STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2314

A Bayesian hierarchical approach to multirater correlated ROC analysis

Timothy D. Johnson*,† and Valen E. Johnson

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.

SUMMARY

In a common ROC study design, several readers are asked to rate diagnostics of the same cases processed under different modalities. We describe a Bayesian hierarchical model that facilitates the analysis of this study design by explicitly modelling the three sources of variation inherent to it. In so doing, we achieve substantial reductions in the posterior uncertainty associated with estimates of the differences in areas under the estimated ROC curves and corresponding reductions in the mean squared error (MSE) of these estimates. Based on simulation studies, both the widths of coverage intervals and MSE of estimates of differences in the area under the curves appear to be reduced by a factor that often exceeds five. Thus, our methodology has important implications for increasing the power of analyses based on ROC data collected from an available study population. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: receiver operating characteristics; ROC analysis; area under the ROC curve; Bayesian analysis; hierarchical model

INTRODUCTION

Multirater correlated receiver operating characteristics (ROC) analysis refers to a particular type of study in which multiple readers rate several diagnostic tests generated from data collected on the same subject. This design is common in radiologic studies where, for example, radiologists evaluate images collected from the same patient using distinct image modalities (e.g. PET, CT and MRI) or different reconstruction algorithms within the same imaging modality. Outcomes from such a study design represent correlated ordinal data.

As a concrete example of such a study (the motivating example for this manuscript) an unpublished ROC study was conducted in 1993 in the UCLA Department of Radiology. The

*Correspondence to: Timothy D. Johnson, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A.
†E-mail: tdjtdj@umich.edu

Contract/grant sponsor: U.S. National Institute of Health; contract/grant number: IPOI CA 87684

Received July 2004
Accepted May 2005

purpose of this study was to compare the diagnostic capabilities of chest film radiographs to digitized images displayed on a 1K by 1K video display. The basis of comparison were radiologists' ability to detect lung nodules in the radiographs. A panel of expert radiologists determined 'truth' by consensus on 772 archived chest radiographs (59 cases with nodules and 713 disease free cases). Each of the 772 radiographs were then digitized for video display. Three experienced radiologists and two radiology residents read each case under both systems (film *versus* video display) and ranked the presence of nodules on a scale of 1–5. Large ratings represented high confidence that nodules were present.

Much of the recent research in ROC methodology has focused on the inclusion of covariate effects and the combination of independent rating information collected from multiple raters (e.g. References [1–5]). In contrast, we propose a Bayesian hierarchical latent variable model for analysing multirater correlated ordinal ROC data. Related models for data collected from multiple raters but premised on a frequentist perspective are described in References [6–8]. Analyses of data collected in designs where only one reader rates the outcomes, both using parametric [9] and non-parametric methods [10–12] have also been developed. After describing our framework, we use simulation studies to compare our model to one of these alternatives in the 'A Data Analysis and Simulation Study' section.

Bayesian approaches to this problem have also been explored. Among the earlier efforts in this direction are those detailed in Ishwaran and Gatsonis [3] and Johnson and Albert [13]. In this article, we describe a hierarchical latent variable model for analysing multirater correlated ordinal ROC data that combines modelling aspects from each of these two earlier approaches. The primary innovation of this model over more commonly used ROC models is the manner in which it accounts for three sources of variation inherent to this study design; namely, variation in ratings attributable to differences in patient/subject characteristics, variation in ratings introduced by inaccuracies in the procedures used to define the diagnostic measure (modality effects), and variation attributable to readers of the diagnostic test. By explicitly modelling these three sources of variation, Bayesian models for ROC analysis are able to achieve substantial increases in power for detecting modality effects, which are the primary variables of interest in most ROC studies. This partitioning of error variances also facilitates the study of individual reader characteristics and provides a natural mechanism for predicting the diagnostic performance of a test when interpreted by a reader drawn randomly from the larger population of potential readers.

This article is organized as follows. In the next section, we review what is arguably the most widely used ROC model for the analysis of multirater correlated data, that of Dorfman *et al.* [6] (henceforth referred to as DBM). Following that, we present a Bayesian hierarchical model for the analysis of multirater correlated ROC data and highlight its connection to the standard bivariate–binormal model. We then illustrate our model in the analysis of a radiological data set intended to compare lung nodule detection using film *versus* a 1K by 1K video display and compare the performance of our model with that of DBM through a simulation study. We conclude with a sensitivity analysis and a brief discussion of results.

Throughout the remainder of the paper we fix ideas by assuming that the ROC study has been designed to compare two (or more) imaging modalities as diagnostic tools for some disease. We also assume that images from both modalities have been obtained from subjects drawn from one of two populations: a population of healthy subjects and a population of diseased subjects (here it is tacitly assumed that the true disease status of all subjects is known). Since each subject contributes one image from each modality, the data (images) are

correlated. We further assume that a random sample of radiologists (typically a small sample), who are expert in reading images from both modalities, are drawn from the population of all radiologists. Each radiologist is asked to rate, or score, both images contributed from all subjects in a random order. The rating scale is an ordinal scale, say 1–5, where 1 represents the radiologist's belief that disease is definitely absent, 2 represents the belief that disease is probably absent, 3 the belief that the absence/presence of disease cannot be determined, 4 the belief that disease is probably present and 5 represents the belief that disease is definitely present.

THE JACKKNIFE METHOD OF DBM

DBM's approach toward analysing multirater correlated ROC data [6] is probably the most widely used method for analysing mulitrater correlated ROC data. This method utilizes a two-stage procedure to evaluate multirater correlated ROC data. In the first stage, jackknifed pseudo-values [14] of the area under the ROC curve, A_Z, are obtained from the bivariate–binormal ROC model [15]. These pseudo-values are then post-processed using standard mixed effect analysis of variance (ME-ANOVA) software in the second stage.

Stage 1: In the first stage of modelling, the bivariate–binormal model is employed, one radiologist at a time, to obtain jackknifed pseudo-A_Z values for each radiologist. The bivariate–binormal model assumes that there are two underlying latent bivariate normal distributions: one for the healthy population and one for the diseased population. The components of each bivariate distribution describe the joint distribution of latent ratings obtained from the modalities on a continuous scale. These ratings represent 'extent of disease.' However, observed ratings are not recorded on this continuous scale. Instead, raters are assumed to group images into diagnostic categories, assigning an ordinal score to each. Assignment of images to ordinal categories is assumed to depend on a set of latent thresholds, unique to the radiologist and perhaps different for each modality. The latent traits from diseased subjects are assumed to be drawn from a bivariate normal distribution with mean $(\mu_1, \mu_2)^{\mathrm{T}}$ and covariance matrix Σ. The latent traits from healthy subjects are assumed to be drawn from a bivariate normal distribution with mean $(0,0)^{\mathrm{T}}$, marginal variances equal to 1, and covariance (or correlation) ρ. The $(0,0)^{\mathrm{T}}$ mean for the healthy population serves to centre the latent scale while the marginal variances for both modalities are set to 1 to establish a scale. Figure 1 illustrates these assumptions.

Based on this model, the area under the ROC curve for a modality (for a particular rater) is defined as the probability that a random variable drawn from the disease distribution is greater than an independently drawn variable from the healthy distribution [10]. That is, if U is a random variable drawn from the diseased distribution, and V is a random variable drawn from the 'healthy' distribution, then $A_Z = \Pr(U > V) = \Phi(\mu/\sqrt{1+\sigma^2})$, where $\Phi(x)$ is the standard normal distribution function and σ^2 is the marginal variance of the diseased population for the modality under consideration.

This classical model for the analysis of correlated ROC data lacks a component for rater variability. The marginal variances that result from the bivariate–binormal model are a convolution of the three components of variation: one due to the disease process itself, one for the modality, and one for the rater. Further, the marginal variances cannot be deconvolved into their constituent components. The model also lacks a natural mechanism for combining

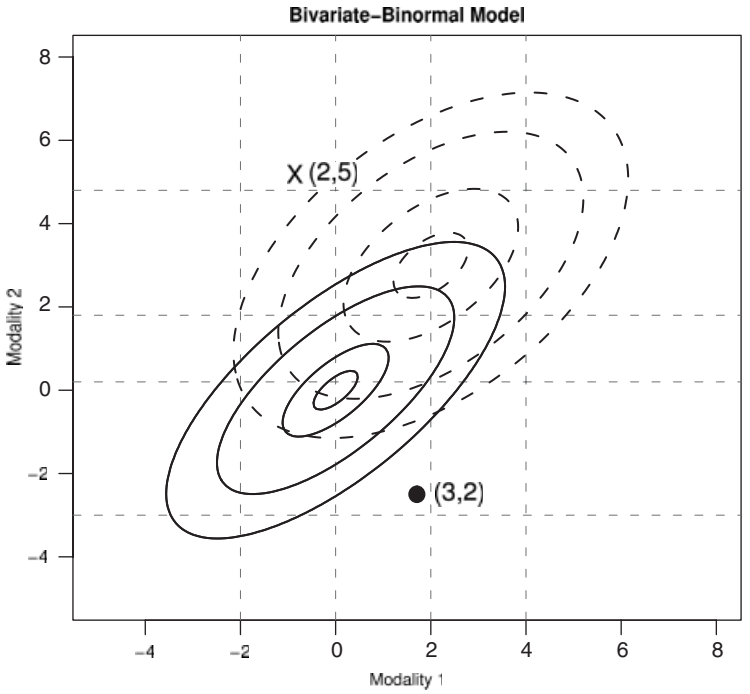

Figure 1. A graphical representation of the bivariate–binormal model. The healthy population is assumed to have a bivariate normal distribution with mean $(0,0)^{\mathrm{T}}$, marginal variances of 1 and a correlation of ρ. Contour lines of this population are solid. The disease population has a bivariate normal distribution with mean $(\mu_1,\mu_2)^{\mathrm{T}}$ and covariance matrix Σ. Contour lines of this population are dashed. The light vertical and horizontal dashed lines depict the rater threshold values. For example, a case whose latent value lies at the 'X', would be classified in category 2 under modality 1 and in category 5 under modality 2.

information across raters. Thus, several authors have proposed models that take resulting A_Z values (or pseudo-A_Z values) from this model and use them as observed data in a second stage model.

The DBM method uses the bivariate–binormal model to obtain jackknifed pseudo-A_Z values for each case and modality, one radiologist at a time. These pseudo-values are assumed to behave as independent observations in the second stage of modelling.

Stage 2: The psuedo-A_Z values are treated as observations in a ME-ANOVA model. Thus, there is one 'observation' for each combination of radiologist, modality, and case. The particular model fitted under the DBM framework can be expressed

$$\hat{A}_{ijk} = \mu + \alpha_k + B_j + C_i + (\alpha B)_{kj} + (\alpha C)_{ki} + (BC)_{ji} + (\alpha BC)_{kji} + \varepsilon_{ijk} \tag{1}$$

In this equation, $\hat{A}_{ijk}$ represents the pseudo-A_Z value for case i, modality k and radiologist j. We note that when each radiologist rates each image under each modality only once (a common study design), the terms $(\alpha BC)_{kji}$ and ε_{ijk} are inseparable. In DBM, the overall mean μ and the modality effects $\{\alpha_k\}$ are considered to be fixed effects under the

constraint that $\sum_k \alpha_k = 0$. The radiologist effects, $\{B_j\}$, case effects, $\{C_i\}$, interaction terms and model error terms are assumed to be mutually independent, mean zero normal random deviates with variances σ_B^2, σ_C^2, $\sigma_{\alpha B}^2$, $\sigma_{\alpha C}^2$, σ_{BC}^2, $\sigma_{\alpha BC}^2$ and σ_ε^2, respectively. Typically, differences between treatment means are assessed using Satterthwaite approximate F tests [16]. Confidence intervals for parameters of interest are constructed using an approximate Student-t distribution, although approximate confidence intervals for treatment means may also be derived using a reduced model defined by omitting all interaction terms except the rater-by-case interactions.

Zhou, Obuchowski, McClish ([17], henceforth ZOM) summarize three major shortcomings of this approach. First, they note that pseudo-values are treated as observed data. Using pseudo-values as observed data has only limited utility, and previous attempts to extract more than variance estimates from pseudo-values have not been successful [14]. Second, pseudo-values are, in general, correlated. Assuming they are not, violates an assumption of the ME-ANOVA. Third, this method applies the one-sample jackknife to a two sample problem (diseased and healthy cases). Also, we note that each A_Z value is supported on the interval $[0, 1]$, but observed pseudo-values often take values outside this interval.

Subsequent to the work of DBM, two alternative likelihood-based approaches for analysing this type of data were developed [7, 8]. Both approaches are similar to that proposed by DBM in that they both use a two-stage modelling procedure. At the first stage of modelling, A_Z values are computed one rater at a time using either the bivariate–binormal model or a non-parametric alternative (e.g. Reference [11]). In the second stage of modelling, the A_Z values are combined across raters and modalities using a ME-ANOVA model. A more detailed comparison and critique of these models is provided in ZOM. Because both methods produce estimates that have statistical properties similar to estimates produced by the DBM method, these models are not considered further here.

A BAYESIAN HIERARCHICAL MODEL

The model we propose is closely related to a simpler model described in Johnson and Albert [13]; it is also close to being a special case of the model proposed in Ishwaran and Gatsonis [3]. The primary generalization of this model over that described in Johnson and Albert is the inclusion of a more flexible class of prior distributions on model parameters. In contrast to the model proposed in Ishwaran and Gatsonis, we do not incorporate a regression model for the underlying latent variables, nor do we consider semi-parametric link functions to account for non-normality of the latent trait distributions. However, as Ishwaran and Gatsonis point out, such link functions are probably not necessary (or estimable) when ROC data are collected using a small number of categories. This is the case of primary interest here, as well as in most applications. We do, however, extend the Ishwaran and Gatsonis model by allowing for distinct rater thresholds for each rater.

Suppose then that there are N_h healthy cases and N_d disease cases for total of $N = N_\mathrm{h} + N_\mathrm{d}$ subjects. Let $\mathscr{D}$ denote the set of subjects classified as diseased and let $\mathscr{H}$ denote the set of subjects classified as healthy. Let $J > 1$ denote the number of readers of each diagnostic test and assume that each reader rates subjects who are diseased and healthy using measurements derived from each of $K > 1$ diagnostic tests. For notational simplicity, suppose that each subject is placed into one of C ordered categories by each reader under each test. The observed rating

from reader j scoring case i under test k is denoted by Y_{ijk}. We adopt the convention that larger values of Y_{ijk} are indicative of a higher degree of confidence that the subject has the disease. We assume the latent variable representation for the data $\mathbf{Y} = \{Y_{ijk}\}$ detailed in Reference [13]. Under this representation, the ordinal ratings of each case by each reader are hypothesized to result from noisy observations of a continuous, scalar-valued random variable representing the presence of a disease attribute. The distribution of this latent disease attribute is assumed to be drawn from one of two distributions, one for healthy subjects and one for diseased individuals. We adopt the binormal assumption and assume that these distributions are Gaussian. The practicality of this assumption is discussed in Reference [15], where an argument is presented to suggest that even non-Gaussian continuous data can be adequately represented under this model (when thresholds for the ordinal categories are estimated from data). The generality of this assumption is clarified further in Reference [18], who shows that there exists a monotone transformation of the continuous data to make the distributions of the healthy and diseased groups normal.

With these comments in mind, our hierarchical prior model for multirater correlated ROC data may be specified in stages as follows.

Stage 1: At the first level of the hierarchy we introduce variability inherent in the two populations of interest: the healthy and diseased populations. We assume that the latent (disease) trait for subject i ($i = 1, \ldots, N$), denoted by Z_i, follows a normal distribution. We assume that the latent value for healthy cases is marginally distributed as a $\mathrm{N}(0, 1)$ random variable, while the latent value for a diseased individual is distributed as a $\mathrm{N}(\mu, \psi^2)$ random variable. The parameters μ and ψ^2 are unknown and so are estimated from data. The first level of the model may be written as

$$Z_i \mid \mu, \psi^2 \overset{\text{iid}}{\sim} \mathrm{N}[\mu I(i \in \mathscr{D}), (\psi^2 - 1) I(i \in \mathscr{D}) + 1] \quad \forall i \tag{2}$$

where $I(A)$ is the indicator function equal to 1 if A holds and 0 otherwise.

Stage 2: At the second level of the hierarchy, we introduce error terms that reflect the infidelity of each diagnostic test in representing the true disease state of a subject. Specifically, we assume that modality k contributes an independent $\mathrm{N}(0, \phi_k^2)$ error to the observation of latent disease trait Z_i. This term accounts for inaccuracies and distortions introduced by the diagnostic modality. The parameter ϕ_k^2 denotes the variance of this error for modality k ($k = 1, \ldots, K$). We let Z_{ik} denote the value of the latent trait of case i that would be observed by an ideal rater (a rater who scores the cases with no variability) using modality k. That is

$$Z_{ik} \mid Z_i, \phi_k^2 \overset{\text{iid}}{\sim} \mathrm{N}(Z_i, \phi_k^2), \quad \forall k \tag{3}$$

Stage 3: At the third level of the model hierarchy, we model reader errors, allowing for the fact that different readers have different levels of expertise in interpreting images. Thus, we assume that the value Z_{ik} is further distorted by the addition of independent $\mathrm{N}(0, \theta_j^2)$ random variables that represent reader errors. The parameter θ_j^2 denotes the error variance particular to the jth reader, $j = 1, \ldots, J$. This stage of the model may be expressed as

$$Z_{ijk} \mid Z_{ik}, \theta_j^2 \overset{\text{iid}}{\sim} \mathrm{N}(Z_{ik}, \theta_j^2), \quad \forall j \tag{4}$$

Likelihood function: Given this prior structure, we complete the joint model specification by defining a sampling distribution for the observed data as follows. In so doing, we assume that each reader uses a unique set of thresholds γ_j to assign cases to diagnostic categories. The components of these threshold vectors satisfy $-\infty = \gamma_{j0} < \gamma_{j1} < \cdots < \gamma_{jC-1} < \gamma_{jC} = \infty$. Reader j assigns case i under modality k to category c if Z_{ijk} falls between the $(c-1)$st and cth threshold. That is,

$$Y_{ijk} = c \text{ if and only if } \gamma_{jc-1} < Z_{ijk} \leqslant \gamma_{jc}$$

From this assumption it follows that

$$\Pr(Y_{ijk} = c \mid \gamma_j, Z_{ijk}) = I(\gamma_{jc-1} < Z_{ijk} \leqslant \gamma_{jc}) \tag{5}$$

from which it follows that each observation contributes a term of the form

$$\begin{aligned} f(y_{ijk}, z_i, z_{ik}, z_{ijk} \mid \mu, \psi^2, \phi_k^2, \theta_j^2, \gamma_j) = I(\gamma_{jy_{ijk}-1} < z_{ijk} \leqslant \gamma_{jy_{ijk}}) \\ \times \frac{1}{\theta_j}\phi[(z_{ijk} - z_{ik})/\theta_j] \times \frac{1}{\phi_k}\phi[(z_{ik} - z_i)/\phi_k] \\ \times \left\{ \phi[(z_i - \mu I(i \in \mathscr{D}))/\sqrt{(\psi^2 - 1)I(i \in \mathscr{D}) + 1}] \right. \\ \left. \times \frac{1}{\sqrt{(\psi^2 - 1)I(i \in \mathscr{D}) + 1}} \right\} \end{aligned}$$

to the joint distribution over observations and parameters. In this expression, $\phi[\cdot]$ denotes the density function of the standard normal distribution. From the assumed independence of error terms in (2)–(4), we obtain

$$\begin{aligned} &f(\mathbf{y}, \{z_i\}, \{z_{ik}\}, \{z_{ijk}\} \mid \mu, \psi^2, \{\phi_k^2\}, \{\theta_j^2\}, \{\gamma_j\}) \\ &\quad = \prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K} f(y_{ijk}, z_i, z_{ik}, z_{ijk} \mid \mu, \psi^2, \phi_k^2, \theta_j^2, \gamma_j) \end{aligned}$$

Integrating over the latent observations yields

$$\begin{aligned} \Pr(\mathbf{Y} = \mathbf{y} \mid \mu, \psi^2, \{\phi_k^2\}, \{\theta_j^2\}, \{\gamma_j\}) = \prod_{ijk}\left[\Phi\left(\frac{\gamma_{jy_{ijk}} - \mu I(i \in \mathscr{D})}{s_{jk}} \right) \right. \\ \left. - \Phi\left(\frac{\gamma_{jy_{ijk}-1} - \mu I(i \in \mathscr{D})}{s_{jk}} \right) \right] \end{aligned}$$

where $s_{jk} = \sqrt{(\psi^2 - 1)I(i \in \mathscr{D}) + 1 + \phi_k^2 + \theta_j^2}$, and $\Phi(x)$ denotes the standard normal distribution function.

Prior assumptions on hyperparameters: We assume that the joint prior distribution on model parameters satisfies the following factorization:

$$\pi(\mu,\psi^2,\{\phi_k^2\},\{\theta_j^2\},\{\gamma_j\}) = \prod_{j=1}^{J}\left[\pi(\theta_j^2)\pi(\gamma_{j1},\gamma_{jC-1})\prod_{i=2}^{C-2}\pi(\gamma_{ji}\,|\,\gamma_{j1},\gamma_{jC-1})\right] \times \left[\prod_{k=1}^{K}\pi(\phi_k^2)\right]\pi(\mu)\pi(\psi^2)$$

There is, of course, some arbitrariness in defining the particular prior densities that appear in this factorization. The choices specified below are based on determining prior densities on variance parameters that yield simulated data that reflect the approximate inter- and intra-rater concordance typically observed in ROC studies when moderate values of the disease population mean are assumed. However, due to posterior sensitivity to the prior on the disease mean, we take a relatively vague prior for this parameter. Sensitivity to prior assumptions is discussed briefly in the 'Sensitivity Analysis' section.

The specific prior densities assumed for model hyperparameters are as follows. We assume an inverse gamma prior distribution with parameters 3 and 3 (i.e. IG(3,3)) for ψ^2. Under the parametrization of the inverse gamma distribution adopted here, this distribution has a mean of 1.5 and a mode of 0.75. This reflects a prior belief that variability in the disease population is typically larger than in the healthy population. Ninety per cent of the mass of this prior lies between 0.48 and 3.7 (equal tail areas). We place a proper, but vague, normal prior (mean 1, variance 10 000) on the disease population mean, μ. Inverse gamma prior distributions with parameters 2 and 1 (i.e. IG(2,1)) are assumed for both the rater variances $\{\theta_j^2\}$ and the modality variances $\{\phi_k^2\}$.

With this choice of priors for the reader and modality variances, propriety of the posterior distribution depends on the choice of prior for the reader thresholds. The prior densities specified for the thresholds in Ishwaran and Gatsonis [3]—uniform for the components of γ_{jc} on a finite interval (subject to an obvious order constraint)—is adequate to establish a proper posterior. However, simulation studies suggest that the marginal posterior distributions over γ_j, $\{\phi_k^2\}$ and $\{\theta_j^2\}$ are sensitive to the length of the intervals chosen to bound the rater thresholds. To reduce this posterior sensitivity, we assume a joint prior on the two extreme thresholds, γ_{j1} and γ_{jC-1}, for each j. Conditionally on these extreme thresholds, we assume a uniform prior on the remaining interior thresholds, $\pi(\gamma_{ji}\,|\,\gamma_{j1},\gamma_{jC-1}) = (\gamma_{jC-1}-\gamma_{j1})^{-1}$, for $i = 2,\ldots,C-2$, subject to the ordering constraint $\gamma_{j2} \leqslant \cdots \leqslant \gamma_{jC-2}$. We assume that the joint prior distribution on $(\gamma_{j1},\gamma_{jC-1})$ has the distribution of maximum and minimum values from a random sample of size $C-1$ from a normal distribution with mean 0.5 and variance 6.5. This implies a prior density on the rater thresholds of the form

$$\pi(\gamma_{j1},\gamma_{jC-1}) = (C-1)(C-2)\left[\Phi\left(\frac{\gamma_{jC-1}-0.5}{\sqrt{6.5}}\right)-\Phi\left(\frac{\gamma_{j1}-0.5}{\sqrt{6.5}}\right)\right]^{C-3} \times \frac{1}{\sqrt{6.5}}\,\phi\left(\frac{\gamma_{j1}-0.5}{\sqrt{6.5}}\right)\frac{1}{\sqrt{6.5}}\,\phi\left(\frac{\gamma_{jC-1}-0.5}{\sqrt{6.5}}\right)$$

Note that the mean of the normal distribution used to bound the extreme thresholds represents the average of the prior mean (1) for the diseased population and the healthy population mean (0). The variance of this normal distribution was obtained by summing the expected prior marginal variances for a rater scoring a healthy subject $E(1+\phi_k^2+\theta_j^2)=(1+1+1)=3$ and a diseased subject $E(\psi^2+\phi_k^2+\theta_j^2)=(1.5+1+1)=3.5$. This completes the specification of the model.

Connection to the bivariate–binormal model

We now compare the distributional assumptions implicit to our hierarchical Bayesian model for ROC data with the assumptions implicit to the classical bivariate–binormal model. To this end, we adopt the following simplified notation and assume that interest focuses on a comparison of only two diagnostic tests for one particular rater, say rater j. Let A_j be a two-by-two diagonal matrix with diagonal elements $(1+\phi_1^2+\theta_j^2)^{-1/2}$ and $(1+\phi_2^2+\theta_j^2)^{-1/2}$. By marginalizing over $\{Z_i\}$ and $\{Z_{ik}\}$ in (2)–(4) and applying the transformation of variables $(X_{ij1},X_{ij2})^{\mathrm{T}}=A_j(Z_{ij1},Z_{ij2})^{\mathrm{T}}$, we find that the marginal distribution for the latent traits observed under each modality for a healthy case can be expressed as

$$(X_{ij1},X_{ij2})^{\mathrm{T}}\sim \mathrm{N}((0,0)^{\mathrm{T}},\Sigma_h) \tag{6}$$

Similarly, the marginal distribution for the latent traits observed under each modality for a diseased case is

$$(X_{ij1},X_{ij2})^{\mathrm{T}}\sim \mathrm{N}((\mu_1,\mu_2)^{\mathrm{T}},\Sigma_d) \tag{7}$$

Conditionally on the observed values (y_{ij1},y_{ij2}), the latent trait distributions are truncated to the interval $(\gamma_{jy_{ij1}-1},\gamma_{jy_{ij1}}]A_j\times(\gamma_{jy_{ij2}-1},\gamma_{jy_{ij2}}]A_j$, where

$$\mu_1=\frac{\mu}{\sqrt{1+\phi_1^2+\theta_j^2}}\quad\text{and}\quad \mu_2=\frac{\mu}{\sqrt{1+\phi_2^2+\theta_j^2}}$$

and

$$\Sigma_h=\begin{pmatrix} 1 & \dfrac{1+\theta_j^2}{\sqrt{(1+\phi_1^2+\theta_j^2)(1+\phi_2^2+\theta_j^2)}} \\ \dfrac{1+\theta_j^2}{\sqrt{(1+\phi_1^2+\theta_j^2)(1+\phi_2^2+\theta_j^2)}} & 1 \end{pmatrix}$$

$$\Sigma_d=\begin{pmatrix} \dfrac{\psi^2+\phi_1^2+\theta_j^2}{1+\phi_1^2+\theta_j^2} & \dfrac{\psi^2+\theta_j^2}{\sqrt{(1+\phi_1^2+\theta_j^2)(1+\phi_2^2+\theta_j^2)}} \\ \dfrac{\psi^2+\theta_j^2}{\sqrt{(1+\phi_1^2+\theta_j^2)(1+\phi_2^2+\theta_j^2)}} & \dfrac{\psi^2+\phi_2^2+\theta_j^2}{1+\phi_2^2+\theta_j^2} \end{pmatrix}$$

Equations (6) and (7) reflect the distributional assumptions made for the latent variables in the standard bivariate–binormal model. The primary difference between the classical bivariate–binormal model and the model specified here involves implicit constraints made on the covariance parameters in the Bayesian model. In the standard bivariate–binormal model, the covariance matrix of the latent traits for the disease population between the two diagnostic tests is completely arbitrary and the correlation between diagnostic tests in the healthy population is unconstrained in the interval $[-1, 1]$. In contrast, the expressions above demonstrate that the marginal correlations in the Bayesian model are constrained to lie in $[0, 1]$. We feel that this constraint is natural, and so represents a feature of this model, rather than a drawback. Of course, the major difference between the two approaches is that the standard bivariate–binormal model must be applied to data from only one rater at a time, making it necessary to fit a second stage ME-ANOVA model. The Bayesian model has the advantage of making integration of data from more than one rater seamless. Apart from these differences, the distributional assumptions underlying the two approaches are quite similar.

A DATA ANALYSIS AND SIMULATION STUDY

In this section we perform a data analysis and examine the frequentist properties of our model. The particular parameters that we examine include A_Z values obtained for individual modalities and differences in A_Z values obtained from an ideal rater, as well as the relative values of rater variances. We also compare the coverage of posterior probability intervals to their nominal frequentist values, and compare the lengths of these intervals to the lengths of the corresponding confidence intervals generated using the DBM method. Finally, we compare the mean squared error (MSE) of estimates of A_Z values and differences of A_Z values computed from our Bayesian hierarchical model to the multirater method of DBM.

A general problem that arises in performing this type of simulation study involves the selection of appropriate study populations. Clearly, if we chose to simulate data according to our prior model, then the posterior properties of parameter estimates would be optimal and little would be learned concerning the relative performance of our model against alternative formulations. Alternatively, we could compare parameter estimates obtained from different models for actual ROC data, but this is also problematic since the baseline truth for the data is then not known. To overcome this difficulty, we decided to perform a simulation study where we identified a real ROC data set in which both models generated similar estimates of the difference in A_Z values on a rater-by-rater basis. After adding random noise to these data, we then used a resampling procedure to obtain smaller samples from this data set, and then compared A_Z estimates obtained under each model based on the sub sampled data to A_Z estimates obtained under the given model using the full data set.

Data analysis

An unpublished ROC study was conducted in 1993 in the UCLA Department of Radiology. The purpose of this study was to compare the diagnostic capabilities of chest film radiographs to digitized images displayed on a 1K by 1K video display. The basis of comparison were radiologists' ability to detect lung nodules in the radiographs. A panel of expert radiologists

determined 'truth' by consensus on 772 archived chest radiographs (59 cases with nodules and 713 disease free cases). Each of the 772 radiographs were then digitized for video display.

Three experienced radiologists and two radiology residents read each case under both systems (film *versus* video display) and ranked the presence of nodules on a scale of 1–5. Large ratings represented high confidence that nodules were present. We analysed this data set with both the BHM and DBM methods. The outcome of variable of interest was assumed to be the difference in A_Z values. For the complete data set, the estimated posterior mean difference in A_Z values (film–video display) under the BHM model was -0.0002 with a 95 per cent posterior probability interval $(-0.003, 0.002)$ (equal tail areas). Estimated median rater variance ratios are displayed in Figure 2 along with the 95 per cent probability intervals. Rater 4 had the smallest relative variance of any of the raters followed by raters 5, 1, 3 and 2, respectively. Both rater 4 and rater 5 had relative variances that can be considered substantively smaller than either rater 2 or 3. It is interesting to note that raters 2 and 3 were

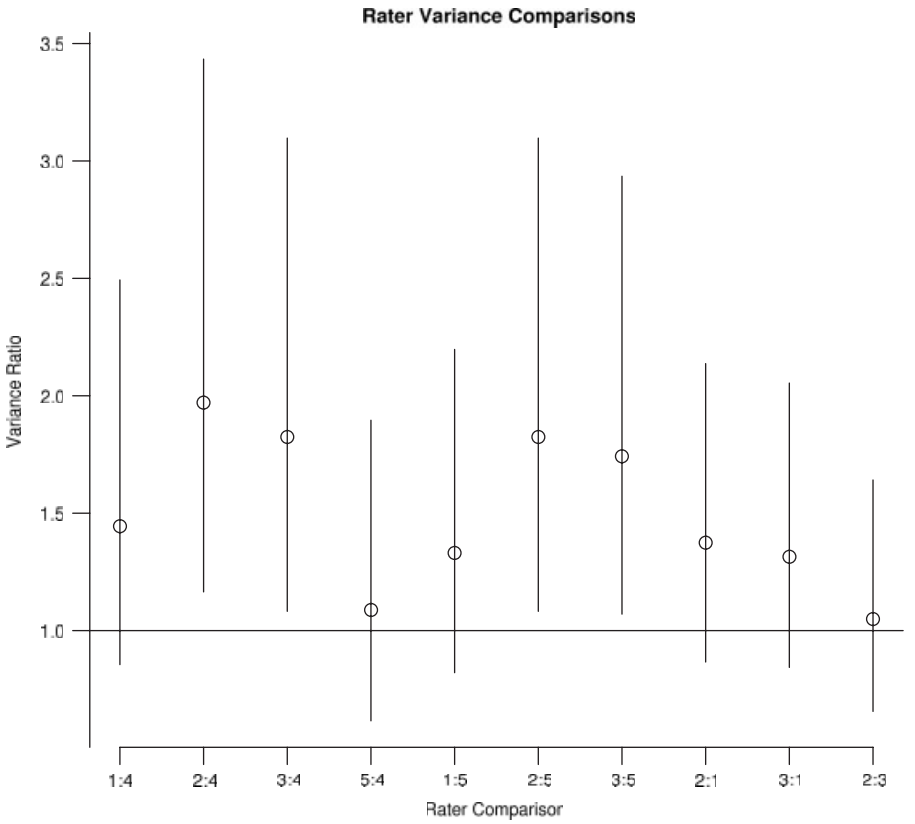

Figure 2. Ratios of rater variance (rater j *versus* rater i denoted $j : i$) can be used to determine the relative precision in which readers rate cases. Circles depict the median variance ratios, and vertical lines reflect the posterior 95 per cent Bayesian probability interval. The horizontal line at 1.0 indicates no difference in rater variances. Rater 4 had the greatest precision in rating cases (smallest variance).

radiology residents, both of whom had considerably less experience than the three senior radiologists.

When we attempted to apply the DBM method to these data, we experienced two numerical problems. First, the ratings obtained from one of the radiologists were 'degenerate'. That is to say, the likelihood-based estimation method failed due to the fact that this radiologists ratings could be separated perfectly into disease and healthy populations. Second, while jackknifing several of the cases for the other radiologists, the likelihood-based estimation methods failed to converge. As a consequence, we were unable to obtain reliable estimates of A_Z values for the full data set using the DBM method.

Resampling corrupted ROC data

Because of these convergence problems, a simulation study was performed using a contaminated version of these data. The contaminated version of the data was generated by randomly selecting 50 per cent of observations, and then perturbing the radiologist's ratings of these selected cases as follows: If a selected rating was between 2 and 4, it was changed by ± 1 unit with probability 0.2; if between 1 and 5 it was changed to 2 and 4 and with probability 0.1, respectively. In other words, if a selected case's rating by a radiologist was 2, then with probability 0.1 it was changed to a 1 and with probability 0.1 it was changed to a 3. For the extreme ratings of 1 and 5, with probability 0.1 the rating was changed by 1 unit toward the centre of the rating scale.

The contaminated data set was then analysed under each model, and the point estimates so obtained were assumed to represent the 'truth' for the corresponding modality. For the DBM method, the average rater A_Z values were 0.717 and 0.698 for film and video display, respectively. The difference in values was 0.019. For BHM, the ideal rater's A_Z values were 0.801 and 0.807 for film and video display, respectively. The difference in A_Z values was -0.006. In neither case were the differences in A_Z values significantly different from zero. Note that the ideal rater A_Z values under the BHM model were about 0.1 larger than the A_Z values obtained by the DBM method. This difference is due to the variability in reader ratings which is not included in the ideal rater A_Z values.

The contaminated data set was then repeatedly resampled with replacement. In each data set simulated in this way, 150 samples from the 'healthy population' were sampled and 50 samples from the 'diseased population' were sampled. A total of 1000 ROC data sets were generated in this way.

A summary of A_Z coverage rates, interval lengths and MSE values obtained from these 1000 simulated data sets appear in Table I. Coverage rates for individual modality A_Z values are slightly low for BHM. For this simulation study, the bias incurred by the BHM was, on average, -1.3 per cent for both film and video modalities, and this bias caused lower than nominal coverage rates. Conversely, individual modality coverage rates for the DBM method were high. For differences in A_Z values, the DBM model provides higher than nominal coverage; coverage for the difference in A_Z values is also higher than its nominal value for the BHM. Although the individual modality A_Z values are biased downward, these biases approximately cancelled when differences in A_Z values were estimated.

In this example/simulation study an important advantage is gained by explicitly modelling the correlation between ratings. By modelling this correlation in the BHM, posterior probability interval lengths (and variances) are smaller for the BHM than for the DBM. In fact, the BHM

Table I. A_Z 95 per cent Coverage rates, interval lengths and MSE under resampling from the contaminated data set.

	Coverage rates			Interval lengths			MSE		
	Film	Video	Diff.	Film	Video	Diff.	Film	Video	Diff.
Bayes	93.2	93.9	>99.9	0.176	0.178	0.035	0.002	0.002	0.00002
DBM	96.7	97.5	97.6	0.266	0.321	0.290	0.004	0.006	0.00390

Table II. Results from the sensitivity analysis to the prior specifications for modality and rater variances as well as the prior specification to the rater thresholds.

	Coverage rates			Interval lengths			MSE		
	Film	Video	Diff.	Film	Video	Diff.	Film	Video	Diff.
Orig. priors	93.2	93.9	> 99.9	0.176	0.178	0.035	0.002	0.002	0.00002
Scenario 1	93.8	93.8	> 99.9	0.175	0.176	0.036	0.002	0.002	0.00003
Scenario 2	93.9	93.7	> 99.9	0.176	0.178	0.036	0.002	0.002	0.00002

produces probability intervals that attain the correct nominal coverage but are, on average, about 8 times shorter than the corresponding intervals generated by DBM. Similar comments apply also to the MSE of estimated differences of A_Z values.

The average coverage of 95 per cent posterior probability intervals for ratios of rater variances was 91.3 per cent in repeated sampling in this simulation study.

Sensitivity analysis

To assess the sensitivity of our model to the choice of prior specifications for the rater and modality variances, we also re-ran the simulations above using two different hyperprior specifications. In the first scenario, we changed the prior densities on the rater and modality variances from IG$(2,1)$ densities to IG$(2,0.5)$ densities. This shifted the prior means of these parameters from 1.0 to 0.5, and reduced the variances by a factor of 4. In the second sensitivity analysis, we changed the priors on the rater and modality variances to IG$(2,2)$ densities; this effectively doubled the prior mean from 1 to 2 and increased the variance by a factor of 4. Furthermore, because the prior specification on the first and last category thresholds depend on the expectations of these priors, the prior specifications for these threshold values were also changed to reflect these differences. In particular, for the first scenario, the joint prior distribution on $(\gamma_{j1}, \gamma_{jC-1})$ has the joint distribution of the minimum and maximum values from a random sample of size $C-1$ from a normal distribution with mean 0.5 and variance 5.5 while those for the first scenario from a normal distribution with mean 0.5 and variance 8.5.

Posterior estimates obtained using these alternative prior specifications did not differ substantively from the simulation results stated previously. Results appear in Table II along with the results from our BHM model for ease of comparison. The average coverage of 95 per cent posterior probability intervals for ratios of rater variances were also similar: 91.3 and 91.5 per cent, respectively. We conclude that the posterior distributions defined in these studies was not sensitive to the choice of prior specification within this range of values.

DISCUSSION

The Bayesian hierarchical model described in this article provides a new approach towards analysing multirater correlated ROC data. The primary advantage of this model over existing methods is the dramatic decrease in the length of uncertainty intervals associated with differences in A_Z values, and corresponding decreases in the MSE of estimates of these differences. In our simulation studies, interval lengths and MSEs for differences in A_Z values were reduced by a factor of more than 5. Such gains in efficiency have important implications for study design and the power of ROC analyses for detecting differences in A_Z values.

Apart from increased efficiency, our model framework also provides reliable estimates of ratios of rater variances, and so offers the potential for providing feedback to readers regarding their precision in rating subjects relative to their peers. A similar potential also exists for improving the calibration of category thresholds across readers.

User friendly computer programs to implement the models described in this paper are available from the authors' website.

REFERENCES

1. Dodd LE, Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association* 2003; **98**(462):409–417.
2. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**:337–341.
3. Ishwaran H, Gatsonis CA. A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics* 2000; **28**(4):731–750.
4. Toledano AY, Gatsonis CA. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* 1996; **15**:1807–1826.
5. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* 1998; **54**:124–135.
6. Dorfman DD, Berbaum KS, Metz CE. Receive operating characteristic rating analysis: generalization to the population of readers and patients 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. *Communication in Statistics-Simulation and Computation* 1995; **24**(2):285–308.
8. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. *Academic Radiology* 2000; **7**:341–349.
9. Metz CE, Herman BA, Roe C. A new approach for testing the significant differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging*, Deconinck F (ed.). Nijihoff: The Hague, The Netherlands, 1984; 432–445.
10. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
11. DeLong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
13. Johnson VE, Albert JH. *Ordinal Data Modeling*. Springer: New York, 1999.
14. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
15. Swets JA, Pickett RM. *Evaluation of Diagnostic Systems*: *Methods from Signal Detection Theory*. Academic Press: New York, 1982.
16. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 1946; **2**:110–114.
17. Zhou X, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
18. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.

