe their confidence in the presence of a dissection. For example, reader 4 with SE-MRI used categories 1–3 for patients without a dissection and categories 3–5 for patients with a dissection. CORROC2 software (31) for fitting a bivariate binormal model to ROC rating data produced a degenerate ROC curve for reader 4 with SE-MRI; the OR and BWC methods, which use this algorithm, replaced the parametric estimate from this degenerate curve with a nonparametric estimate. On the other hand, the DBM estimates were not degenerate. In general, DBM and CORROC2 estimates can differ because DBM estimates the ROC curve for each treatment-reader combination separately, assuming a binormal model, while CORROC2 estimates ROC curves for two treatments and one reader or two readers and one treatment simultaneously, under the assumption of a bivariate binormal model.

The analytical results of the five methods for the aortic dissection example are presented in Table 4. For the DBM, OR, and BWC methods, we present re-

Table 4
Empirical Results From Aortic Dissection Example

Method	CI for Test Accuracy		*P* Value	CI for Difference (CINE-SE)
	CINE	SE		
Parametric, random*				
DBM	[0.849, 0.988]	[0.912, 0.990]	.236	[−0.089, 0.024]
OR	[0.830, 0.999]	[0.886, 1.0]	.199	[−0.104, 0.030]
BWC	[0.825, 0.988]	[0.904, 0.985]	.217	[−0.127, 0.001]
Nonparametric, random†				
DBM	[0.825, 0.969]	[0.894, 0.988]	.053	[−0.088, 0.001]
OR	[0.811, 0.983]	[0.873, 1.0]	.102	[−0.101, 0.014]
BWC	[0.824, 0.956]	[0.888, 0.980]	.087	[−0.105, −0.005]
Parametric, fixed‡				
DBM	[0.865, 0.972]	[0.920, 0.982]	.159	[−0.078, 0.013]
OR	[0.870, 0.960]	[0.906, 0.997]	.077	[−0.080, 0.004]
BWC	[0.851, 0.954]	[0.919, 0.978]	.075	[−0.096, −0.007]
HROC	[0.811, 0.874]	[0.889, 0.939]	.001	[−0.112, −0.031]
Nonparametric, fixed§				
DBM	[0.849, 0.945]	[0.908, 0.974]	.021	[−0.081, −0.007]
OR	[0.857, 0.937]	[0.901, 0.981]	.019	[−0.080, −0.007]
Song's WMW	[0.861, 0.933]	[0.906, 0.976]	.019	[−0.079, −0.008]
BWC	[0.848, 0.941]	[0.906, 0.970]	.018	[−0.084, −0.012]

*Parametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
†Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
‡Parametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.
§Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.

sults based on random-reader and fixed-reader models with both parametric and nonparametric estimates of accuracy. Note that the DBM and OR methods gave very similar results: the CIs are similar, although the OR CIs are often wider; a significant difference was detected only for the fixed-reader model with nonparametric estimates. A significant difference was, in fact, found with all of the approaches that used the fixed-reader model with nonparametric estimates (ie, DBM, Song's multivariate WMW methods, OR, and BWC).

Unlike the DBM and OR methods, the BWC model yielded CIs that did not contain zero for both the random-reader model with nonparametric estimates and the fixed-reader model with parametric estimates (see Fig 2). (Note that the *P* values from the BWC method do not always coincide with the CIs because the CIs are based on the nonparametric bootstrap results, while the *P* values are derived by assuming normality.) The HROC method also yielded a significant difference for the fixed-reader model with parametric estimates.

Film Versus Digitized Images for Detecting Breast Cancer

The empirical ROC curves of the six readers are illustrated in Figure 3. The ROC curves of film and digitized-film mammography are similarly located, with arithmetic means of the readers' areas under the empirical ROC curves of 0.754 and 0.747, respectively.

Table 5 summarizes the results of the different methods for this example. The DBM, OR, and BWC methods gave similar results for CIs and *P* values for all four models. The widths of the CIs sometimes differed, but none of these methods always yielded substantially narrower CIs. The multivariate WMW method yielded results similar to these three for the fixed-reader model with nonparametric estimates. Song's jack-knife gave the narrowest CIs for the fixed-reader model with nonparametric estimates, but the *P* value was similar to the other methods. In fact, for all of the methods and models (except the HROC method), the *P* values were in the range of 0.739−0.875. The HROC method gave much narrower CIs than the other methods for the fixed-reader model with parametric estimates (see Fig 4), and its *P* value, .569, was the smallest of all the methods.

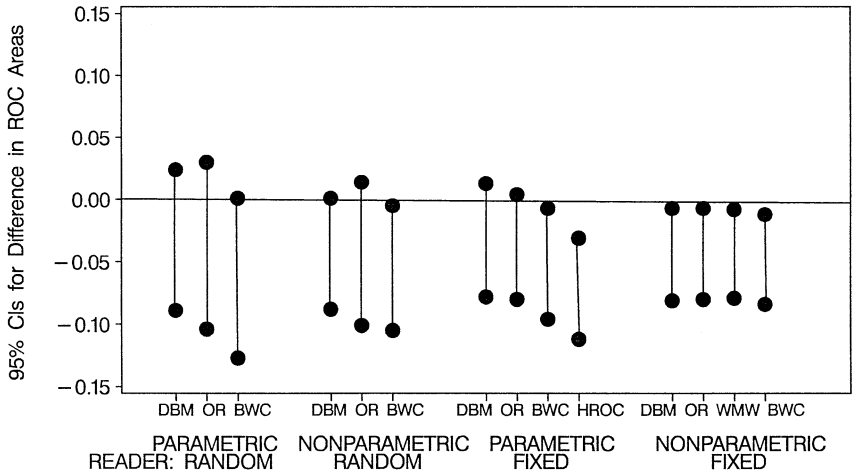

Figure 2. 95% Confidence intervals for the difference in mean ROC areas of SE-MRI and CINE based on the various multireader approaches, treating readers as random- or fixed-effects, and using parametric or nonparametric estimates of the ROC areas.

High, Medium, and Low Resolution Chest Images for Detecting Interstitial Disease

The empirical ROC curves of the five readers are illustrated in Figure 5. There is large inter-reader variability, with much smaller differences between curves at different resolutions. The arithmetic means of the readers' areas under the empirical ROC curves for high, medium, and low resolutions are 0.725, 0.718, and 0.699, respectively.

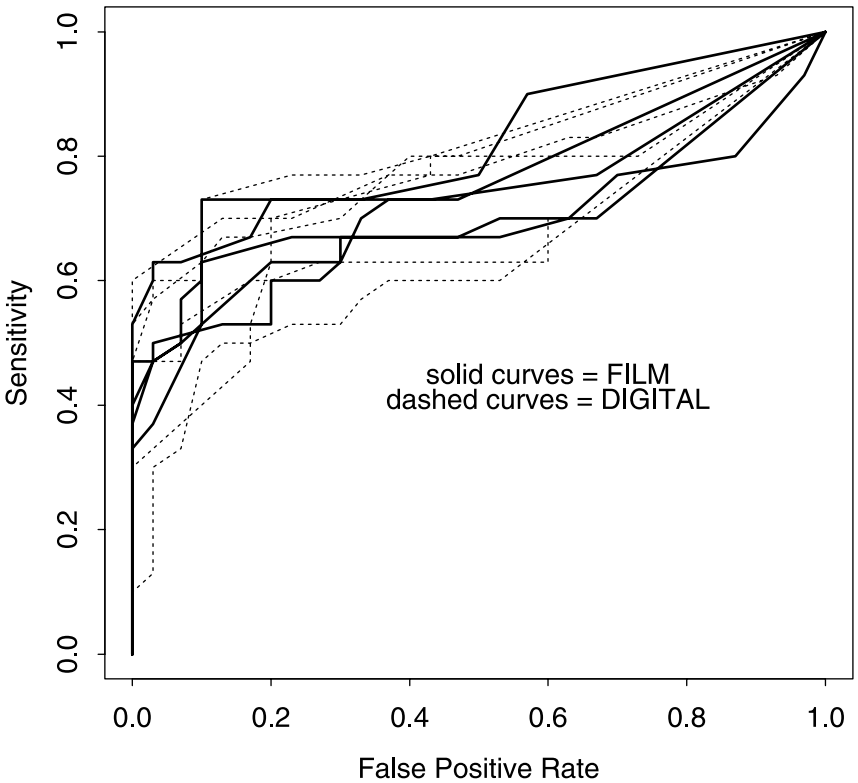

Figure 3. Empirical ROC curves of the six readers (example 2) for film (solid) and digitized-film (dashed) mammography in the detection of breast cancer. The average areas under the empirical curves are 0.754 (film) and 0.747 (digitized-film).

Table 5
Empirical Results From Mammography Example

Method	CI for Test Accuracy		*P* Value	CI for Difference (Film-Digitized)
	Film	Digitized		
Parametric, random*				
DBM	[0.626, 0.832]	[0.627, 0.847]	.745	[−0.058, 0.079]
OR	[0.608, 0.886]	[0.582, 0.889]	.741	[−0.072, 0.095]
BWC	[0.624, 0.849]	[0.601, 0.845]	.818	[−0.093, 0.116]
Nonparametric, random†				
DBM	[0.658, 0.842]	[0.652, 0.848]	.807	[−0.061, 0.076]
OR	[0.638, 0.871]	[0.621, 0.873]	.809	[−0.066, 0.081]
BWC	[0.648, 0.850]	[0.634, 0.848]	.875	[−0.087, 0.096]
Parametric, fixed‡				
DBM	[0.642, 0.849]	[0.625, 0.846]	.790	[−0.066, 0.087]
OR	[0.643, 0.851]	[0.632, 0.839]	.750	[−0.058, 0.081]
BWC	[0.644, 0.837]	[0.631, 0.826]	.739	[−0.060, 0.084]
HROC	[0.691, 0.787]	[0.704, 0.801]	.569	[−0.059, 0.033]
Nonparametric, fixed§				
DBM	[0.662, 0.847]	[0.647, 0.847]	.818	[−0.056, 0.071]
Song's ANOVA	[0.700, 0.808]	[0.706, 0.788]	.781	[−0.028, 0.043]
OR	[0.665, 0.844]	[0.658, 0.836]	.816	[−0.054, 0.069]
Song's WMW	[0.664, 0.844]	[0.658, 0.836]	.816	[−0.054, 0.069]
BWC	[0.664, 0.839]	[0.656, 0.832]	.815	[−0.054, 0.069]

*Parametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
†Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
‡Parametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.
§Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.

Table 6 summarizes the results of the five methods. Again, the DBM, OR, and BWC methods gave similar results, although for the random-reader models, the widths of the BWC CIs for the differences in mean accuracy were considerably greater (see Fig 6). The HROC method yielded a lower *P* value and narrower CIs than the other fixed-reader models with parametric estimates of accuracy. The multivariate WMW method

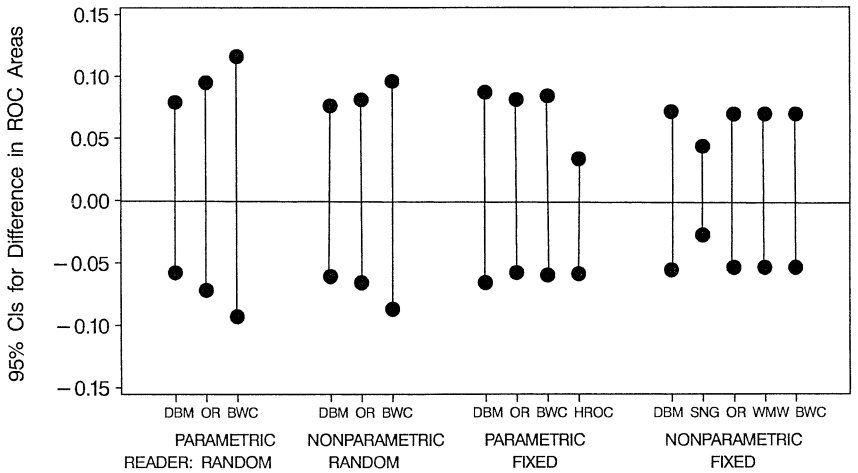

Figure 4. 95% Confidence intervals for the difference in mean ROC areas of film and digitized-film mammography based on the various multireader approaches, treating readers as random- or fixed-effects, and using parametric or nonparametric estimates of the ROC areas.

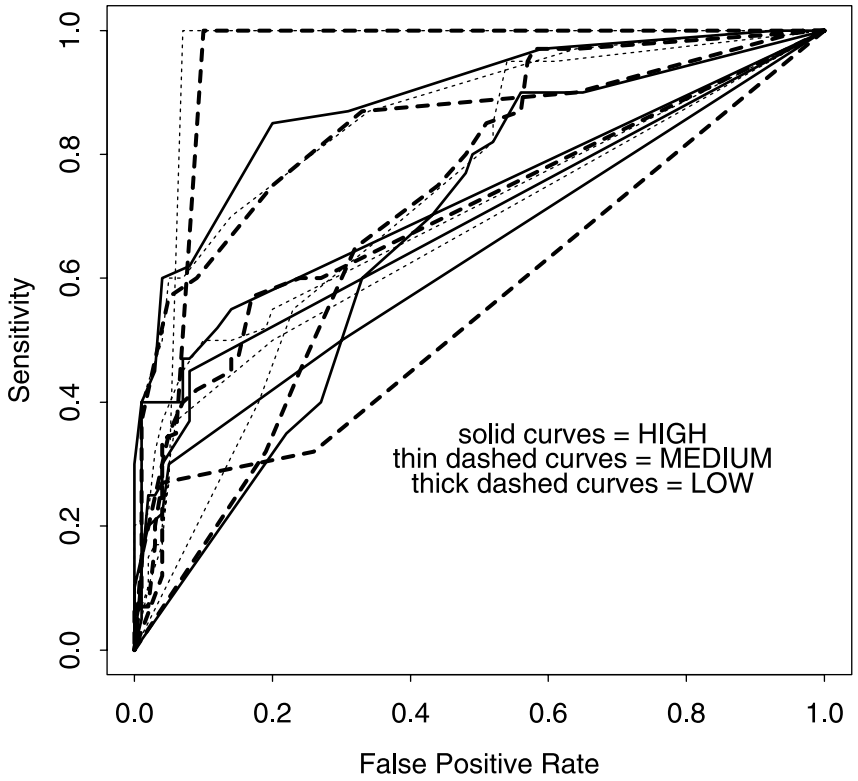

Figure 5. Empirical ROC curves of the five readers (example 3) for high (solid), medium (dotted), and low (dashed) resolution chest images for detecting interstitial disease. The average areas under the empirical curves are 0.725 (high), 0.718 (medium), and 0.699 (low).

agreed well with the other nonparametric fixed-reader models.

DISCUSSION

For analyzing MRMC ROC datasets, investigators currently have available to them six unique approaches. These approaches are based on different models for explaining the differences in accuracy between diagnostic tests, they use different methods of estimation, and make different assumptions about the data. At one end of the spectrum is the fully nonparametric method of Song, which makes no assumptions about the sources of variability. The DBM (12,14) and BWC (21,23,24) methods use simple additive models to describe the variability in the data; the models include main effects and interactions for diagnostic test, patients and readers; the methods differ mainly in how they estimate the parameters of the model. Similarly, the OR method (18,19) uses a simple additive model but focuses on the correlations between and within readers and modalities, and adjusts traditional ANOVA methods to account for these correlations. At the other end of the spectrum is the HROC method (13), which imposes constraints on the data to study specific reader differences, eg, differences in readers' cutpoints, which helps to understand the large variation observed among readers.

In this study, with participation from the authors of each approach, we were able to apply the different approaches to three challenging datasets. The sample sizes of these studies were fairly small (five or six readers, 60 to 150 patients). We expect that some of the differences that we saw between methods are because of finite-sample uncertainties, in addition to systematic effects of the different methods. These sample sizes are, however, what are often seen in studies of diagnostic tests and, therefore, a comparison of the methods on datasets of these sizes is particularly relevant.

Overall, we found that for both parametric and nonparametric ROC area estimates, the fixed-reader models usually give narrower CIs than the random-reader models, although not always (see Fig 6). This is expected because there are fewer sources of variability to esti-

Table 6
Empirical Results From Chest Image Resolution Example

Method	CIs for Test Accuracy			Overall	CIs for Differences Between Resolutions		
	High	Medium	Low	*P* Value	Hi-Med	Hi-Low	Med-Low
Parametric, random*							
DBM	0.627, 0.876	0.663, 0.870	0.626, 0.867	.562	−0.061, 0.031	−0.041, 0.052	−0.026, 0.067
OR	0.627, 0.884	0.645, 0.863	0.609, 0.870	.638	−0.032, 0.035	−0.042, 0.074	−0.040, 0.068
BWC	0.629, 0.851	0.642, 0.846	0.622, 0.828	≥.701	−0.106, 0.111	−0.096, 0.116	−0.110, 0.105
Nonparametric, random†							
DBM	0.614, 0.836	0.620, 0.817	0.581, 0.816	.397	−0.037, 0.051	−0.018, 0.071	−0.024, 0.064
OR	0.599, 0.851	0.604, 0.833	0.566, 0.831	.397	−0.043, 0.056	−0.031, 0.084	−0.032, 0.072
BWC	0.632, 0.821	0.632, 0.808	0.601, 0.793	≥.467,	−0.059, 0.071	−0.041, 0.102	−0.045, 0.083
Parametric, fixed‡							
DBM	0.677, 0.826	0.695, 0.838	0.683, 0.810	.849	−0.088, 0.058	−0.067, 0.078	−0.052, 0.093
OR	0.684, 0.827	0.682, 0.826	0.668, 0.811	≥.616	−0.061, 0.064	−0.046, 0.078	−0.048, 0.076
BWC	0.670, 0.814	0.671, 0.817	0.661, 0.789	≥.576	−0.081, 0.076	−0.063, 0.085	−0.054, 0.087
HROC	0.643, 0.711	0.641, 0.706	0.625, 0.695	≥.309	−0.028, 0.036	−0.015, 0.049	−0.018, 0.045
Nonparametric, fixed§							
DBM	0.669, 0.782	0.665, 0.771	0.648, 0.749	.452	−0.036, 0.050	−0.017, 0.070	−0.023, 0.063
OR	0.672, 0.778	0.666, 0.771	0.646, 0.751	≥.228	−0.036, 0.050	−0.017, 0.070	−0.023, 0.063
Song's WMW	0.669, 0.781	0.666, 0.771	0.649, 0.748	.433	−0.049, 0.062	−0.026, 0.079	−0.030, 0.069
BWC	0.670, 0.781	0.666, 0.770	0.649, 0.748	≥.254	−0.038, 0.051	−0.019, 0.072	−0.019, 0.056

*Parametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
†Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as random effects.
‡Parametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.
§Nonparametric estimates of the ROC area were used in an analysis where the readers were treated as fixed effects.

mate in the fixed-reader model. The choice between a random-reader and fixed-reader model, however, should be based on whether it is appropriate to generalize the results to a broader reader population or to just the readers in the sample. We note that recent submissions to the US Food and Drug Administration involving imaging technologies have all used the random-reader model.

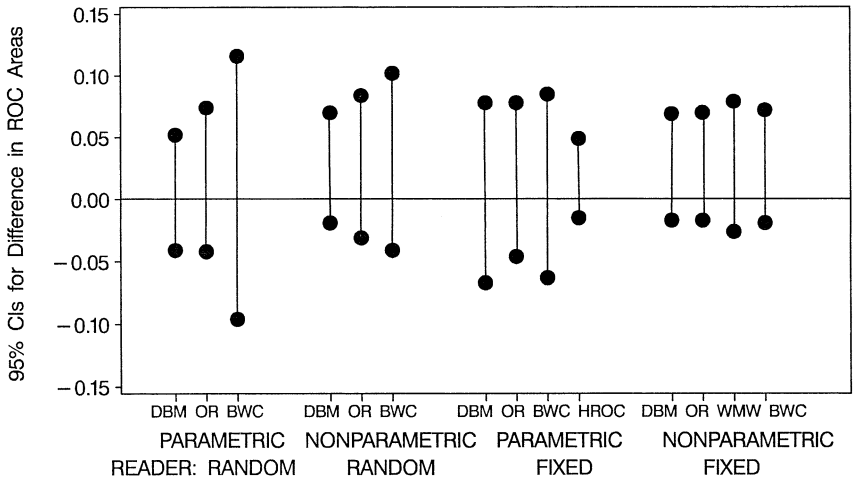

Figure 6. 95% Confidence intervals for the difference in mean ROC areas of high-resolution and low-resolution chest images based on the various multireader approaches, treating readers as random- or fixed-effects, and using parametric or nonparametric estimates of the ROC areas.

We also observed that CIs for the difference in mean accuracy based on nonparametric estimates of accuracy were narrower than CIs based on parametric estimates. This is an unexpected finding but consistent for all three of our examples. We did observe marked deviations from the binormal model used to fit the ROC curves, particularly for examples 2 and 3.

The DBM, OR, and BWC methods often give similar results. This is not surprising because they are based on similar adaptions of the same model. In fact, Hillis (32) recently compared the DBM and OR approaches, concluding that the main difference between the methods is in the degrees of freedom for the test statistic. Song's multivariate WMW yielded results similar to these methods when they used fixed-reader models with nonparametric estimates of accuracy.

Song suggested a simple modification of the DBM jack-knifing procedure. In its current development, however, it can only be applied to studies with an equal sample size for the normal and diseased patients.

The HROC method is a very comprehensive method; it can handle covariates on patients and readers and can detect differences currently not tested by the other methods. In particular, it can detect differences between two tests in how the readers use cutpoints differently. This extension is accomplished by modeling the readers' cutpoints for the latent variable and then comparing the cutpoints between readers, as well as between diagnostic tests (26). For example, in the aortic dissection study (example 1) the HROC method identified fundamental differences in the way readers used the lower ordinal confidence scores (scores 2 and 3); the differences occur between readers in the same modality and between modalities for the same reader. In contrast, the readers used the higher values on the ordinal scale, ie, scores 4 and 5, more consistently, suggesting that the readers are more definitive in calling a case positive (an aortic dissection). We note that the HROC method may also be the most difficult to compute.

The BWC model in Table 2 has been extended (23,24) to a nine-variance component version to account for differences in variance structure across modalities. This extended model can be very useful for quantifying the reduction in the variance when a computer-assist modality is used to aid readers. For the three examples in this study, however, the extended model added little because the variance structures were similar across modalities. Note that the DBM method could also be used to look at differences in variances between modalities by modeling the variance structure of the pseudovalues accordingly in the ANOVA analysis. For example, one could specify the desired error variance structure using the SAS Institute Inc (Cary, NC) procedure PROC MIXED (33).

Another important consideration when choosing which MRMC method to use is their type I error rate and power. There have been a few studies looking at the power and type I error rates of these methods. Roe and Metz (34) and Dorfman et al (14) performed simulation studies of the DBM method to investigate its type I error rate. They simulated patient sample sizes of 50 to 400, reader sample sizes of 3, 5, and 10, different magnitudes for the variances, different distributions (ie, gaussian and non-gaussian), and different ROC areas. They found that the DBM method performs at the correct significance level, or conservatively for small sample size or large ROC areas; it is rarely liberal. Obuchowski and Rockette (18) performed a simulation study to evaluate the type I error rate and power of their method. They considered patient sample sizes of 50 to 200, reader sample sizes of 4 to 12, different covariance structures and variances, and rating and continuous data. They found that their type I error rate is at the correct level for eight or more readers and conservative for fewer readers; it too is rarely liberal. The power of their test drops dramatically with $\leq$ four readers. Beiden et al (35) found, for the case of 10 readers, that the BWC approach gave unbiased estimates even to second order (ie, the estimates of the variance of the variance components were themselves almost unbiased). The results were similar for the case of five readers.

The DBM, OR, and BWC methods can be used to size future studies. In the Appendix we discuss these three methods of sample size determination and compare their results using data from the first example.

Finally, we draw attention to the size of the reader samples in these three examples and their implications. Similar to many other MRMC studies in the literature, our example datasets used five or six readers. It has been the almost universal experience of the present authors that reader variability and its impact have been underestimated by investigators in sizing MRMC studies. For example, in our third dataset, for the high versus low resolution comparison, BWC estimate the uncertainty in the modality-by-reader component to be an order of magnitude greater than the uncertainty in the modality-by-case component. This point cannot be overlooked when considering the obvious differences among the methods for that example. (Note that the modality-by-reader term does not contribute when the readers are considered fixed effects. Not

surprisingly, the methods all produce similar results for this example when the readers are considered fixed.) We hope to address these issues in greater detail in future work.

We suggest caution regarding three aspects of the task of comparing competing modalities in MRMC ROC studies. First, as discussed earlier, the choice between using a model with readers random versus readers fixed is a fundamental one–independent of which analytical tools and software will then apply. Next, some of the approaches depend on assumptions regarding the underlying distributions of the random variables. It is not yet well-known how robust the methods are to departures from these assumptions. Finally, we have remarked on the possible limitations of a five- or six-reader study when the reader variability is great. It is possible that the finite-sample uncertainties resulting from that small number may dominate other effects observed here.

ACKNOWLEDGMENT

The authors greatly appreciate the contributions of Carolyn Van Dyke, MD, Kimerly Powell, PhD, and David Gur, ScD. Without their willingness to share their datasets, this study would not have been possible. RFW acknowledges very helpful conversations with Gregory Campbell, PhD and Brandon D. Gallas, PhD, both of CDRH/US Food and Drug Administration. We also thank two reviewers for their helpful critique of an earlier draft.

REFERENCES

1. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
2. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory–a direct solution. Psychometrika 1968; 33:117–124.
3. Dorfman DD, Alf E. Maximum–likelihood estimation of parameters of signal detection theory and determination of confidence intervals–rating method data. J Math Psychology 1969; 6:487–496.
4. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
5. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148:839–843.
6. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–844.
7. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989; 9:190–195.
8. Zhou HH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley & Sons, 2002.
9. Beam CA, Baker ME, Paine SS, Sostman HD, Sullivan DC. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology research. Radiology 1992; 183:619–620.
10. Obuchowski NA. Multi-reader ROC studies: a comparison of study designs. Acad Radiol 1995; 2:709–716.
11. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging 1989; 29:307–335.
12. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
13. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. Can J Stat 2000; 28:731–750.
14. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol 1998; 5:591–602.
15. LABMRMC. Available at: http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index.htm.
16. Schartz KM, Hillis SL, Berbaum KS, Dorfman DD: MRMC2.0. Available at: http://perception.radiology.uiowa.edu.
17. Song HH. Analysis of correlated ROC areas in diagnostic testing. Biometrics 1997; 53:370–382.
18. Obuchowski NA, Rockette HE. Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: an ANOVA approach with dependent observations. Commun Stat Simul Computation 1995; 24:285–308.
19. Obuchowski NA. Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations with rejoinder. Acad Radiol 1995; 2:S22–S29.
20. OBUMRM. Available at: http://www.bio.ri.ccf.org/OBUMRM/OBUMRM.html.
21. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad Radiol 2000; 7:341–349.
22. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad Radiol 1997; 4:587–600.
23. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structure across modalities. Acad Radiol 2001; 8:605–615.
24. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Acad Radiol 2001; 8:616–622.
25. Efron B, Tibshirani RJ. An introduction to the bootstrap. Monographs on statistics and applied probability 57. New York, NY: Chapman & Hall, 1993.
26. Ishwaran H. Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. Can J Stat 2000; 28:715–730.
27. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. Stat Med 1996; 15:1807–1826.
28. Van Dyke CW, White RD, Obuchowski NA, Geisinger MA, Lorig RJ, Meziane MA. Cine MRI in the diagnosis of thoracic aortic dissection. Chicago, IL: 79th RSNA Meetings, November 28–December 3, 1993.
29. Powell KA, Obuchowski NA, Chilcote WA, Barry MM, Ganobcik SN, Cardenosa G. Clinical evaluation of digital versus film-screen mammograms: diagnostic accuracy and patient management. AJR Am J Roentgenol 1990; 173:889–894.
30. Herron JM, Bender TM, Campbell WL, Sumkin JH, Rockette HE, Gur D. Effects of luminance and resolution on observer performance with chest radiographs. Radiology 2000; 215:169–174.
31. CORROC2. a FORTRAN program written by C. Metz. Replaced by ROCKIT. Available at: http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index.htm.
32. Hillis SL. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Tampa, FL, International Biometric Society ENAR Spring Meeting, March 30–April 2, 2003.
33. PROC MIXED. SAS Institute Inc, Cary, NC. Available at http://www.sas.com
34. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation wtih computer simulation. Acad Radiol 1997; 4:298–303.

35. Beiden SV, Maloof M, Wagner RF. A general model for finite-sample effects in training and testing of competing classifiers. IEEE Trans Patt Anal Machine Intell 2003; 25:1561–1569.
36. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000; 175:603–608.
37. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Acad Radiol (in press).

APPENDIX

One method for sample size calculation for determining both the number of patients and readers needed has been published (19). It is based on the OR model and provides a simple method of estimating sample sizes when there are no pilot data available. Sample size estimation has been developed recently, however, based on two other methods, DBM and BWC. We briefly described each of these three methods of sample size calculation, then compare the methods using data from the aortic dissection example.

The OR method of sample size calculation is derived from OR's corrected F-statistic (19). For comparing two modalities, the power of the test is given by

$$\text{power} = \text{Prob}(F_{1,df,\lambda} > F_{(1-\alpha),1,df}), \tag{A1}$$

where λ is the noncentrality parameter and df is the denominator degrees of freedom. In the OR method, df equals the number of readers minus one, ie, (r-1), and the noncentrality parameter is (19)

$$\lambda = \frac{(A_1 - A_2)^2}{\frac{2}{r}\left[\sigma^2_{\alpha B} + \sigma^2_W/Q + \sigma^2_P[(1-r_1) + (r-1)(r_2 - r_3)]\right]} \tag{A2}$$

where A_i is the expected diagnostic accuracy of modality i under the alternative hypothesis, Q is the number of reading occasions by the same reader using the same test for the same sample of subjects (here, Q = 1), and the variance components and correlations are defined in the text and in Table 3. Some helpful suggestions for estimating these parameters when there are no pilot data are given in (36). When pilot data are available, λ can be estimated (SL Hillis, personal communication, September 2003):

$$\hat{\lambda} = \frac{(A_1 - A_2)^2}{\frac{2}{r}\left[\hat{\sigma}^2_{\alpha B} + \left(\frac{c'}{c}\right)\widehat{(\sigma^2_P + \sigma^2_W)} - \left(\frac{c'}{c}\right)\widehat{covr1} + \left(\frac{c'}{c}\right)(r-1)(\widehat{covr2} - \widehat{covr3})\right]}, \tag{A3}$$

where $\hat{\sigma}^2_{\alpha B} = \text{MS}(\alpha \times \text{B}) - [\widehat{(\sigma^2_P + \sigma^2_W)} - \widehat{covr1} - \widehat{covr2} + \widehat{covr3}]$, and MS($\alpha \times$ B) is the mean square error of the interaction term for readers and modalities, calculated from a two-way ANOVA. $\widehat{(\sigma^2_P + \sigma^2_W)}$ is an estimate of the variance associated with patient samples and the within-reader variability; it can be calculated by taking the average of the variance of the r $\times$ t ROC area estimates. $\widehat{covr1}$, $\widehat{covr2}$, and $\widehat{covr3}$ are the estimated covariances in the error terms of the same reader using different diagnostic tests, of different readers using the same diagnostic test, and of different readers using different diagnostic tests, respectively. These estimates are printed out by OBUMRM (19). c' is the total number of patients in the pilot study, and c and r are the number of total patients and readers, respectively, needed for the pivotal study.

Sample size estimation based on the DBM method (37) is similar to that of the OR method. The power based on the DBM method is given by A1; the degrees of freedom and the noncentrality parameter are:

$$df = \frac{[c\sigma^2_{\alpha B} + r\sigma^2_{\alpha C} + \sigma^2]^2}{\frac{(c\sigma^2_{\alpha B} + \sigma^2)^2}{r-1} + \frac{(r\sigma^2_{\alpha C} + \sigma^2)^2}{c-1} + \frac{(\sigma^2)^2}{(r-1)(c-1)}} \tag{A5}$$

and

$$\lambda = \frac{(A_1 - A_2)^2}{\frac{2}{rc}[c\sigma^2_{\alpha B} + r\sigma^2_{\alpha C} + \sigma^2]}, \tag{A6}$$

where the variance components are defined in Table 3 and estimated from the mean squares from pilot data:

$$\hat{\sigma}^2 = \overline{MS}(\alpha \times B \times C),$$

$$\hat{\sigma}^2_{\alpha B} = \frac{\overline{MS}(\alpha \times B) - \overline{MS}(\alpha \times B \times C)}{c'}, \text{ and}$$

$$\hat{\sigma}^2_{\alpha C} = \frac{\overline{MS}(\alpha \times C) - \overline{MS}(\alpha \times B \times C)}{r'},$$

where c' and r' are the numbers of patients and readers in the pilot study.

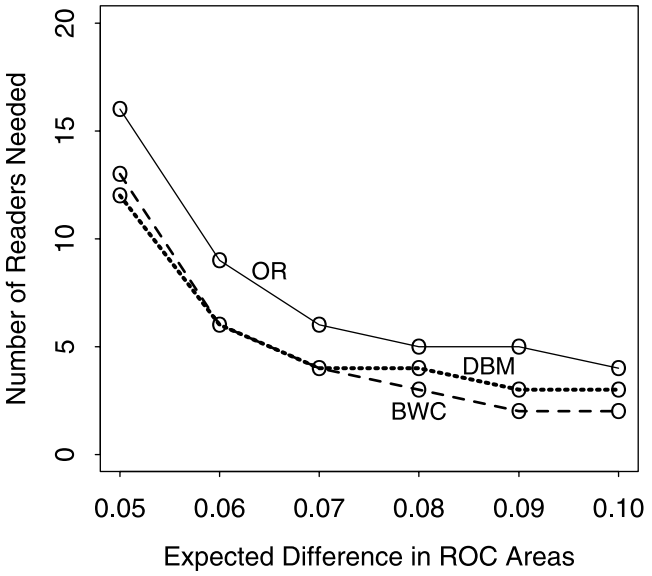

Figure 7. Estimated number of readers needed for aortic dissection example with original patient sample size as a function of the difference in mean reader accuracy. The three sample size methods correspond to the DBM, OR, and BWC methods for MRMC analysis.

For the BWC method, the sample sizes are estimated from:

$$\hat{Var} = \frac{(A_1 - A_2)^2}{[z_{(1-\alpha)} + z_{(1-\beta)}]^2}, \qquad \text{(A9)}$$

where $\hat{Var}$ is the target variance, z_x is the x-percentile of a standard normal distribution, α and (1-β) are the type I error rate and power, respectively. Estimates of the needed number of readers, r, and patients, c, are obtained by requiring that:

$$\hat{Var} = 2[\hat{\sigma}^2_{\alpha C}(c'/c) + \hat{\sigma}^2_{\alpha B}/r + \hat{\sigma}^2_{\alpha BC}(c'/c)/r]. \qquad \text{(A10)}$$

BWC estimate the variance components from the system of six bootstrap experiments described in their articles (20,22,23) and the pilot study data involving c' patients and r' readers.

We now compare the three sample size methods. We used the first example as a pilot study for planning a future study of the accuracy of CINE and SE MRI for detecting aortic dissection. In the analysis of the pilot data we used nonparametric estimates for the ROC areas and treated the readers as random effects. Keeping the number of patients the same as in the original study (ie, 45 patients with a dissection and 69 without), we computed the number of readers needed for a difference in ROC areas between CINE and SE of 0.05 to 0.10 by 0.01. Similarly, keeping the number of readers the same as in the original study (ie, five readers) and the ratio of patients with and without a dissection the same (ie, 45/69, or 0.652), we computed the total number of patients needed for a difference in ROC areas between CINE and SE of 0.05 to 0.10 by 0.01. For each method we determined the minimum number of readers, or patients, needed to achieve at least 80% power with a 5% type I error rate.

The results of these comparisons are illustrated in Figures 7 and 8. The DBM and BWC methods give very similar estimates; the OR method is more conservative, requiring about 70% more patients than the DBM and BWC methods and 1–4 more readers. The differences are due largely to the degrees of freedom used in the OR method (32).

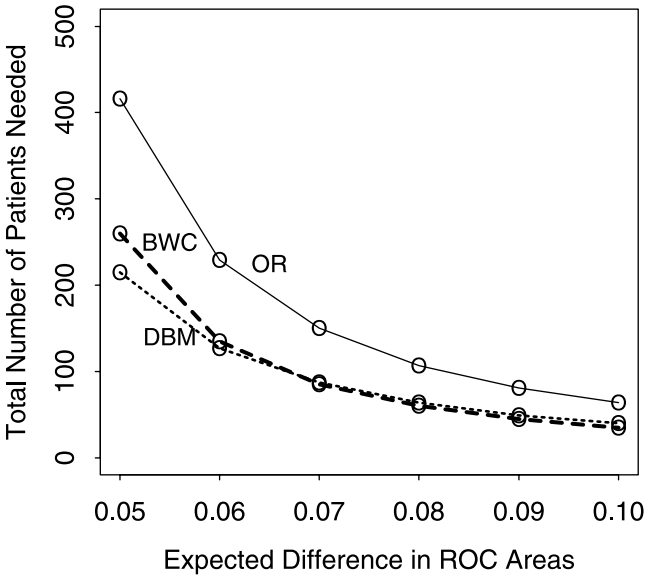

Figure 8. Estimated number of patients needed for aortic dissection example with original reader sample size as a function of the difference in mean reader accuracy.