STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:2919–2933 (DOI: 10.1002/sim.1518)

Three methods for analysing correlated ROC curves: a comparison in real data sets from multi-reader, multi-case studies with a factorial design

Alicia Y. Toledano*,†

Center for Statistical Sciences, Department of Community Health, Brown University, Providence, Rhode Island, U.S.A.

SUMMARY

This paper compares three published methods for analysing multiple correlated ROC curves: a method using generalized estimating equations with marginal non-proportional ordinal regression models; a method using jackknifed pseudovalues of summary statistics; a method using a corrected F-test from analysis of variance of summary statistics. Use of these methods is illustrated through six real data examples from studies with the common factorial design, that is, multiple readers interpreting images obtained with each test modality on each study subject. The issue of the difference between typical summary statistics and summary statistics from typical ROC curves is explored. The examples also address similarities and differences among the analytical methods. In particular, while point estimates of differences between test modalities are similar, the standard errors of these differences do not agree for all three methods. A simulation study supports the standard errors provided by the generalized estimating equations with marginal non-proportional ordinal regression models. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: ROC curve; generalized estimating equations; ordinal categorical data; summary statistic

1. INTRODUCTION

A receiver operating characteristic (ROC) curve is a graph of the sensitivity versus 1 - specificity of a diagnostic test as the criterion for a positive test result varies through all possible values of the test result. ROC curves measure the potential of a test to discriminate between diseased and non-diseased subjects, independent of the criterion for a positive test result. They have been used to evaluate the accuracy of diagnostic imaging modalities for several years, and are being used increasingly in other medical fields. Studies of diagnostic tests are

*Correspondence to: Alicia Y. Toledano, Center for Statistical Sciences, Brown University, 167 Angell Street 2nd Floor, Box G-H, Providence, Rhode Island, 02912, U.S.A.
† E-mail: toledano@stat.brown.edu

Contract/grant sponsor: National Institutes of Health/National Cancer Institute; contract/grant number: CA76310.

Received March 2000
Accepted January 2003

frequently designed to increase power by performing multiple tests on each subject, thus leading to correlated ROC curves. Among the most common study designs is a factorial design in which each subject receives each diagnostic test, and each diagnostic test is interpreted independently by each member of a group of participating readers. In many applications, smooth ROC curves are estimated from ordinal categorical data using a binormal model [1]. ROC curves are frequently compared through summary statistics, the most common of which is the area under the curve, although other measures may also be used [2, 3].

Three methods for analysing multiple correlated ROC curves have recently been described. Toledano and Gatsonis present a method that uses generalized estimating equations (GEEs) with marginal ordinal regression (OR) models with non-linear predictors [4, 5]. Dorfman *et al.* [6] discuss a method in which one makes inferences about a chosen summary statistic through a generalized linear mixed model with parameter estimation based on jackknifed pseudovalues of the summary statistic. Obuchowski and Rockette [7] propose a method for adjusting the F-test obtained in an analysis of variance (ANOVA) of summary statistics for each of the correlated ROC curves. Obuchowski also discusses sample size estimation for this method [8].

The current paper is motivated by noting similarities and differences in results obtained from the three analytical methods in several real data sets. In order to use available software, these data sets arise from studies with factorial designs and ordinal test results, and analyses focus on the area under the binormal ROC curve. The models are presented in Section 2. In Section 3 we use several real data examples to illustrate the different capabilities of the methods and to gain insight into their relative performance. A simulation study supporting the results obtained by the GEE/OR method is presented in Section 4. The paper concludes with a discussion of these results.

2. THE METHODS

Consider data to arise from a study in which M tests (indexed by $m=1,\ldots,M$) are obtained from each of i subjects ($i=1,\ldots,n$). For example, in radiology the M tests represent imaging modalities. This paper considers tests with ordinal categorical results in C categories, indexed by the subscript c. For the current paper, we assume that the number of possible categories is the same for all of the tests. Each test is scored by R readers, indexed by the subscript r. Thus, there are $M\times R$ test results for each subject. Subjects are considered to be a random sample from the population of subjects. Readers may be considered to be a random sample from the population of readers [9], or may be considered as fixed choices. Test modalities are considered to be fixed choices. The observed test result for subject i and modality–reader combination mr is $Y^{\text{o}}_{i,mr}=c$ for $c=1,\ldots,C$. Finally, the indicator of true disease status is denoted x_i, taking the value 1 if abnormal (diseased) or 0 if normal (non-diseased).

The methods of Dorfman *et al.* [6] and Obuchowski and Rockette [7] analyse summary statistics. The most often used is the maximum likelihood estimate (MLE) of the area under the ROC curve for each modality–reader combination, denoted $\hat{A}_{mr}$. To facilitate comparison with these methods, we focus analyses using the method of Toledano and Gatsonis [4, 5] on $\hat{A}_{m.}=(\sum_{r=1}^{R}\hat{A}_{mr})/R$. The estimates $\hat{A}_{mr}$ are obtained as functions of the parameters estimated using the GEEs [10]. The variance of $\hat{A}_{mr}$ and the covariance of $\hat{A}_{mr},\hat{A}_{m'r'}$ are obtained by Taylor series expansion.

2.1. Generalized estimating equations with ordinal regression (GEE/OR)

Let $\mu_{i,mr,c} = \Pr(Y^{\mathrm{o}}_{i,mr} \leqslant c)$, and apply a non-proportional ordinal regression model [11] with a probit link:

$$\mu_{i,mr,c} = \Phi\left(\frac{\theta_{mr,c} - \alpha_{mr}x_i}{\exp(\beta_{mr}x_i)}\right) \tag{1}$$

Let $B_{mr} = [\theta_{mr,1}, \ldots, \theta_{mr,C-1}, \alpha_{mr}, \beta_{mr}]^{\mathrm{T}}$, where ${}^{\mathrm{T}}$ indicates the transpose, and $B = [B_1^{\mathrm{T}}, \ldots, B_{MR}^{\mathrm{T}}]^{\mathrm{T}}$. Consistent and asymptotically normal estimates of B are obtained as the solution to the GEEs

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} (Y_i - \mu_i) = 0 \tag{2}$$

where $Y_i = [Y_{i1}^{\mathrm{T}}, \ldots, Y_{i,MR}^{\mathrm{T}}]^{\mathrm{T}}$ for $Y_{i,mr} = [Y_{i,mr,1}, \ldots, Y_{i,mr,C-1}]^{\mathrm{T}}$ and $Y_{i,mr,c} = 1$ if $Y^{\mathrm{o}}_{i,mr} \leqslant c$ and 0 otherwise, $\mu_i = [\mu_{i1}^{\mathrm{T}}, \ldots, \mu_{i,MR}^{\mathrm{T}}]^{\mathrm{T}}$ for $\mu_{i,mr} = [\mu_{i,mr,1}, \ldots, \mu_{i,mr,C-1}]^{\mathrm{T}}$, $V_i = \mathrm{var}(Y_i)$ and $D_i = \partial\mu_i/\partial B$.

Pairwise correlations between test results can be modelled as polychoric correlations. Regarding the structure of the correlation matrix, the correlation between two test results $Y^{\mathrm{o}}_{i,mr}$ and $Y^{\mathrm{o}}_{i,(mr)'}$ can be modelled as belonging to one of three types: correlation due to modality and subject, ρ_m; due to reader and subject, ρ_r; and due only to subject, ρ_s. This correlation structure is denoted 3PC. Correlation estimates are obtained using a second set of GEEs. Alternatively, one can use a working assumption of independence among the $M \times R$ test results obtained for each subject.

Under either working independence or 3PC, a consistent estimate of the asymptotic variance of $\hat{B}$ is given by $I_B^{-1}\{\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^{\mathrm{T}} V_i^{-1} D_i\} I_B^{-1}$ evaluated at $\hat{B}$ and, when modelling correlations, estimates of the correlation parameters [12], where the model-based variance estimator is $I_B^{-1} = (\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} D_i)^{-1}$. The two variance estimates approach each other as the correlation model approaches the observed correlation.

Using the GEE method allows choices beyond the average area within each modality, $\hat{A}_{m.}$. One might instead be interested in the area under a typical, or average, ROC curve itself. These will be different, as the area under the curve is a highly non-linear function of the parameters used to estimate it. First, the average ROC curve for test modality m can be defined by $\hat{a}_{m.} = (\sum_{r=1}^{R} \hat{a}_{mr})/R$ and $\hat{b}_{m.} = (\sum_{r=1}^{R} \hat{b}_{mr})/R$. To obtain the area under it, we take $\check{A}_{m.} = \Phi(\hat{a}_{m.}/\sqrt{(1 + \hat{b}_{m.}^2)})$. Alternatively, we can define $\tilde{a}_{m.} = \hat{\alpha}_{m.} \exp(-\hat{\beta}_{m.})$ and $\tilde{b}_{m.} = \exp(-\hat{\beta}_{m.})$ for $\hat{\alpha}_{m.} = (\sum_{r=1}^{R} \hat{\alpha}_{mr})/R$ and $\hat{\beta}_{m.} = (\sum_{r=1}^{R} \hat{\beta}_{mr})/R$. The area under this curve is $\tilde{A}_{m.} = \Phi(\tilde{a}_{m.}/\sqrt{(1 + \tilde{b}_{m.}^2)})$. Differences among $\hat{A}_{m.}$, $\check{A}_{m.}$ and $\tilde{A}_{m.}$ are explored in Section 3. For comparison with the other methods, we estimate $A_{1.} - A_{2.}$ by $\hat{A}_{1.} - \hat{A}_{2.}$. Equality of areas across modalities is tested through an F-test when $M > 2$, or a z-test when $M = 2$.

2.2. Jackknifed pseudovalues (JP)

Dorfman *et al.* [6] define the pseudovalue for each case as

$$\hat{A}_{mr*i} = n\hat{A}_{mr} - (n-1)\hat{A}_{mr(i)} \tag{3}$$

where $\hat{A}_{mr(i)}$ is the MLE of the area under the ROC curve for the combination of modality m and reader r with the ith subject deleted.

The pseudovalues are analysed using a generalized linear model, usually mixed-model ANOVA:

$$A_{mr*i} = \mu + \alpha_m + B_r + C_i + (\alpha B)_{mr} + (\alpha C)_{mi} + (BC)_{ri} + (\alpha BC)_{mri} + e_{mri} \tag{4}$$

where μ and α_m are fixed effects and the other effects are random. B_r is the effect for reader r, and C_i is the effect for subject i. Random effects are assumed to be mutually independent and normally distributed with mean zero and variances σ_B^2, σ_C^2, σ_{aB}^2, σ_{aC}^2, σ_{BC}^2, σ_{aBC}^2 and σ_e^2. In practice, each reader typically interprets each modality for each subject only once, such that $(\alpha BC)_{mri}$ and e_{mri} are not identifiable.

We test the null hypothesis that $\alpha_1 = \cdots = \alpha_M$ by testing whether the sum of squares for modality is zero. One can also test reader effects and/or subject effects, and estimate confidence intervals for modality means and differences between them.

2.3. *Adjusted F-test* (F^*)

A third method described by Obuchowski and Rockette [7] and Obuchowski [8] uses ANOVA of $\hat{A}_{mr}$s, modelling the correlations among them. The $\hat{A}_{mr}$s are most often obtained from a bivariate binormal model [13]. The correlation for summary measures from the same reader, different modalities is ρ_r^*; ρ_m^* is the correlation for different readers, same modality; and ρ_s^* is the correlation for different readers, different modalities.

One analyses the data using a mixed-model ANOVA:

$$A_{mr} = \mu + \alpha_m + B_r + (\alpha B)_{mr} + e_{mr} \tag{5}$$

where μ and α_m are fixed effects. The other effects are random, mutually independent, with mean zero and variances $\sigma_B^2, \sigma_{aB}^2$ and Σ. The diagonal elements of Σ are $\sigma_w^2 + \sigma_e^2$, where σ_w^2 is the within-reader variance. The off-diagonal elements of Σ are $E(e_{mr}e_{m'r'}) = \sigma_e^2\rho_s^*$, $E(e_{mr}e_{mr'}) = \sigma_e^2\rho_m^*$ and $E(e_{mr}e_{m'r}) = \sigma_e^2\rho_r^*$. Estimates of elements of Σ are obtained by averaging appropriate covariance estimates from all possible bivariate binormal models. Again, in typical studies where each reader interprets each modality for each subject only once, σ_{aB}^2 and σ_w^2 will not be identifiable.

In such situations, the null hypothesis $\alpha_1 = \cdots = \alpha_M$ is tested through the statistic

$$F^* = \frac{R\sum_m(\hat{A}_{m.} - \hat{A}_{..})^2/(M-1)}{(M-1)^{-1}(R-1)^{-1}\Sigma_m\Sigma_r(\hat{A}_{mr} - \hat{A}_{m.} - \hat{A}_{.r} + \hat{A}_{..})^2 + R\hat{\sigma}_e^2(\hat{\rho}_m^* - \hat{\rho}_s^*)} \tag{6}$$

where $\hat{A}_{.r} = (\sum_{m=1}^M \hat{A}_{mr})/M$ and $\hat{A}_{..} = (\sum_{m=1}^M \sum_{r=1}^R \hat{A}_{mr})/(MR)$.

The value of F^* is referenced to an F-distribution with $M-1, (M-1)\times(R-1)$ degrees of freedom. The second term in the denominator adjusts for correlations among the areas under the ROC curves. It is set to zero if $\hat{\rho}_m^* \leqslant \hat{\rho}_s^*$. The estimate of var$(\hat{A}_{1.} - \hat{A}_{2.})$ involves a similar correction factor. Confidence intervals for $A_{1.} - A_{2.}$ are computed through a t-distribution with $(R-1)$ degrees of freedom.

2.4. *Comparison of model assumptions and sources of variability*

The robust variance of parameter estimates obtained from the GEE/OR method accounts for correlations among the responses for each subject, using either working correlation described above. Using the 3PC working correlation imposes additional structure. All sources of within-subject variation are included in the variance of parameter estimates. If readers are random, subjects are no longer independent of each other; the GEEs then ignore this source of correlation. The estimate $(\hat{A}_{1.} - \hat{A}_{2.})$ is obtained by transforming the GEE/OR parameter estimates, and the variance of $(\hat{A}_{1.} - \hat{A}_{2.})$ is estimated by Taylor series expansion. If we consider readers to be random, that variance will then also ignore correlations between subjects due to common readers.

The JP method treats the $\hat{A}_{mr*i}$s as independent and identically distributed. In contrast to jackknife estimation of a sample mean, jackknifed pseudovalues are generally correlated [14]. This is true of the $\hat{A}_{mr*i}$s. Ignoring this correlation also affects estimation of $\text{var}(\hat{A}_{1.} - \hat{A}_{2.})$.

The F^* method is based on the work of Thompson and Zucchini [15]. The F^* method shrinks σ_e^2 by ρ_m^*, ρ_r^* or ρ_s^* in $\text{cov}(A_{mr}, A_{m'r'})$, whereas the method of Thompson and Zucchini does not. In either method, two areas estimated form the same sample of subjects are correlated. However, neither method includes terms for the interaction of sample with modality, reader or the modality-by-reader interaction. Thompson and Zucchini hypothesized that the variance components associated with these random interaction terms were not likely to be substantial when dealing with samples of subjects (given the sample size) [15]. The variance component for the reader-by-sample interaction is not a component of $\text{var}(\hat{A}_{1.} - \hat{A}_{2.})$; the other two are.

When the variance components for the other interactions are positive, assuming that they are zero leads to underestimation of $\text{var}(\hat{A}_{1.} - \hat{A}_{2.})$. These variance components cannot be directly estimated from the data [15]. However, one can conceive of reasonable situations in which they would be positive. For example, in comparing digital and screen-film mammography, we expect digital to perform better in dense breasts. Thus, if the distribution of breast density can vary across samples, the component of variance for modality-by-sample interaction may be positive. Similarly, we expect dedicated mammographers to perform better than general radiologists for breasts with lesions that are difficult to detect. If the spectrum of difficulty can vary across samples, the component of variance for the reader-by-sample interaction may be positive. Combining the two situations would lead to a possible positive component of variance for the three-way interaction. We thus expect the F^* method to underestimate $\text{var}(\hat{A}_{1.} - \hat{A}_{2.})$, as some components of variance associated with the subject sample are assumed to be zero.

3. APPLICATIONS TO REAL DATA

Table I summarizes the characteristics of six data sets used to illustrate analyses following the methods described above. The test results in data sets 1–3 are five-point ratings of suspicion of abnormality: definitely normal; probably normal; possibly abnormal; probably abnormal; definitely abnormal. The test results in data sets 4–6 are six-point ratings of suspicion of abnormality from ‘lesion absent, definite’ through ‘lesion present, definite.’ Each data set was analysed using the GEE/OR, JP, and F^* methods. The GEE/OR method was run with a working independence assumption and with the 3PC structure, as described in Section 2.

Table II shows that estimates of $\hat{A}_{1.}$ and $\hat{A}_{2.}$ for each data set agreed across the three methods. Table II also illustrates differences between average areas and areas under ‘typical’ ROC curves. Note that $\hat{A}_{m.}$s, $\check{A}_{m.}$s, and $\tilde{A}_{m.}$s from GEE/OR using independence and using 3PC are very similar, providing empirical support for the robustness of $\hat{B}$ to misspecification of the

Table I. Data sets.

Data set	Description	Modalities*	Readers	Subjects	Range of
				$-:+$	$A_1 - A_2^{\dagger}$
1	Colorectal cancer [16]	CT, MR	2	16:45	0.181, 0.309
2	CT for head injury [17]	with, without clinical history	4	54:35	0.021, 0.047
3	Neonatal radiographs [18]	PACS, plain film	4	33:67	0.003, 0.037
4	US for gallstones [19] (wet better)	Wet, dry	4	20:20	0.022, 0.036
5	US for gallstones [19] (dry better)	Dry, wet	6	20:20	0.005, 0.069
6	Calibration curves for mammograms [20]	Barten, SMPTE	6	14:36	0.016, 0.087

* CT, computed tomography; MR, magnetic resonance imaging; PACS, picture archiving and communication system; US, ultrasound, wet is traditional chemical film processing, dry is without external chemicals; SMPTE, Society for Motion Picture and Television Engineers.

† Based on bivariate binormal models within each reader.

correlation model. The average of the areas under individual reader curves, $\hat{A}_{m.}$, was generally less than the area under the average curve as defined by averaging the a and b parameters, $\check{A}_{m.}$ Differences were generally less than 0.01. Also, $\check{A}_{m.}$ was generally less than the area under the average curve defined by obtaining a and b from average $\hat{\alpha}_{m.}$ and $\hat{\beta}_{m.}$, $\tilde{A}_{m.}$. Differences between $\hat{A}_{m.}$ and $\tilde{A}_{m.}$ were also mainly less than 0.01. Noticeable differences occur for CT in data set 1 ($\hat{A}_{m.}$ versus $\check{A}_{m.}$), and for each modality in data set 5 ($\hat{A}_{m.}$ versus $\tilde{A}_{m.}$). In data set 1, the position of the ‘typical’ ROC curve obtained from $(\hat{a}_{m.}, \hat{b}_{m.})$ is heavily influenced by the higher of the two individual ROC curves. In data set 5, the positions of the ‘typical’ ROC curves obtained from $(\hat{\alpha}_{m.}, \hat{\beta}_{m.})$ are heavily influenced by two highly asymmetric individual ROC curves.

Obtaining estimates of typical ROC curves has the benefit of indicating ranges of sensitivity and specificity where differences between modalities are likely to occur. This information can be combined with the location of observed operating points (sensitivity, 1 - specificity pairs) to determine whether differences are likely to be realized in practice. If the ROC curves for different modalities are similar throughout the range of observed operating points, then we would not expect to observe differences in accuracy across modalities in practice, whether or not the overall areas under the curves are different. However, if the ROC curves are different within the range of observed operating points, particularly those which may be used to define criteria for medical management, then differences in accuracy may be observed in practice. Again, this will be true whether or not the overall areas under the ROC curves are different.

Figure 1 illustrates estimated ROC curves from GEE/OR for data set 6. The observed operating points indicate that for each test modality, each reader operates to some degree in the region of high sensitivity and high specificity. For the Barten monitor calibration curve (left panel), readers 4 and 6 perform worse than the other readers. ROC curves for the other four readers cross each other at sensitivity slightly less than 0.8 and specificity nearly 1.0. For the Society of Motion Picture and Television Engineers (SMPTE) monitor calibration curve (right panel), reader 2 performs better than the other readers. ROC curves for the other five readers cross each other, mainly for sensitivity and specificity both slightly less than 0.8. The average ROC curve based on $\hat{a}_{m.}, \hat{b}_{m.}$ and the average ROC curve based on $\tilde{a}_{m.}, \tilde{b}_{m.}$ are similar. Where differences occur, the ROC curve based on $\tilde{a}_{m.}, \tilde{b}_{m.}$ tends to be closer to the majority

Table II. Estimated areas under ROC curves within each test modality.

Data set			GEE/OR	GEE/OR		
	Modality*	Result	Independence	3PC†	JP	F^*
1	CT	$\hat{A}_{m.}$	0.7355	0.7352	0.7543	0.7445
		$\check{A}_{m.}$	0.7708	0.7705		
		$\tilde{A}_{m.}$	0.7417	0.7413		
	MR	$\hat{A}_{m.}$	0.4988	0.4990	0.4950	0.4993
		$\check{A}_{m.}$	0.4988	0.4990		
		$\tilde{A}_{m.}$	0.4987	0.4989		
2	With History	$\hat{A}_{m.}$	0.9774	0.9773	0.9800	0.9762
		$\check{A}_{m.}$	0.9804	0.9803		
		$\tilde{A}_{m.}$	0.9807	0.9806		
	Without History	$\hat{A}_{m.}$	0.9422	0.9425	0.9453	0.9433
		$\check{A}_{m.}$	0.9476	0.9477		
		$\tilde{A}_{m.}$	0.9539	0.9541		
3	PACS	$\hat{A}_{m.}$	0.8684	0.8683	0.8681	0.8674
		$\check{A}_{m.}$	0.8699	0.8698		
		$\tilde{A}_{m.}$	0.8742	0.8741		
	Plain Film	$\hat{A}_{m.}$	0.8466	0.8463	0.8503	0.8468
		$\check{A}_{m.}$	0.8477	0.8474		
		$\tilde{A}_{m.}$	0.8477	0.8474		
4	Wet	$\hat{A}_{m.}$	0.8790	0.8789	0.8772	0.8804
		$\check{A}_{m.}$	0.8832	0.8832		
		$\tilde{A}_{m.}$	0.8847	0.8847		
	Dry	$\hat{A}_{m.}$	0.8468	0.8468	0.8472	0.8504
		$\check{A}_{m.}$	0.8486	0.8487		
		$\tilde{A}_{m.}$	0.8561	0.8561		
5	Dry	$\hat{A}_{m.}$	0.8915	0.8915	0.8941	0.8927
		$\check{A}_{m.}$	0.9067	0.9067		
		$\tilde{A}_{m.}$	0.9422	0.9422		
	Wet	$\hat{A}_{m.}$	0.8588	0.8588	0.8619	0.8610
		$\check{A}_{m.}$	0.8668	0.8668		
		$\tilde{A}_{m.}$	0.8853	0.8853		
6	Barten	$\hat{A}_{m.}$	0.9373	0.9375	0.9377	0.9393
		$\check{A}_{m.}$	0.9484	0.9489		
		$\tilde{A}_{m.}$	0.9511	0.9516		
	SMPTE	$\hat{A}_{m.}$	0.8834	0.8835	0.8836	0.8873
		$\check{A}_{m.}$	0.8969	0.8972		
		$\tilde{A}_{m.}$	0.8934	0.8936		

* CT, computed tomography; MR, magnetic resonance imaging; PACS, picture archiving and communication system; US, ultrasound, wet is traditional chemical film processing, dry is without external chemicals; SMPTE, Society for Motion Picture and Television Engineers.

† 3PC, three types of polychoric correlations.

of the individual reader ROC curves within each modality. The average ROC curve for the Barten monitor calibration curve dominates the average ROC curve for the SMPTE monitor calibration curve. The locations of the observed operating points indicate that the difference

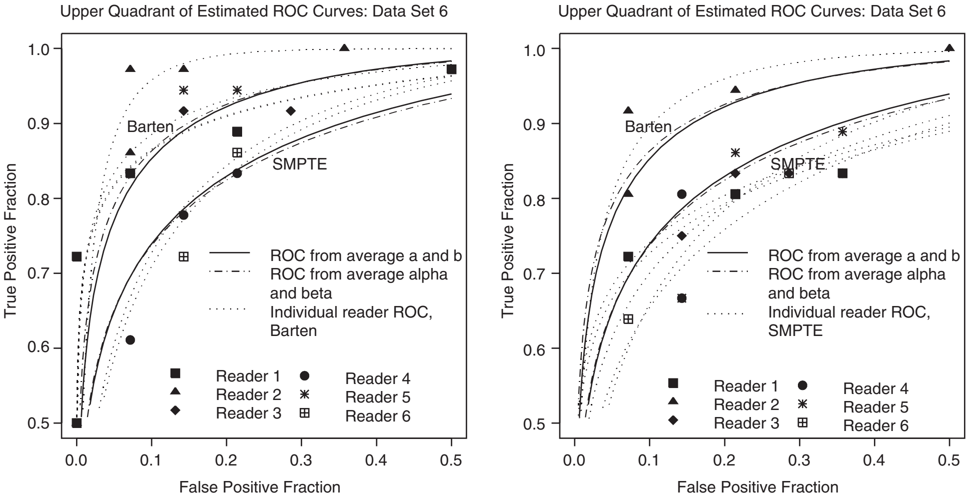

Figure 1. High sensitivity, high specificity quadrant of estimated ROC curves from GEE/OR for data set 6. Parameter estimates from the working independence assumption are used, because robust and model-based standard errors for $\hat{A}_{1.} - \hat{A}_{2.}$ matched more closely than they did for 3PC. The left panel includes individual reader ROC curves for the Barten monitor calibration curve, and the right panel includes individual reader ROC curves for the SMPTE Society for Motion Picture and Television Engineers monitor calibration curve.

in performance between the Barten and SMPTE monitor calibration curves is likely to be realized in practice.

Table III shows the estimated difference in areas beneath the ROC curve for the two test modalities, and standard errors and confidence intervals for those differences. Robustness of $\hat{B}$ to the choice of correlation model is supported by the similarity of $\hat{A}_{1.} - \hat{A}_{2.}$ for GEE/OR using either a working assumption of independence or using 3PC. The robust standard error (SE) of $(\hat{A}_{1.} - \hat{A}_{2.})$ is also similar for the two correlation models, while the model-based SE varies between models. By noting similarities of robust and model-based SEs for each data set, we see that for comparing average areas, 3PC fits better for data sets 2, 3 and 4, whereas independence fits better for data sets 1, 5 and 6.

Table III also shows that point estimates of $\hat{A}_{1.} - \hat{A}_{2.}$ are similar across the three analytical methods. Robust estimates of $\mathrm{SE}(\hat{A}_{1.} - \hat{A}_{2.})$ obtained from GEE/OR using either correlation model and obtained from JP also tend to agree with each other. Estimates of $\mathrm{SE}(\hat{A}_{1.} - \hat{A}_{2.})$ obtained from F^* tend to be noticeably smaller, as do widths of confidence intervals (CIs) for $\hat{A}_{1.} - \hat{A}_{2.}$. The exception is for data set 1, where F^* uses a t-distribution with 1 d.f., that is, the Cauchy distribution, while CIs under JP and GEE/OR are based on a t-distribution with 60 d.f. and the standard normal distribution, respectively.

4. SIMULATION STUDIES

We carried out simulations to evaluate estimates $\hat{A}_{1.} - \hat{A}_{2.}$ and $\mathrm{var}(\hat{A}_{1.} - \hat{A}_{2.})$ provided by the GEE/OR method. Sample sizes and parameter values were selected to reflect the data

Table III. Estimated differences in areas under the ROC curve for the two test modalities.

Data set	Model	Difference	Standard error	Confidence interval
1	GEE/OR: Ind.*	0.2367	0.0980, 0.0981	(0.0446, 0.4289); (0.0446, 0.4289)
	GEE/OR: 3PC*	0.2362	0.0991, 0.0982	(0.0419, 0.4305); (0.0437, 0.4287)
	JP	0.2594	0.1043	(0.0587, 0.4600)
	$F^{*\dagger}$	0.2325	0.0740	$(-0.7072, 1.1721)$
2	GEE/OR: Ind.	0.0352	0.0169, 0.0139	(0.0197, 0.0684); (0.0080, 0.0624)
	GEE/OR: 3PC	0.0348	0.0155, 0.0138	(0.0045, 0.0652); (0.0078, 0.0619)
	JP	0.0347	0.0135	(0.0078, 0.0617)
	F^*	0.0329	0.0059	(0.0141, 0.0517)
3	GEE/OR: Ind.	0.0217	0.0276, 0.0193	$(-0.0323, 0.0758)$; $(-0.0161, 0.0596)$
	GEE/OR: 3PC	0.0220	0.0200, 0.0194	$(-0.0173, 0.0612)$; $(-0.0162, 0.0601)$
	JP	0.0178	0.0204	$(-0.0224, 0.0581)$
	$F^{*\dagger}$	0.0206	0.0072	$(-0.0024, 0.0436)$
4	GEE/OR: Ind.	0.0321	0.0427, 0.0387	$(-0.0515, 0.1158)$; $(-0.0438, 0.1081)$
	GEE/OR: 3PC	0.0321	0.0362, 0.0387	$(-0.0389, 0.1031)$; $(-0.0438, 0.1080)$
	JP	0.0300	0.0423	$(-0.0537, 0.1137)$
	F^*	0.0300	0.0032	(0.0197, 0.0404)
5	GEE/OR: Ind.	0.0326	0.0331, 0.0322	$(-0.0322, 0.0975)$; $(-0.0304, 0.0957)$
	GEE/OR: 3PC	0.0327	0.0276, 0.0322	$(-0.0215, 0.0868)$; $(-0.0304, 0.0957)$
	JP	0.0321	0.0324	$(-0.0319, 0.0961)$
	F^*	0.0317	0.0100	(0.0060, 0.0574)
6	GEE/OR: Ind.	0.0539	0.0249, 0.0250	(0.0051, 0.1027); (0.0049,0.1029)
	GEE/OR: 3PC	0.0540	0.0216, 0.0248	(0.0117, 0.0962); (0.0054,0.1025)
	JP	0.0542	0.0260	(0.0020, 0.1063)
	F^*	0.0520	0.0153	(0.0127, 0.0913)

* Ind., independence; 3PC, three types of polychoric correlations. For GEE models, robust standard errors (SEs) are preceded by model-based SEs, and similarly for confidence intervals.

† $\hat{\rho}_m^* < \hat{\rho}_s^*$, so no adjustment is made.

sets in Table I. Each data set contained an octavariate response for each of 20 negative and 40 positive subjects. The octavariate response represented results from four readers in two modalities. Data were generated under two scenarios, regarding readers as fixed or as random (see Appendix).

We analysed the data sets using GEE/OR models and both the working independence (misspecified) and 3PC (true) correlation models. For the fixed reader setting, the true difference is 0.0551; 2438 data sets (97.52 per cent) provided estimates under WI, and 2360 (94.40 per cent) provided estimates under PC. For the random reader setting, the average of the true differences among the first 2500 non-degenerate data sets was 0.0408. Under WI, 2265 data sets (90.60 per cent) provided estimates; the average of the true differences among these was 0.0405. Under PC, 2137 data sets (85.48 per cent) provided estimates; the average of the true differences among these was 0.0403. The proportions of data sets providing results are reasonable in light of the complexity of the models. Table IV summarizes the results.

Table IV. Results of simulation studies.

Model*	Per cent bias	Sampling SE†	Model-based average SE	Robust average SE	Model-based CP‡	Robust CP
Fixed reader scenario						
Independence	−1.5268	0.0286	0.0368	0.0283	0.9877	0.9467
3PC	−3.1752	0.0284	0.0288	0.0282	0.9466	0.9453
Random reader scenario						
Independence	−0.9162	0.0394				
Conditional			0.0342	0.0265	0.9868	0.9497
Marginal			0.0342	0.0265	0.9188	0.8260
Adjusted			0.0511	0.0463	0.9788	0.9629
3PC	−3.7225	0.0388				
Conditional			0.0271	0.0265	0.9560	0.9527
Marginal			0.0271	0.0265	0.8437	0.8297
Adjusted			0.0462	0.0459	0.9649	0.9616

* 3PC, three types of polychoric correlations.

† SE, standard error; adjusted SEs allow for between-reader correlations.

‡ CP, coverage probability; conditional CP is for within-data set true differences: marginal and adjusted CPs are for average of within-data set true differences.

4.1. Bias

There was a slight negative bias in estimating the difference in average area between the two modalities. This was more pronounced for estimation under the polychoric correlation model (PC) than for estimation under the working independence model (WI). Part of this bias may be attributed to larger differences among the subset of data sets that did not provide estimates, as is noted above for the random reader situation. Under WI, the mean estimated difference was within −1.53 per cent error for fixed readers and within −0.92 per cent error for random readers.

4.2. Confidence interval coverage and estimated variance

Under WI, coverage rates for model-based CIs are notably higher than for robust CIs. This reflects the fact that the model-based estimate of the variance of the difference does not take into account correlations among the eight responses on each subject, and is thus too high. Under PC, coverage levels for model-based CIs were similar to those for robust CIs. In the fixed reader setting, robust CIs covered the true difference 94.67 per cent of the time under WI and 94.53 per cent of the time under PC. In the random reader setting, 94.97 per cent of the robust CIs covered the within data set truth under WI, and 95.27 per cent of the robust CIs covered the within data set truth under PC. However, coverage of the overall true difference was decidedly too low, reflecting the fact that estimates of var($\hat{A}_{1.} - \hat{A}_{2.}$) are made under a model of fixed effects for readers.

An adjusted estimator of the variance of the difference can be used when considering effects of readers to be random. For R readers, and Δ the true difference between areas for the two modalities, consider the observed differences $\hat{\Delta}_1, \ldots, \hat{\Delta}_R$ as independent draws from a distribution with mean Δ and variance σ_b^2, where σ_b^2 represents the variance component for readers. We estimate Δ by $\hat{\Delta}=(\sum\hat{\Delta}_r)/R$. Given the observed differences, $\hat{\Delta}|\hat{\Delta}_1,\ldots,\hat{\Delta}_R$ has mean Δ and variance σ_d^2; σ_d^2 is the quantity estimated by the GEEs. Under this hierarchy

$$\text{var}(\hat{\Delta})=\sigma_d^2+\sigma_b^2/R$$

This leads to the suggestion of estimating the variance of the within-reader differences, dividing by R, and adding it to the variance estimate obtained from the GEEs. The adjusted estimator provides slightly conservative coverage levels for robust CIs: 96.29 per cent under WI, and 96.16 per cent under PC. The slight conservativeness results from the fact that the adjusted SE is higher than the observed SE, to a degree that goes beyond any necessary inflation to overcome the slight bias in the estimate of the difference.

5. DISCUSSION

This paper compared three methods for analysing multiple correlated ROC curves. All three methods can be used when ROC curves are estimated from test results that occur in ordered categories, as in the current paper. The JP and F^* methods may also be used for continuous test results. The JP and F^* methods analyse summary measures, while the GEE/OR method analyses the observed test results. For all three methods, summary statistics other than the area under the curve can be used. For the data sets in this paper, the average of the areas under the individual reader curves within each modality was generally less than the area under an average ROC curve defined in either of two ways. This points out the difference between estimating 'typical' ROC curves and summary statistics thereof, versus estimating 'typical' values of a chosen summary statistic. While the magnitude of the differences observed in our six examples was generally small, there were notable exceptions. Further, an important question is the scale on which any random effects for readers are operating. If one believes that there is a distribution of summary statistics, then averaging those summary statistics is sensible. If, however, one believes that the random effects distribution pertains to the ordinal test results, then averaging over the parameter estimates from an ordinal regression model (including the binormal model as a special case) would be more appropriate. This issue is one which deserves further research.

An important concern when working with real data is how to perform analyses when data for responses and/or covariates, particularly true disease status, are incomplete. Toledano and Gatsonis have extended the GEE/OR method to these situations [21]. The JP method can be extended to these situations when univariate procedures for estimating ROC curves in the presence of missing data exist. The F^* method can be extended to accommodate missing data when an appropriate bivariate or M-variate procedure for estimating ROC curves exists.

The GEE/OR approach can allow the ROC curve, including the placement of operating points, to depend on characteristics of readers and of subjects in addition to true disease status. This is important, since such characteristics often affect diagnostic accuracy [5]. The JP method allows summary measures to depend on these characteristics. The F^* method allows summary measures to depend on characteristics of readers, but not of subjects. Owing to these differences, in the current paper we allowed the ROC curve to depend only on true disease status. The correlations modelled in GEE/OR are for the underlying hypothetical continuous degree-of-suspicion variables. These correlations may also be modelled as varying within type, and may depend on covariates [4, 5]. The F^* method models correlations among the summary measures. These correlations must be common within correlation type, and may not depend on covariates. Owing to this difference, the current paper did not use more complex correlation models with GEE/OR. The JP method uses random effects to account for correlations.

Both the JP and F^* methods model the effects of readers as random. This is appropriate when we conceive of the readers as being drawn from a larger population [9]. Practical considerations, however, often result in studies that are limited to a small number of selected expert readers, such that generalization to the larger population of readers may not be appropriate. Further, even if the sample of readers were representative of the larger population, the small number of readers could negatively impact the quality of the estimation of the variance of the random effects distribution. The flexible nature of the correlation model in the GEE/OR method can be exploited to specify a structure of correlations across the repeated measures on each subject equivalent to what we would observe in a model with random effects for readers. As we saw in Table III, this resulted in SE($\hat{A}_{1.} - \hat{A}_{2.}$)s that were highly similar for GEE/OR and JP.

Six data sets were used to compare numerical results across the three methods. For the GEE/OR method, parameter estimates and robust SEs were indeed robust to choice of correlation model. Estimated differences in average area under the ROC curve for the two test modalities were similar for all three methods. A major difference in the numerical results obtained by the three methods is that SE($\hat{A}_{1.} - \hat{A}_{2.}$) obtained from F^* was considerably smaller than from JP and GEE/OR. As discussed in Section 2.4, this is due in part to the form of the ANOVA model used in the F^* method, and in part to ignoring between-subject correlations in the JP and GEE/OR methods.

The significance of modality-by-subject and reader-by-subject interactions estimated using the JP method does not indicate whether modality-by-sample or reader-by-sample interactions for the F^* method would be significant. For example, whether or not there is a modality-by-subject interaction, if the case mix (for example, spectrum of disease severity, difficulty of interpretation etc.) differs across samples, there may be a modality-by-sample interaction. The interactions at issue can be estimated in an experiment that uses multiple subject samples for each modality–reader combination, which does not generally occur in medical studies, or by obtaining separate $\hat{A}_{mr}$s for partitions of a single sample of subjects [15]. The latter method requires deciding the number of subgroups to use, and may not have adequate power to detect interactions when they exist. It also implicitly assumes that the case mix in the sample obtained is representative of that in the larger population.

Obuchowski and Rockette [7] performed a Monte Carlo simulation study which showed that the F^* test approached nominal α levels only when R became as large as 8 to 12. Dorfman *et al.* [22] performed a bootstrap analysis of data set 3 that provides empirical evidence for the numerical results obtained by the JP and GEE/OR methods for those data. Dorfman *et al.* [23] also performed a Monte Carlo simulation study of the empirical type I error rate for the JP method under the null hypothesis that the diagnostic accuracies of two modalities are the same for the factorial experimental design. Empirical type I error rates were within the 95 per cent probability bands for a nominal α level of 0.05 when $A_{m.}$ was 0.855 (similar to many of the $\hat{A}_{m.}$s in our examples), n was at least 100, and there were 5 or 10 readers. Lipsitz *et al.* [24] performed simulation studies for $n = 15$, 30 and 45, with three binary responses per subject, where the true correlation for all pairs of responses on a given subject was 0.15, 0.45 or 0.60. They showed that when using the correct models, 95 per cent confidence intervals from GEEs had appropriate coverage probabilities, and that the type I error rate was not significantly different from a nominal type I error rate of 0.05.

We performed simulation studies directed at determining whether $\mathrm{SE}(\hat{A}_{1.}-\hat{A}_{2.})$ was estimated appropriately by the GEE/OR method. We conclude that if one is interested primarily in estimating the difference in areas across modalities, and not in estimating correlations, the working independence model and accompanying robust standard errors perform very well. Estimates of the difference had on average a slight negative bias, of a magnitude that is likely negligible in practical significance. The standard error of $(\hat{A}_{1.}-\hat{A}_{2.})$ was estimated appropriately under the conception of fixed readers, providing 94.67 per cent coverage for nominal 95 per cent CIs in the fixed reader simulations, and 94.97 per cent within data set coverage in the random reader simulations. A simple and intuitive augmentation of $\mathrm{SE}\,(\hat{A}_{1.}-\hat{A}_{2.})$ provided by the GEEs was proposed for the situation of random reader effects conceived such that, similar to the JP and F^* methods, there is a distribution of summary statistics across readers. The augmented SEs performed well, providing a conservative estimate of $\mathrm{SE}(\hat{A}_{1.}-\hat{A}_{2.})$ and 96.29 per cent coverage of the overall true difference.

Software for analysing data using the GEE/OR method is available by contacting Toledano, and for analysing data using the JP method by contacting Metz (Department of Radiology, the University of Chicago). Data analysis using the F^* method is based on combining results from available software to perform ANOVA and M-variate ROC curve analysis.

APPENDIX

Data were generated under two scenarios, reflecting fixed readers and random readers. For the fixed reader scenario, four draws $(\alpha_{1r},\beta_{1r},\alpha_{2r},\beta_{2r})$, $r=1,2,3,4$ were generated from a quadrivariate normal distribution with mean $(2.0,0.2,1.9,0.4)$ and variance

$$\mathscr{A}\begin{bmatrix}1.00 & 0.78 & 0.43 & 0.25\\ 0.78 & 1.00 & 0.27 & 0.30\\ 0.43 & 0.27 & 1.00 & 0.78\\ 0.25 & 0.30 & 0.78 & 1.00\end{bmatrix}\mathscr{A}$$

for $\mathscr{A}$ the matrix with $(0.50,0.05,0.50,0.10)$ on the diagonal and zero elsewhere. Responses for negative subjects were generated using repeated samples from a standard octavariate normal distribution with correlations $\rho_{-m}=0.30$, $\rho_{-r}=0.60$, and $\rho_{-s}=0.25$. Responses for positive subjects were generated using repeated samples from an octavariate normal distribution with $E(Y_{mr})=\alpha_{mr}$, $\mathrm{var}(Y_{mr})=\mathrm{e}^{2\beta_{mr}}$, $\mathrm{cov}(Y_{mr},Y_{mr'})=0.60\mathrm{e}^{\beta_{mr}+\beta_{mr'}}$, $\mathrm{cov}(Y_{mr},Y_{m'r})=0.75\mathrm{e}^{\beta_{mr}+\beta_{m'r}}$ and $\mathrm{cov}(Y_{mr},Y_{m'r'})=0.55\mathrm{e}^{\beta_{mr}+\beta_{m'r'}}$; $m=1,2$, $m\neq m'$, $r\neq r'$. Continuous data were categorized using cutpoints equally spaced in an equal-component-weight mixture distribution of $\mathrm{N}(0,1)$ and $\mathrm{N}(\alpha_{mr},\mathrm{e}^{2\beta_{mr}})$. The resulting categorical data was then assembled into data sets consisting of 20 negative and 40 positive subjects, and evaluated to ensure non-degeneracy. The GEE/OR method was used to analyse 2500 non-degenerate data sets.

For the random reader scenario, four draws $(\alpha_{1r},\beta_{1r},\alpha_{2r},\beta_{2r})$, $r=1,2,3,4$, were generated as above, and a single set of responses for 20 negative and 40 positive subjects was drawn and categorized similar to the above. The entire process of parameter and response draws was repeated, and the GEE/OR method was used to analyse 2500 non-degenerate data sets.

ACKNOWLEDGEMENTS

This research was supported in part by grant CA76310 from the National Cancer Institute. The author thanks James M. Hanley, Elizabeth Krupinski, and the Radiology Diagnostic Oncology Group for providing data sets, Benjamin A. Herman for programming support, and Paul Rathouz, Charles E. Metz and the referee for helpful comments on the manuscript. The author extends special thanks to Donald D. Dorfman for providing a data set and for many thought-provoking conversations.

REFERENCES

1. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. Rating method data. *Journal of Mathematical Psychology* 1969; **6**:487–496.
2. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996; **201**:745–750.
3. McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; **9**:190–195.
4. Toledano A, Gatsonis C. Regression analysis of correlated receiver operating characteristic data. *Academic Radiology* 1995; **2**:S14–S21.
5. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996; **15**:1807–1826.
6. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. *Investigative Radiology* 1992; **27**:723–731.
7. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Communications in Statistics, Part B: Simulation and Computation* 1995; **24**:285–308.
8. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Academic Radiology* 1995; **2**:S22–S29.
9. Beam CA. Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. *Academic Radiology* 1995; **2**:S4–S13.
10. Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988; **8**:204–215.
11. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B* 1980; **42**:109–142.
12. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
13. Metz CE, Wang P, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging*. Martinus Nijhoff: The Hague, 1984; 432–445.
14. Hinkley D. Jackknife methods. In *Encyclopedia of Statistical Sciences, Volume 4: Icing the Tails to Limit Theorems.* Wiley: New York, 1983; 280–287.
15. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. *Statistics in Medicine* 1989; **8**: 1277–1290.
16. Zerhouni EA, Rutter C, Hamilton SR, Balfe DM, Megibow AJ, Francis IR, Moss AA, Heiken JP, Tempany CMC, Aisen AA, Weinreb JC, Gatsonis C, McNeil BJ. CT and MR imaging in the staging of colorectal carcinoma: report of the Radiology Diagnostic Oncology Group II. *Radiology* 1996; **200**:443–451.
17. McNeil BJ, Hanley JA, Funkenstein HH, Wallman J. The use of paired ROC curves in studying the impact of history on radiographic interpretation: CT of the head as a case study. *Radiology* 1983; **149**:75–77.
18. Franken EA, Berbaum KS, Marley SM, Smith WL, Sato Y, Kao SC, Milam SG. Evaluation of a digital workstation for interpreting neonatal examinations: a receiver operating characteristic study. *Investigative Radiology* 1992; **27**:732–737.
19. Krupinski EA. Clinical assessment of dry laser-processed film versus traditional wet-processed film with computed tomography, magnetic resonance imaging, and ultrasound. *Academic Radiology* 1996; **3**:855–858.
20. Krupinski EA, Roehrig H. The influence of a perceptually linearized display on observer performance and visual search. *Academic Radiology* 2000; **7**:8–13.

21. Toledano AY, Gatsonis C. Missing data in receiver operating characteristic curve analysis. *Biometrics* 1999; **55**:488–496.
22. Dorfman DD, Berbaum KS, Lenth RV. Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. *Academic Radiology* 1995; **2**:626–633.
23. Dorfman DD, Berbaum KS, Lenth RV, Chen Y-F, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. *Academic Radiology* 1998; **5**:591–602.
24. Lipsitz SR, Fitzmaurice GM, Orav EJ, Laird NM. Performance of generalized estimating equations in practical situations. *Biometrics* 1994; **50**:270–278.