A Bayesian Hierarchical Approach to Multirater Correlated ROC Analysis

Tim Johnson[*] Valen Johnson[†]

[*] University of Michigan Biostatistics, tdjtdj@umich.edu

[†] University of Michigan School of Public Health, valenj@umich.edu

A Bayesian Hierarchical Approach to Multirater Correlated ROC Analysis

Tim Johnson and Valen Johnson

Abstract

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 modeling 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 confidence intervals and MSE of estimates of differences in the area under the curves appear to be reduced by a factor that often exceeds two. Thus, our methodology has important implications for increasing the power of analyses based on ROC data collected from an available study population.

STATISTICS IN MEDICINE

Statist. Med. 2004; **00**:1–28 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

A Bayesian Hierarchical Approach to Multirater Correlated ROC Analysis

Timothy D. Johnson*, Valen E. Johnson

Department of Biostatistics
School of Public Health
University of Michigican
Ann Arbor, MI 48109-2029

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 modeling 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 confidence intervals and MSE of estimates of differences in the area under the curves appear to be reduced by a factor that often exceeds two. Thus, our methodology has important implications for increasing the

*Correspondence to: Timothy D. Johnson, Department of Biostatistics, Univ. of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029

Contract/grant sponsor: U.S. National Institute of Health; contract/grant number: 1PO1 CA 87684

power of analyses based on ROC data collected from an available study population.

INTRODUCTION

We propose a hierarchical latent variable model for analyzing multirater correlated ordinal receiver operating characteristics (ROC) data. 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., [1, 2, 3, 4, 5]). An important and efficient study design that has received less attention is one 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. For the analysis of data collected in designs where only one reader rates the outcomes, both parametric [6] and non-parametric methods [7, 8, 9] have been developed.

From a classical perspective, it is difficult to combine ROC data collected from several raters. This difficulty is caused by the absence of a model component for rater variability, although several methods have now been proposed to account for this source of variability (see, for example [10, 11, 12]). Each of these methods require two stages of modeling. In the first stage, estimates of the area under the ROC curve, commonly referred to as A_Z, (and jackknife pseudo-values of A_Z) for each rater are obtained. In the second stage, these estimates are used as observations in a mixed-effects analysis of variance model. Further discussion of these approaches may be found in Zhou, Obuchowshi and McClish (ZOM) [13].

The problem of combining information across raters is more transparent from the Bayesian perspective, and several Bayesian approaches have now also been explored. Among the earlier efforts in this direction are those detailed in Ishwaran and Gatsonis [3] and Johnson and Albert [14]. In this article, we describe an hierarchical latent variable model for analyzing multirater correlated ordinal ROC data that combines modeling aspects from each of these approaches and others currently being developed. 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 in 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 modeling 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 the 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, arguably, the most widely used ROC model for the analysis of mulitrater correlated data, that of Dorfman, Berbaum and Metz [10] (henceforth referred to as DBM). In Section we present a Bayesian hierarchical model for the analysis of multirater correlated ROC data and highlight its connection to the standard bivariate-binormal model (e.g., [15]). We compare performance of our model with that of DBM in Section . Finally, we illustrate our model in the analysis of a radiological data set intended to compare lung nodule detection using film versus a 1K display in Section . We conclude the manuscript with a short discussion.

THE JACKKNIFE METHOD OF DBM

Arguably the most widely used method to analyze mulitrater correlated ROC data is that due to DBM [10]. In this approach, reader ratings are jackknifed [16] to obtain pseudo-values for the A_Z for each case and modality one rater at a time. The pseudo-values are assumed to behave as independent observations, and are subsequently entered into standard ME-ANOVA software to fit a model of the form

$$\hat{A}_{ijk} = \mu + \alpha_i + B_j + C_k + (\alpha B)_{ik} + (\alpha C)_{ik} + (BC)_{jk} + (\alpha BC)_{ijk} + \varepsilon_{ijk}. \quad (1)$$

In this equation, $\hat{A}_{ijk}$ represents the A_Z pseudo-value for test or modality i, reader j and case k. We note that when each reader rates each image under each test only once (a common study design), the terms $(\alpha BC)_{ijk}$ and ε_{ijk} are inseparable. In DBM, the overall mean μ and the test effects α_i are fixed with $\sum_i \alpha_i = 0$. The reader effects, B_j, case effects, C_k, interaction terms and model error 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 [17]. 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 but the rater-by-case interaction terms.

ZOM summarize three shortcomings of this approach which we paraphrase here. First, 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 [16]. Second, pseudo-values are, in general, correlated. This means that the ME-ANOVA assumption of independent observations is violated. Third, this method applies the one-sample jackknife to a two sample problem (diseased and healthy cases). Finally, we note that the A_Z is supported on the interval $[0, 1]$, but observed pseudo-values often take values outside this interval. In practice, then, there are important differences between pseudo-values for A_Z and independent A_Z observations. Despite these theoretical shortcomings, this method enjoys much success in the radiological sciences.

Subsequent to the work of DBM, two other maximum-likelihood-based approaches have been developed [11, 12]. Both approaches follow along the same lines as that of DBM. At the first stage of modeling, A_Z values are computed one rater at a time. At the second stage of modeling, the A_Z values are combined across raters and modalities using a ME-ANOVA model. Both models have been compared to the DBM model with results comparable to that of DBM [11, 12] and so are not presently compared. A short summary and critique of each is provided in ZOM.

A BAYESIAN HIERARCHICAL MODEL

For the remainder of the paper, we assume that a panel of readers have assessed the disease status of a population of controls and cases to produce ordinal ratings based on two or more diagnostic tests (modalities). In performing our analysis, we assume that we know the true disease status of all subjects. The model proposed here is closely related to a simpler model described in Johnson and Albert [14], and may be considered approximately as a special case of the model proposed in Ishwaran and Gatsonsis [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 (the case in which we are interested). We do, however, extend the Ishwaran and Gatsonis model by allowing for distinct rater thresholds for each rater.

Suppose then there are N_h healthy cases and N_d disease cases for total of $N = N_h + N_d$ subjects. Let $\mathcal{D}$ denote the set of subjects classified as diseased and let $\mathcal{H}$ denote the set of subjects classified as being 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 diagnostic test. Let $K > 1$ denote the number of tests or diagnostic measures available to readers for evaluating each case, and suppose, for simplicity, 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 is diseased. We assume the latent variable representation for the data $\mathbf{Y} = \{Y_{ijk}\}$ detailed in Johnson and Albert. Under this representation, the ordinal ratings of each case by each reader are hypothesized to result from the (possibly distorted) 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 Swets and Pickett [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 Pepe [18], who provides a proof that there exists a monotone transformation of the continuous data to make the distributions of the healthy group and that of the diseased group normal.

In the first level of our model, we assume that the latent (disease) trait for the ith subject, denoted by Z_i, follows a normal distribution. We assume that the latent value for healthy cases is marginally distributed as a $N(0,1)$ random variable, while the latent value for a diseased individual is distributed as a $N(\mu, \psi^2)$ random variable. We treat the parameters μ and ψ^2 as unknown parameters and estimate them from data (see Figure 1, panel A). At the second level of the hierarchy, we assume that a $N(0, \phi_k^2)$ error is added to each latent disease trait. This error 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 (Figure 1, panel B). The parameter Z_{ik} denotes 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. At the final stage of the model hierarchy, we assume that the value of Z_{ik} is further distorted by the addition of a $N(0, \theta_j^2)$ random variable that represents error attributable to the jth reader's observation of Z_{ik}. The parameter θ_j^2 denotes the error variance particular to the jth reader. The sum of Z_{ik} and the reader error for the jth reader is denoted by Z_{ijk} (Figure 1, panel C).

In assigning cases to categories, we assume that each reader uses a unique set of thresholds $\boldsymbol{\gamma}_j$ having components that 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 $(j-1)$st and jth threshold. That is, $Y_{ijk} = c$ if and only if $\gamma_{jc-1} < Z_{ijk} \le \gamma_{jc}$.

To summarize the hierarchical model as specified thus far, we have

$$Z_i \overset{iid}{\sim} N(0,1),\ \forall\, i \in \mathcal{H} \quad \text{and} \quad Z_i \mid \mu, \psi^2 \overset{iid}{\sim} N(\mu, \psi^2),\ \forall\, i \in \mathcal{D} \qquad \text{Level 1} \qquad (2)$$

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

$$p(Z_{ijk} \mid Z_{ik}, \theta_j^2) = (\mathcal{K}/\theta_j) \exp\left[-0.5(Z_{ijk} - Z_{ik})/\theta_j)^2\right] \times \qquad \text{Level 3} \qquad (4)$$

$$\mathrm{I}(\gamma_{jy_{ijk}-1} < Z_{ijk} \le \gamma_{jy_{ijk}})$$

The right hand side of (4) is a truncated normal density with normalizing constant $\mathcal{K}$ and $\mathrm{I}(A)$ is the indicator function with $\mathrm{I}(A) = 1$ if A is true and is equal to zero otherwise.

Heuristically, this model may be interpreted as follows. First, latent disease traits for individuals in the population have inherent variability, and the magnitude of this variability is different among diseased and non-diseased individuals. Without loss of generality, we assume the variance of the latent traits among the non-diseased population is one, and among the diseased population is ψ^2. Second, measurements made by a diagnostic test or modality introduce errors in the observation of a case's latent disease trait. The variance of the error for the kth diagnostic test is denoted by ϕ_k^2. Finally, readers' interpretations of diagnostic tests are subject to error, and we allow for the possibility that different readers may have different expertise. The variance of the jth reader's error contribution is denoted by θ_j^2.

To specify prior constraints on parameters appearing in (2)-(4), we adopt the following prior factorization:

$$\left(\prod_{j=1}^{J} \pi(\boldsymbol{\gamma}_j)\pi(\theta_j^2)\right)\left(\prod_{k=1}^{K} \pi(\phi_k^2 \mid \theta_1^2, \ldots, \theta_J^2)\right)\ \pi(\mu)\,\pi(\psi^2).$$

For ψ^2, we take an inverse gamma prior distribution with parameters 3 and 3 (i.e., $IG(3,3)$). 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 constraint that variability in the disease population is typically larger than in the healthy population. Ninety percent of the mass of this prior lies between 0.48 and 3.7 (equal tail areas). We place an improper uniform prior on the disease population mean, μ.

Some care must be exercised in specifying the priors for modality and reader variances. Because the scale of the "observed" latent variables Z_{ijk} is not well defined from the priors specified on the distribution of the latent disease traits Z_i, application of non-informative priors on both the components of $\{\phi_k^2\}$ and $\{\theta_j^2\}$ can result either in a rater or modality variance collapsing to zero—resulting in a marginal posterior density that becomes highly peaked around zero—or variances that becomes arbitrarily large as category thresholds and all "observed" latent traits become large simultaneously.

These problems can be avoided by using uniform shrinkage priors for the modality variances. The Uniform shrinkage prior–first proposed by Strawderman [19]–represent a relatively vague but proper prior specification. These priors derive their name from the fact that they place a uniform distribution on the shrinkage of a posterior mean toward a prior mean in simple linear models. More recently, they were investigated by Daniels [20] and Natarajan and Kass [21]. Daniels showed they possess favorable frequentist properties, particularly when compared to other commonly used non-informative priors for variance components. To define the uniform priors specified here, let H_θ denote the harmonic mean of the θ_j^2: $H_\theta = J / \sum_{j=1}^{J} \theta_j^{-2}$ and consider the conditional posterior expectation of Z_{ik}:

$$E(Z_{ik} \mid Z_i, \phi_k^2, Z_{ijk}, \theta_j^2, j = 1, \ldots, J) = \frac{H_\theta \phi_k^2}{J\phi_k^2 + H_\theta} \sum_{j=1}^{J} Z_{ijk}/\theta_j^2 + \frac{H_\theta}{J\phi_k^2 + H_\theta} Z_i.$$

The uniform shrinkage prior for ϕ_k^2 can be induced by placing a uniform prior on the shrinkage parameter $H_\theta/(J\phi_k^2 + H_\theta)$ and applying the appropriate transformation of variables. Thus

$$\pi(\phi_k^2 \mid \theta_1^2, \ldots, \theta_J^2) = H_\theta/(J\phi_k^2 + H_\theta)^2, \quad k = 1, \ldots, K. \tag{5}$$

In contrast to the improper priors mentioned above, the uniform shrinkage prior puts arbitrarily small mass in neighborhoods of zero, thus avoiding a collapse of one or more of the marginal posterior distributions on the variance parameters to zero. In conjunction with this uniform shrinkage prior on the components of $\{\phi_j^2\}$, we take improper uniform priors on the rater variances $\{\theta_j^2\}$.

With this choice of priors for the reader and modality variances, propriety of the posterior distribution depends on the choice of prior density for the reader thresholds. The prior densities specified for the thresholds in Ishwaran and Gatsonis [3], which are 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, the posterior distribution on both the components of $\boldsymbol{\gamma}_j$ and the variance components $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are sensitive to the length of the interval chosen. To overcome the sensitivity in the posterior to the choice of interval length, we induce a prior density for the category thresholds by assigning probabilities to the event that a randomly selected subject would be categorized into category c by an ideal rater (zero variance) under an ideal modality (zero variance). The prior density induced in this way has support over the real line. To define the specific form of this prior, let

$$F(x) = \frac{N_h}{N}\Phi(x; 0, 1) + \frac{N_d}{N}\Phi(x; 1.07, 1.5)$$

where $\Phi(x, \mu, \sigma^2)$ is the cdf of a normal distribution with mean μ and variance σ^2. Here, the disease group variance has been set to its prior mean and the disease group mean has been set to 1.07, which lead to an ideal A_Z of about 0.75. This value is midway between the minimum informative value of 0.5 and the maximum of 1.0. Define the probability that a subject is placed in category c under this idealized rating scheme by $p_{jc} = F(\gamma_{jc}) - F(\gamma_{jc-1})$, for $c = 1, \ldots, C$. Then our prior density for the category thresholds is obtained by placing a Dirichlet prior on the $\mathbf{p}_j = \{p_{j1}, \ldots, p_{jC}\}$ and transforming to the $\boldsymbol{\gamma}$ scale to obtain

$$\begin{aligned} \pi(\boldsymbol{\gamma}_j) \propto \mathrm{J} \prod_{c=1}^{C} \{ (N_h/N) \left[\Phi(\gamma_{jc}; 0, 1) - \Phi(\gamma_{jc-1}; 0, 1)\right] + \\ (N_d/N) \left[\Phi(\gamma_{jc}; 1.07, 1.5) - \Phi(\gamma_{jc-1}; 1.07, 1.5)\right] \}^{(2-1)} . \end{aligned} \tag{6}$$

Here, J is the Jacobian of the transformation. The parameters of the Dirichlet prior effectively place 2 observations, a priori, in each category for each reader. This prior prevents the escape of the reader thresholds to ±∞, but imposes only weak information about the values of the reader thresholds. This completes the specification of the model.

CONNECTION TO THE BIVARIATE-BINORMAL MODEL

We now compare the distributional assumptions implicit in 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. Let A_j be a two-by-two diagonal matrix with diagonal elements $(1+\phi_1^2+\theta_j^2)^{-1/2}$ and $(1+\phi_2^2+\theta_j^2)^{-1/2}$. If we marginalize over $\{Z_i\}$ and $\{Z_{ik}\}$ in (2), (3), and (4) and apply the transformation of variables $(X_{ij1}, X_{ij2})^T = A_j(Z_{ij1}, Z_{ij2})^T$, we find that the marginal distribution for the latent traits observed under each modality for a healthy case under our Bayesian hierarchical model can be expressed as

$$(X_{ij1}, X_{ij2})^T \sim \Phi_2((0,0)^T, \Sigma_h), \tag{7}$$

where Φ_2 denotes a bivariate normal distribution. Similarly, the marginal distribution for the latent traits observed under each modality of a diseased case under the Bayesian model can be expressed

$$(X_{ij1}, X_{ij2})^T \sim \Phi_2((\mu_1, \mu_2)^T, \Sigma_d). \tag{8}$$

Conditionally on the observed values, y_{ij}, of Y_{ij} the latent trait distributions are truncated to the 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

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

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

Equations (7) and (8) reflect the distributional assumptions made for the latent variables in the standard bivariate-binormal model. However, 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. So is the correlation between the two tests in the healthy population. In the hierarchical Bayesian model, the equations for the marginal covariance matrices given above imply that the correlations between traits is forced to be positive. We feel that this is a reasonable constraint to impose on the covariance between latent traits observed for the same subject, making this a feature rather than a drawback of the model. Furthermore, the standard bivariate-binormal model can only be applied to data from one rater at a time, thus the need for a second stage ME-ANOVA model in the commonly used likelihood-based approaches previously mentioned.

MCMC DETAILS

The complexity of the joint posterior distribution on model parameters precludes the analytical study of posterior expectations. For that reason, we rely on Markov chain Monte Carlo (MCMC) methodology to generate samples from the posterior distribution and base model inferences on these sampled values.

After initializing model parameters, the particular steps in the MCMC scheme we propose may be described as follows.

1. Sample $\psi^2 \sim IG(3 + 0.5N_d, 3 + 0.5\sum_{i\in\mathcal{D}}(Z_i - \mu)^2)$.
2. Sample $\mu \sim N(\sum_{i\in\mathcal{D}} Z_i/N_d, \psi^2/N_d)$.
3. For $i \in \mathcal{H}$, sample $Z_i \sim N(m_h, v_h^2)$ where $m_h = v_h^2 \sum_k Z_{ik}\phi_k^{-2}$ and $v_h^2 = (1 + \sum \phi_k^{-2})^{-1}$.
4. For $i \in \mathcal{D}$ sample $Z_i \sim N(m_d, v_d^2)$ where $m_d = v_d^2(\mu\psi^{-2} + \sum_k Z_{ik}\phi^{-2})$ and $v_d^2 = (\psi^{-2} + \sum_k \phi_k^{-2})^{-1}$.
5. For $k = 1, \ldots, K$ draw a candidate value $\phi_{k*}^2 \sim IG(-1 + N/2, 0.5\sum_{i=1}^{N}(Z_{ik} - Z_i)^2)$. Accept ϕ_{k*}^2 with probability $\min\left[1, (H_\theta + J\phi_k^2)^2/(H_\theta + J\phi_{k*}^2)^2\right]$.
6. For $i = 1, \ldots, N$ and $k = 1, \ldots, K$, sample $Z_{ik} \sim N(m_{ik}, v_{ik}^2)$ where $m_{ik} = v_{ik}^2(Z_i\phi_k^{-2} + \sum_j Z_{ijk}\theta_j^{-2})$ and $v_{ik}^2 = (\phi_k^{-2} + \sum_j \theta_j^{-2})^{-1}$.
7. Draw a candidate value $\theta_{j*}^2 \sim IG(-1 + NK/2, 0.5\sum_{i=1}^{N}\sum_{k=1}^{K}(Z_{ijk} - Z_{ik})^2)$ for $j = 1, \ldots, J$. Accept θ_{j*}^2 with probability $\min\left[1, (H_\theta + J\phi_k^2)^2/(H_{\theta*} + J\phi_k^2)^2\right]$.
8. For $j = 1, \ldots, J$ and for $c = 1, \ldots, C-1$ draw γ_{jc*} from a truncated normal distribution with mean γ_{jc} and variance V, truncated to $(\gamma_{jc-1*}, \gamma_{jc+1})$ where $\gamma_{j0*} \equiv -\infty$ and

$\gamma_{jC*} \equiv \infty$. Let $\boldsymbol{\gamma}_{j*} = (\gamma_{j1*}, \ldots, \gamma_{jC*})$. Accept $\boldsymbol{\gamma}_{j*}$ as the new value of $\boldsymbol{\gamma}_j$ with probability

$$\min\left(1, \frac{\pi(\boldsymbol{\gamma}_{j*} \mid \mu, \psi^2, \theta_j^2, \phi_1^2, \ldots, \phi_K^2)}{\pi(\boldsymbol{\gamma}_j \mid \mu, \psi^2, \theta_j^2, \phi_1^2, \ldots, \phi_K^2)} \prod_{i=1}^{N} \prod_{k=1}^{K} \frac{\Phi(\gamma_{jy_{ijk}*}; Z_{ik}, V) - \Phi(\gamma_{jy_{ijk}-1*}; Z_{ik}, V)}{\Phi(\gamma_{jy_{ijk}}; Z_{ik}, V) - \Phi(\gamma_{jy_{ijk}-1}; Z_{ik}, V)} \right.$$
$$\left. \times \prod_{c=1}^{C-1} \frac{\Phi(\gamma_{jc+1}; \gamma_{jc}, V) - \Phi(\gamma_{jc-1*}; \gamma_{jc}, V)}{\Phi(\gamma_{jc+1*}; \gamma_{jc*}, V) - \Phi(\gamma_{jc-1}; \gamma_{jc*}, V)} \right).$$

V was chosen so that the acceptance rate was approximately 35%. This Metropolis-Hastings strategy for updating the category thresholds was proposed by Cowles [22].

9. For all i, j and k, draw $Z_{ijk} \sim N(Z_{ik}, \theta_j^2)$ truncated to the interval $(\gamma_{jy_{ijk}-1}, \gamma_{jy_{ijk}}]$.

Samples from the posterior distribution on the model parameters can be used to obtain posterior samples of A_Z values as follows. Let W_{jk} denote the latent variable of a randomly chosen healthy individual from rater j under condition k and let U_{jk} denote the latent variable of a randomly chosen diseased individual. Then, $\mathrm{A}_{\mathrm{Z}jk} = \Pr(W_{jk} < U_{jk})$ yields a sample from the posterior distribution of the A_Z for rater j under condition k [7]. From the model assumptions, it follows that $\mathrm{A}_{\mathrm{Z}jk} = 1 - \Phi(0; \mu, 1 + \psi^2 + \phi_k^2 + \theta_j^2)$. Similarly, we define the A_Z for an ideal rater as the A_Z for a rater with zero variance. The A_Z for such a rater is denoted by $\mathrm{A}_{\mathrm{Z}k}$ and is equal to $1 - \Phi(0; \mu, 1 + \psi^2 + \phi_k^2)$.

SIMULATION STUDIES

In this section we examine the frequentist properties of our model. The particular parameters that we examine include A_Zs obtained for individual modalities and differences in A_Zs obtained from an ideal rater, as well as ratios of rater variances (note that absolute magnitudes of rater variances are only defined relative to the prior densities assumed for the latent traits). We also compare the coverage of posterior probability intervals to their nominal values, and compare the lengths of these intervals to the lengths of the corresponding intervals generated using the

DBM models. Finally, we examine the mean squared error (MSE) of the A_Z and differences of A_Zs computed from our Bayesian hierarchical model and the multirater method of DBM.

A difficulty that arose in performing this simulation study involved the selection of study populations. Clearly, if we had chosen 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 to alternative formulations. Alternatively, we might have compared parameter estimates obtained from different models for the same data, but this is also problematic since the baseline truth for the data is then not known. Because of these difficulties, we decided to perform two simulation studies. In the first, we drew model parameters from prior densities that differed markedly from the prior densities assumed in our formulation, and then used these parameters to generate ROC data. These data were then analyzed using our model and the model described by DBM. In the second simulation, we identified a real ROC data set in which both models generated similar estimates of the difference in A_Z values. 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 subsampled data to the A_Z estimates obtained under that model using the full data set.

SYNTHETIC DATA

In this simulation, we repeatedly generated ROC data from a hierarchical model that had a structure similar to our Bayesian hierarchical model, but which used different priors on model parameters. In particular, the data generating model used values of μ drawn from a $U(0, 3)$ distribution (an improper prior was used in the estimation model), values of ψ^2 drawn from a $U(0.5, 3)$ distribution (an IG(3,3) distribution is used in the estimation model), and modality variances ϕ_k^2 and rater variances θ_j^2 drawn independently from a $U(0.1, 3)$ distribution (uniform shrinkage and improper prior densities were assumed for ϕ_k^2 and θ_j^2, respectively, in the estimation model). Figure 2 gives a visual comparison of the model priors and the priors used to generate the synthetic data.

We simulated 1000 data sets, each with parameters drawn from the distributions just described. Each data set consisted of 100 healthy and 100 diseased cases, each rated by four readers based on observations generated from two diagnostic tests. A 5-point rating scale was used. For simplicity, category thresholds were drawn from their actual prior distribution. After generating model parameters, latent variables Z_i, Z_{ik} and Z_{ijk} were generated according to (2)-(4). Ordinal data values Y_{ijk} were then determined according to the rule $Y_{ijk} = c$ if and only if $Z_{ijk} \in (\gamma_{jc-1}, \gamma_{jc}]$. Posterior distributions were estimated from 150,000 iterations after a burn-in of 100,000 iterations. One MCMC simulation took approximately 17.5 minutes on a PowerMac G4, 1.42 GHz processor.

A_Z coverage rate, interval lengths and MSE from our model and those from DBM are presented in Table I. A_Z statistics for the hierarchical Bayesian model correspond to the A_Z for an ideal rater.

For the study of a single modality, from Table I we see that the DBM model provides more accurate coverage of A_Z values than the Bayesian model. Coverage of the Bayesian model is slightly low. When considering only a single modality, the average length of the 95% confidence interval for the A_Z and the MSE of the estimate of the A_Z values were comparable for the DBM and Bayesian models.

In most ROC studies, the parameters of primary interest are differences in A_Z areas between modalities. For such differences, our Bayesian model provides important gains in efficiency. For example, the average length of the confidence interval for the difference in A_Z areas between two modalities obtained from the Bayesian model is only about 1/2 as wide as the corresponding confidence interval obtained using the DBM model. Similarly, the MSE error for the difference in A_Z values under the Bayesian model is only 1/3 as large as the MSE of the DBM models. This gain in precision is attributable to two factors. First, biases inherent in the Bayesian estimates of the A_Z areas largely cancel when differences in areas are examined. In this regard, it is important to note that the distribution of parameter values used to generate the simulated data differed substantially from the prior models used for estimation. Second, because the distribution of differences in A_Z values are computed with respect to the joint posterior distribution in the Bayesian model, the positive correlation between simulated values of the A_Z attributable to common draws of the rater variances reduces the variance of the distribution of the difference, much as it does in a paired t-test. The classical methods cannot exploit this covariance because it is not reflected in the marginal distribution of A_Z (or A_Z pseudo-values) values used in the follow-on ME-ANOVA analyses. Finally, it is interesting to note that the coverage rate for the difference in modality A_Z values is best for the Bayesian model.

Table I. A_Z 95% Coverage Rates, interval lengths and MSE. Comparison with the DBM model.

	Coverage Rates		Interval lengths		MSE	
	Modality	Difference	Modality	Difference	Modality	Difference
Bayes	91.6	94.7	0.113	0.064	0.00115	0.00029
DBM	96.0	93.6	0.132	0.139	0.00097	0.00118

Coverage rates for ratios in the values of rater variances obtained from within the Bayesian model were close to their nominal values. The coverage of 95% posterior probability intervals was 95.4 in repeated sampling in this simulation study.

RESAMPLING CORRUPTED ROC DATA

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 film displayed on a 1K by 1K video display. The basis of comparison were radiologists' abilty 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 vs. video display) and ranked the presence of nodules on a scale of 1 to 5. Large ratings represented high confidence that nodules were present. We analyzed this data with the Bayesian hierarchical model and the method proposed by DBM. The outcome of interest was the difference in A_Zs. For the complete data set, the estimated posterior mean difference in A_Zs (film - video display) under the Bayesian model was -0.0002 with a 95% posterior probability interval $(-0.003, 0.002)$ (equal tail areas). From this, we conclude that the A_Z values for the two methods are not substantively different.

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 was "degenerate". That is to say the likelihood method used to compute A_Zs and pseudo-values was unable to produce A_Zs for that rater because of the lack of variation in his/her ratings of images. Second, while jackknifing several of the cases for the other radiologists, the likelihood 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.

Because of these convergence problems, the second simulation study was carried out using a contaminated version of the data. Noise was added to the data set by randomly choosing half the cases, and for each selected case randomly changing all radiologist's rating of that case by 1 unit with probability .2 for ratings between 2 and 4, and with probability .1 for ratings of 1 and 5. For example, if a case's rating by a radiologist was 2, then with probability .1 it was changed to a 1 and with probability .1 it was changed to a 3. For the extreme ratings of 1 and 5, with probability .1 the rating was changed by 1 unit toward the center of the rating scale. The contaminated data set was then analyzed under each model, and the point estimates so obtained were subsequently assumed to represent the "truth" for the corresponding modality. For the likelihood-based models, the average rater A_Z for film was .717 and for the video display .698. The difference was .019. For the Bayesian model, the ideal rater A_Z for film was .801 and for video display .807; the difference in A_Z values was -0.006. The A_Z differences were not significantly different from zero under all models considered.

The contaminated data 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 1,000 ROC data sets were obtained in this way.

A summary of A_Z coverage rates, interval lengths and MSE values obtained from these 1,000 simulated data sets appear in Table II. Qualitatively, the results in Table II are similar to those reported in Table I. Coverage rates for individual modality A_Z values are more accurate under the DBM model than they are under the Bayesian model. For differences in A_Z values, the DBM model provides somewhat low coverage, while coverage for the difference in A_Z values is high for the Bayesian model. Interval lengths of the individual A_Z values and their difference are shortest for the Bayesian model. Both methods have comparable MSE values for individual modality A_Z values. However, the Bayesian model provides the smallest MSE for the difference in A_Z values. The MSE is nearly 7 times smaller than that provided by the DBM model.

Table II. A_Z 95% 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	91.6	91.5	99.5	0.186	0.188	0.041	0.0032	0.0033	0.00050
DBM	94.8	97.7	93.6	0.267	0.326	0.217	0.0036	0.0029	0.00330

DISCUSSION

The Bayesian hierarchical model described in this article provides a new approach towards analyzing multirater correlated ROC data. The primary advantage of this approach over existing methods is a marked decrease in the length of confidence intervals associated with differences in A_Z values and corresponding decreases in the MSE of these differences. In our simulation studies, confidence interval lengths were reduced by a factor of more than 2, while MSEs were reduced by a factor greater than 3. These finding have important implications for study design and the power of ROC analyses for detecting differences in A_Zs.

Aside from increases in the efficiency of the model, this framework provides reliable estimates of ratios of rater variances, and so has the potential for providing feedback to readers regarding their precision in rating subjects relative to their peers. A similar potential also exists for the model to help readers calibrate their category thresholds.

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 cnesored survival data and a 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. *The 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. Metz CE, Herman BA, Roe C. A new approach for testing the significant differences between ROC curves measured from correlated data. In F. Deconinck, editor, *Information processing in medical imaging*, pages 432–445, Nijihoff: The Hague, The Netherlands, 1984; pp 432–445.
7. 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.
8. 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.
9. Hanley JA, McNeil. BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
10. 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.

11. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. *Commun. Statist.—Simula.* 1995; **24**(2):285–308..
12. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple bootstrap experiements: An alternative method for random-effects, receiver operating characteristic analysis. *Academic Radiology* 2000; **7**:341–349.
13. Zhou X, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* John Wiley & Sons: New York, 2002.
14. Johnson VE, Albert, JH. *Ordinal Data Modeling.* Springer-Verlag: New York, 1999.
15. Swets JA, Pickett RM. *Evaluation of diagnostic systems: methods from signal detection theory.* Academic Press: New York, 1982.
16. Efron B,Tibshirani RJ. *An Introduction to the Bootstrap.* Chapman & Hall: New York, 1993.
17. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 1946; **2**:110–114.
18. Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press: Oxford, 2003.
19. Strawderman WE. Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics* 1971; **42**:385–388.
20. Daniels MJ. A prior for the variance in hierarchical models. *The Candadian Journal of Statistics* 1999; **27**(3):567–578.
21. Natarajan R, Kass RE. Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association* 2000; **95**(449):227–237.
22. Cowles MK. Accelerating monte carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing* 1996; **6**:101–111.

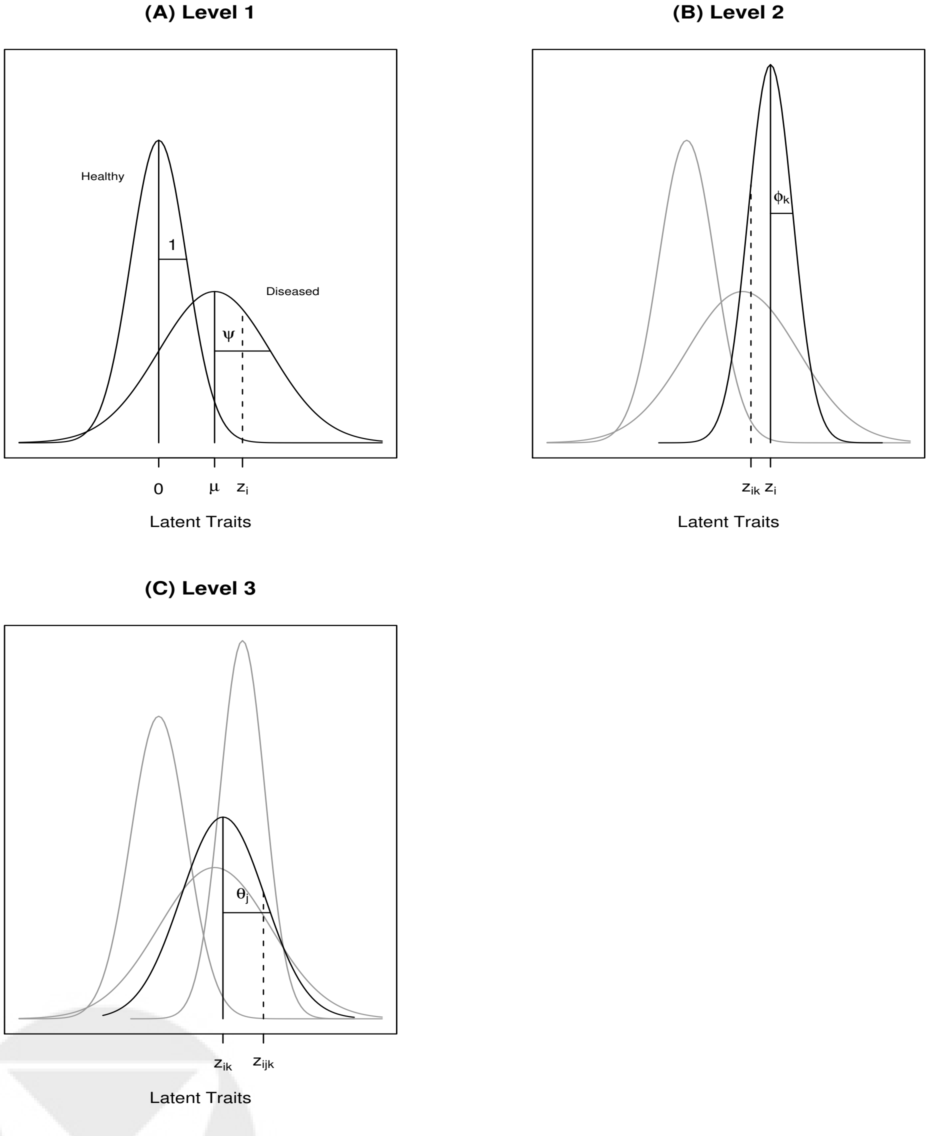

Figure 1. Graphical representation of our Bayesian hierarchical model. Panel A: The distributions of the latent case traits. Panel B: Modality k adds noise to the latent trait Z_i. Panel C: Rater j adds more noise to the system centered at Z_{ik}. The dashed vertical lines represent hypothetical perceived values.

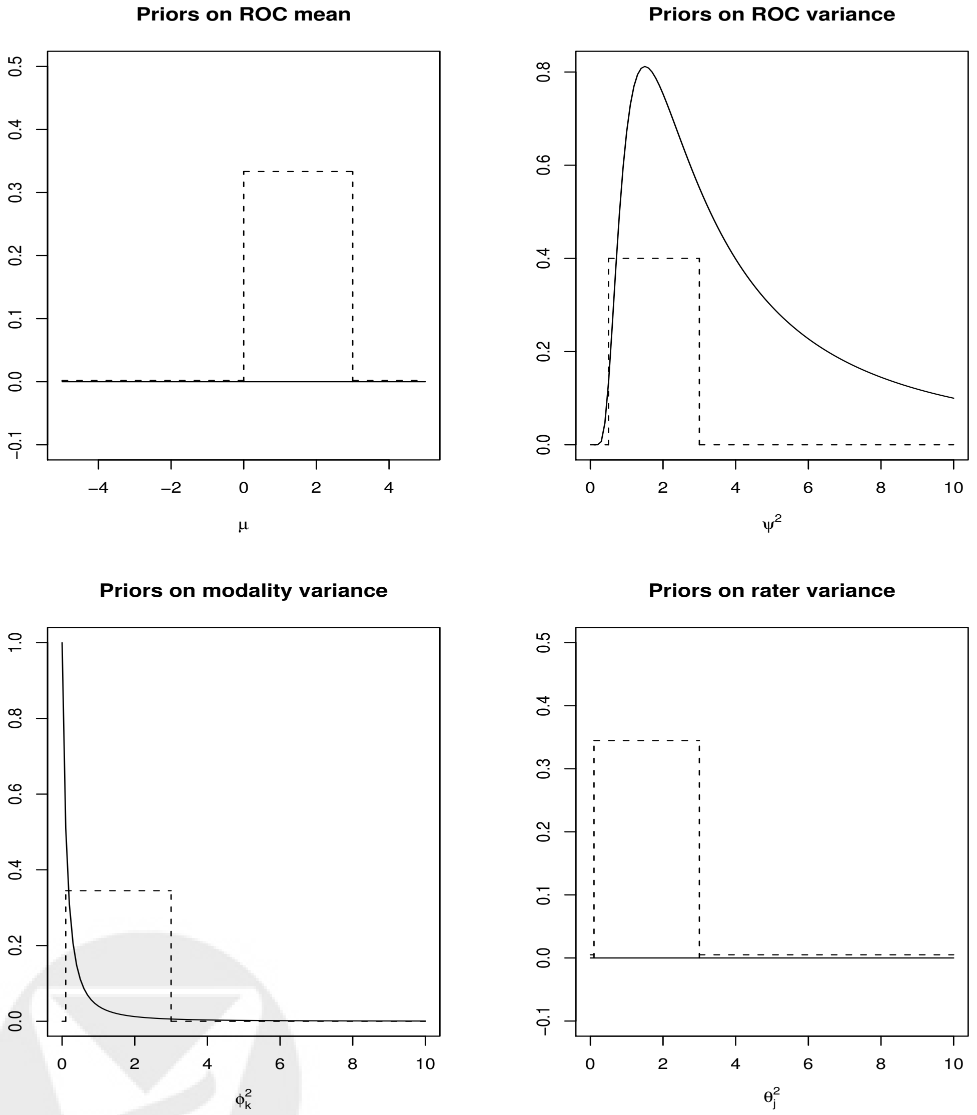

Figure 2. Graphical representation of the priors used for data generation (dashed lines) and the model priors (solid lines). UL: Priors for μ. UR: Priors for ψ^2. LL: Priors for ϕ_k^2. The uniform shrinkage prior assumes $J = 4$ and $\theta_j^2 = 1$, $j = 1, 2, 3, 4$ [see (5)]. LR: Priors for θ_j^2.