 DOI: 10.1111/j.1541-0420.2006.00712.x

Random Effects Modeling Approaches for Estimating ROC Curves from Repeated Ordinal Tests without a Gold Standard

Paul S. Albert

Biometric Research Branch, National Cancer Institute,
Bethesda, Maryland 20892, U.S.A.
email: albertp@mail.nih.gov

Summary. Estimating diagnostic accuracy without a gold standard is an important problem in medical testing. Although there is a fairly large literature on this problem for the case of repeated binary tests, there is substantially less work for the case of ordinal tests. A noted exception is the work by Zhou, Castelluccio, and Zhou (2005, *Biometrics* **61,** 600–609), which proposed a methodology for estimating receiver operating characteristic (ROC) curves without a gold standard from multiple ordinal tests. A key assumption in their work was that the test results are independent conditional on the true test result. I propose random effects modeling approaches that incorporate dependence between the ordinal tests, and I show through asymptotic results and simulations the importance of correctly accounting for the dependence between tests. These modeling approaches, along with the importance of accounting for the dependence between tests, are illustrated by analyzing the uterine cancer pathology data analyzed by Zhou et al. (2005).

Key words: Diagnostic accuracy; Latent class analysis; Mixture models; Random effects models for repeated ordinal data; ROC curves.

1. Introduction

The lack of gold standard diagnostic truth often complicates evaluation of diagnostic accuracy for new medical tests. In some cases, gold standard evaluation may be too costly to obtain, while in others, a method for establishing true disease status may not exist. Modeling diagnostic accuracy without a gold standard test remains an active area of biostatistics research. A majority of the work in this area involves estimating the diagnostic accuracy of binary tests (i.e., sensitivity and specificity) without a gold standard test. Albert and Dodd (2004) discussed a number of latent class modeling approaches for binary tests which express the joint distribution of test results conditional on the true disease status with different models. They showed that estimates of diagnostic accuracy may be sensitive to the choice of the model for the conditional joint distribution. Further, they showed that it is difficult to distinguish between these different models with a small number of binary tests. A natural question is whether there are similar inferential problems with ordinal tests as compared with binary tests. Because ordinal variables contain more information than binary variables, there is the possibility that we may be able to better distinguish between competing models for the dependence structure between tests with a small number of ordinal rather than binary tests. For ordinal tests, interest is in estimating receiver operating characteristic (ROC) curves rather than sensitivity and specificity. Specific questions that might be asked include the following: (i) are estimators of ROC curves robust to modeling assumptions about the dependence between tests, and (ii) can we distinguish between models for the dependence structure based on data from only a limited number of ordinal tests?

There has been limited work on modeling ROC curves without a gold standard from repeat continuous or ordinal tests. Henkelman, Kay, and Bronskill (1990) proposed a maximum likelihood method based on a multivariate normal mixture. Choi et al. (2006) proposed a Bayesian modeling approach for estimating ROC curves from two normally distributed tests that are assumed to be correlated. Zhou, Castelluccio, and Zhou (2005) proposed a nonparametric estimation procedure for estimating the ROC curve from repeat ordinal tests. This approach is very flexible in parameterizing the mean structure, but it assumes conditional independence between tests (i.e., tests are independent conditional on the gold standard test). This article proposes random effects models for estimating ROC curves from multiple ordinal tests on the same subject. These approaches incorporate dependence between tests through the random effects.

I begin with a random effects model which extends the model for binary tests proposed by Qu, Tan, and Kutner (1996) to ordinal test data, and I then discuss models where the random effects distribution is non-Gaussian. These models fit into a broader framework of random effects models for repeated ordinal data, whereby the probability of an ordinal response is modeled with fixed and random effects, which are additive on the cumulative probit or cumulative logit scale. For many of these models, the ordinal responses are characterized by thresholding a latent continuous variable (with cutpoints corresponding to ordinal categories), which is modeled

with a linear mixed model (Harville and Mee, 1984; Hedeker and Gibbons, 1994; Tutz and Hennevogl, 1996). More recent literature includes Bayesian approaches for increased numerical stability (Ishwaran, 2000), easier implementation (Qiu, Song, and Tan, 2002), and more general random effects distributions (Kottas, Muller, and Quintana, 2005). Ishwaran and Gatsonis (2000) presented a general class of ordinal random effects models with applications to correlated ROC analyses when true disease status is known. In this article, the focus is on developing approaches for correlated ROC analyses when disease status is unknown. Because the unknown disease status is treated as a latent variable, this article is related to a broad literature in latent class models for discrete data (e.g., Lazarsfeld and Henry, 1968; Goodman, 1974; Espeland and Handelman, 1989; Formann, 1992; Uebersax, 1999).

In Section 2, I develop random effects models for estimating ROC curves without a gold standard. In Section 3, I show that the asymptotic bias occurs if we assume conditional independence when in fact there is strong dependence between tests. More generally, it is shown that assuming the correct dependence between tests is important for estimating diagnostic accuracy. Fortunately, unlike with binary tests, one can more easily distinguish between competing models for the dependence between tests with a small number of ordinal tests. In Section 4, simulations demonstrating the importance of incorporating dependence between tests are presented. In Section 5, I illustrate these approaches with a study on assessing the accuracies of seven specific pathologists in detecting carcinoma in situ of the uterine cervix. These data were used by Zhou et al. (2005) for estimating rater-specific ROC curves using a nonparametric approach under a conditional independence assumption. Using the proposed models, I show the importance of correctly modeling the dependence structure for estimating the ROC curve without a gold standard in this application. A discussion follows in Section 6.

2. Random Effects Models

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be a vector of J ordinal test results for the ith subject ($i = 1, 2, 3, \ldots, I$), where Y_{ij} takes on the ordinal values 1 to K. The joint distribution can be expressed by marginalizing over the latent true disease status. Specifically, the marginal distribution of $\mathbf{Y}_i$ can be written as

$$P(\mathbf{Y}_i) = \sum_{l=0}^{1} P(\mathbf{Y}_i \mid d_i = l) P(d_i = l), \tag{1}$$

where d_i is the unknown disease status for the ith subject, $P(\mathbf{Y}_i \mid d_i)$ is the joint distribution of $\mathbf{Y}_i$ conditional on d_i, and $P(d_i = 1)$ is the probability of disease, which is commonly referred to as disease prevalence. The remaining part of this section is concerned with modeling $P(\mathbf{Y}_i \mid d_i)$.

In the next subsection, the situation in which a common ROC curve across all J tests is estimated is discussed.

2.1 *A Common ROC Curve across J Tests or Raters*

This structure is sensible when the test can be considered as replicate tests or ratings. For example, a single test is repeated J times. In this case, we can parameterize $P(\mathbf{Y}_i \mid d_i)$ as a cumulative probit random effects model, which exploits the ordinal nature of the data. Specifically, we can model the conditional distribution as

$$\Phi^{-1}\big\{P\big(Y_{ij} \le k \mid d_i, b_{d_i,i}\big)\big\} = C_{d_i,k} + b_{d_i,i}, \tag{2}$$

where $C_{d_i,k}$ are monotonically nondecreasing cutpoints (i.e., $-\infty = C_{0,0} \le C_{0,1} \le C_{0,2} \le \cdots \le C_{0,K-1} \le C_{0,K} = \infty$ and $-\infty = C_{1,0} \le C_{1,1} \le \cdots \le C_{1,K-1} \le C_{1,K} = \infty$ for $d_i = 0$ and 1, respectively) and $b_{d_i,i}$ is a random effect which characterizes the dependence in the conditional distribution $P(\mathbf{Y}_i \mid d_i)$. Further, $\Phi(x)$ and $\Phi^{-1}(x)$ denote the cumulative and inverse cumulative distributions of a standard normal. The cumulative probit link function has been discussed by McCullagh and Nelder (1989) and has commonly been used for regression modeling of univariate and multivariate ordinal data. Cumulative probit random effects models have a natural interpretation corresponding to the thresholding of a continuous underlying latent random variable which is modeled with a linear mixed model. The random effect $b_{d_i,i}$ varies by individual i, and its distribution depends on the true disease status (d_i). There are a number of choices for the random effects distribution: first, a Gaussian random effects (GRE) model can be developed whereby $b_{d_i,i} = \sigma_{d_i} b_i$, and where b_i has a standard normal distribution. Such a random effects model generalizes the approach of Qu et al. (1996) for analyzing binary test data to ordinal test data. For the GRE model, the conditional distribution of $\mathbf{Y}_i$ ($\mathbf{Y}_i \mid d_i$) can be expressed as

$$P(\mathbf{Y}_i \mid d_i) = \int \prod_{j=1}^{J} \big\{\Phi\big(C_{d_i,y_{ij}} + b\big) - \Phi\big(C_{d_i,y_{ij}-1} + b\big)\big\} \phi_{\sigma_{d_i}}(b)\, db, \tag{3}$$

where $\phi_\sigma(b)$ is a normal density with mean zero and variance σ^2. Equation (3) needs to be evaluated numerically. I have evaluated (3) using Gaussian quadrature with 50 Gaussian quadrature points (Abramowitz and Stegun, 1972). The conditional independence model is a special case of the GRE model when $\sigma_0 = \sigma_1 = 0$. In other words, equation (2) reduces to the conditional independence model when $b_{d_i,i} = 0$, for all d_i and i. The GRE model is similar to the random effects model for ordinal data proposed by Kottas et al. (2005), whereby the random effects distribution is specified as a finite mixture (FM) of normals. It is also closely related to the probit latent class model proposed by Uebersax (1999).

The FM model is an alternative approach for incorporating conditional dependence between tests, whereby dependence is incorporated with an FM as compared with a continuous mixture in the GRE model. Specifically, conditional on $d_i = 0$, a fraction η_0 of individuals will always be rated with the lowest rating (rating = 1 in our example), and the remainder with probability $1 - \eta_0$ are subject to equation (2) with a conditional independence structure. Similarly, conditional on $d_i = 1$, a fraction η_1 of individuals will always be rated with the highest rating (rating = 5 in our example), and the remainder with probability $1 - \eta_1$ will be subject to equation (2) with a conditional independence structure. The FM model generalizes an

FM model proposed for binary tests (Albert et al., 2001). For the FM model with five ordinal categories as in our example, the conditional distribution of $\mathbf{Y}_i$ given d_i can be expressed as follows. If $\mathbf{Y}_i = (1, 1, \ldots, 1)'$ and $d_i = 0$, then

$$P(\mathbf{Y}_i \mid d_i = 0) = \eta_0 + (1 - \eta_0) \prod_{j=1}^{J} \Phi(C_{0,1}). \tag{4}$$

If $\mathbf{Y}_i \neq (1, 1, \ldots, 1)'$ and $d_i = 0$,

$$P(\mathbf{Y}_i \mid d_i = 0) = (1 - \eta_0) \prod_{j=1}^{J} \left\{ \Phi\left(C_{0,y_{ij}}\right) - \Phi\left(C_{0,y_{ij}-1}\right) \right\}. \tag{5}$$

If $\mathbf{Y}_i = (5, 5, \ldots, 5)'$ and $d_i = 1$,

$$P(\mathbf{Y}_i \mid d_i = 1) = \eta_1 + (1 - \eta_1) \prod_{j=1}^{J} \{1 - \Phi(C_{1,4})\}. \tag{6}$$

Last, if $\mathbf{Y}_i \neq (5, 5, \ldots, 5)'$ and $d_i = 1$,

$$P(\mathbf{Y}_i \mid d_i = 1) = (1 - \eta_1) \prod_{j=1}^{J} \left\{ \Phi\left(C_{1,y_{ij}}\right) - \Phi\left(C_{1,y_{ij}-1}\right) \right\}. \tag{7}$$

Note that when η_0 or η_1 is positive, the FM model accounts for positive conditional dependence between tests. This follows by noting that $\mathrm{cov}(y_{ij}, y_{ik} \mid d_i = 1) = (K + C)^2 \eta_1 (1 - \eta_1)$, where $C = \sum_{l=1}^{K} l\{\Phi(C_{1,l}) - \Phi(C_{1,l-1})\}$.

A generalization of both the GRE and FM models is a combined model in which both types of heterogeneity are incorporated into the random effects distribution. Specifically, conditional on $d_i = 0$,

$$b_{0,i} = \begin{cases} +\infty & \text{with probability } \eta_0 \\ N(0, \sigma_0^2) & \text{with probability } 1 - \eta_0 \end{cases}, \tag{8}$$

where $N(0,\ \sigma^2)$ denotes a normal distribution with mean 0 and variance σ^2. Conditional on $d_i = 1$,

$$b_{1,i} = \begin{cases} -\infty & \text{with probability } \eta_1 \\ N(0, \sigma_1^2) & \text{with probability } 1 - \eta_1 \end{cases}. \tag{9}$$

The GRE model is a special case of the combined model (8) and (9) when $\eta_0 = \eta_1 = 0$. Further, the FM model results when $\sigma_1 = \sigma_0 = 0$. Primarily the GRE and FM models will be considered in this article.

The model parameters themselves are usually not of direct interest. Focus is on obtaining model-based estimates of the ROC curve. For a general random effects distribution, the coordinates of the ROC curve, corresponding to (1-Specificity, Sensitivity) at each of the ordinal categories ($k = 1, 2, 3, \ldots, K$), can be expressed as ($[1 - E\{P(Y_{ij} \leq k \mid d_i = 0)\}]$, $E\{P(Y_{ij} > k \mid d_i = 1)\}$), which can be evaluated as ($[1 - E\{\Phi(C_{0,k} + b_{0,i})\}]$, $[1 - E\{\Phi(C_{1,k} + b_{1,i})\}]$), where the expectation is taken over the random effects distribution. The area under the ROC curve (AUC), which is an important summary measure of the ROC curve, can be expressed as

$$AUC = \sum_{k=1}^{K-1} \left\{ P(Y_{ij} = k \mid d_i = 0) \sum_{l=k+1}^{K} P(Y_{ij} = l \mid d_i = 1) \right\} + \frac{1}{2} \sum_{k=1}^{K} P(Y_{ij} = k \mid d_i = 0) P(Y_{ij} = k \mid d_i = 1), \tag{10}$$

where

$$P(Y_{ij} = k \mid d_i) = E\left\{\Phi\left(C_{d_i,k} + b_{d_i,i}\right)\right\} - E\left\{\Phi\left(C_{d_i,k-1} + b_{d_i,i}\right)\right\}. \tag{11}$$

Under the more general combined GRE/FM model (8) and (9), $E\{\Phi(C_{d_i,k} + b_{b_i,i})\} = (1 - \eta_{d_i})\Phi(C_{d_i,k}/\sqrt{1 + \sigma_{d_i}^2}) + \eta_{d_i} I(d_i = 0)$, where $I(x)$ is an indicator function that is equal to 1 when the condition x is met and is otherwise equal to 0. Under a GRE assumption, $E\{\Phi(C_{d_i,k} + b_{d_i,i})\} = \Phi(C_{d_i,k}/\sqrt{1 + \sigma_{d_i}^2})$. Under an FM model, $E\{\Phi(C_{d_i,k} + b_{d_i,i})\} = (1 - \eta_{d_i})\Phi(C_{d_i,k}) + \eta_{d_i} I(d_i = 0)$. For conditional independence, equation (10) reduces to the expression for AUC developed by Zhou et al. (2005).

2.2 *Rater-Specific ROC Curves*

In this subsection, I generalize the random effects models to allow for different ROC curves for each of the J tests or raters. One parsimonious model would be to introduce rater-specific effects to the cumulative probit random effects model (2) as

$$\Phi^{-1}\left\{P\left(Y_{ij} \leq k \mid d_i, b_{d_i,i}\right)\right\} = C_{d_i,k} + \sum_{r=2}^{J} \beta_{d_i,r} I_{(r=j)} + b_{d_i,i}, \tag{12}$$

$k = 1, 2, 3, \ldots, K - 1$, where $I_{(x)}$ is an indicator function that is equal to 1 if the condition x is met. The regression coefficients $\beta_{d_i,r}$ ($\beta_{0,2}$, $\beta_{0,3}, \ldots, \beta_{0,J}$ and $\beta_{1,2}$, $\beta_{1,3}, \ldots, \beta_{1,J}$) allow for rater-specific shifts in the cutpoints depending on d_i for raters 2 to J relative to rater 1. Model (12) assumes that both the random effects and rater effects act on the cumulative probability of an ordinal response in an additive fashion on the probit scale. One could generalize model (12) so that we estimate the cutpoints separately for each rater. For this model,

$$\Phi^{-1}\left\{P\left(Y_{ij} \leq k \mid d_i, b_{d_i,i}\right)\right\} = C_{d_i,k,j} + b_{d_i,i}, \tag{13}$$

where for each rater the cutpoints are constrained to be monotonically nondecreasing (i.e., for each rater j, $-\infty = C_{0,0,j} \leq C_{0,1,j} \leq \cdots \leq C_{0,K-1,j} \leq C_{0,K,j} = \infty$ and $-\infty = C_{1,0,j} \leq C_{1,1,j} \leq \cdots \leq C_{1,K-1,j} \leq C_{1,K,j} = \infty$ for $d_i = 0$ and 1, respectively) and $b_{d_i,i}$ is a random effect as described previously. Zhou et al.'s (2005) nonparametric conditional independence approach is an alternative parameterization to (13) when there are no random effects. Thus, model (13) can be thought of as a generalization of Zhou et al.'s nonparametric approach, which allows for conditional dependence between tests.

2.3 *Estimation*

The log likelihood can be written as $\log L = \sum_{i=1}^{I} \log L_i$, where

$$\log L_i = \sum_{i_1=1}^{K} \sum_{i_2=1}^{K} \cdots \sum_{i_J=1}^{K} I_{\{\boldsymbol{Y}_i=(i_1,i_2,\ldots,i_J)\}} \times \log\{P(\boldsymbol{Y}_i = (i_1, i_2, \ldots, i_J))\}, \quad (14)$$

where $P(\boldsymbol{Y}_i)$ is given by equation (1). The log likelihood $\log L = \sum_{i=1}^{I} \log L_i$ can be maximized using a quasi-Newton–Rhaphson technique by using `GAUSS` (Aptech Systems, 1992). As with other latent class models (Zhou et al., 2005, for example), the likelihood (14) is invariant to a relabeling of the latent variable (e.g., $d = 0$ corresponding to a positive gold standard test). Thus, there are two solutions which maximize equation (14). Fortunately, in most cases, only one of the solutions provides reasonable prevalence and ROC curve estimates.

Although estimators of the standard errors for both estimates of the ROC curve and estimates of the AUC could be derived from variance estimates of the model parameters (as is suggested by Zhou et al., 2005) using the delta method, I propose using a bootstrap. The bootstrap (Efron and Tibshirani, 1993) is a flexible approach that avoids the laborious delta method calculations required for more complex models. Specifically, I constructed a bootstrap sample of I vectors of ordinal tests $(\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*, \ldots, \boldsymbol{Y}_I^*)$, where each vector is drawn with replacement from the original set of I vectors, $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_I$ (Moulton and Zeger, 1989). Standard errors for parameter estimates were estimated with 800 resampled datasets. The choice of 800 bootstrap datasets was based on the results and recommendations of Booth and Sarkar (1998), who showed that this number was sufficient to achieve a relative error in the variance estimate of less than 10% with probability 95%.

3. Asymptotic Bias for Misspecified Random Effects Distribution

I examine the asymptotic bias in misspecifying the dependence structure between tests. For simplicity, I will consider the situation in which all ROC curves are constant across four raters or tests ($J = 4$). The approach is similar to the one taken for assessing asymptotic bias for misspecified random effects distribution for repeated binary tests (Albert and Dodd, 2004). The misspecified maximum likelihood estimator for the model parameters, denoted as $\hat{\boldsymbol{\theta}}^*$, converges to the value $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} E_T\{\log L_M(\boldsymbol{Y}_i, \boldsymbol{\theta})\}, \quad (15)$$

and $\log L_M(\boldsymbol{Y}_i\boldsymbol{\theta})$ is the individual contribution to the log likelihood under the assumed model M and the expectation is taken with respect to the true model T. An expression for $E_T\{\log L_M(\boldsymbol{Y}_i, \boldsymbol{\theta})\}$ is provided in the Appendix. The notation

$$E_T(\log L_M) = E_T\{\log L(\boldsymbol{Y}_i, \boldsymbol{\theta})\}|_{\theta=\theta^*} \quad (16)$$

denotes the expectation (with respect to the true model T) of the individual contribution to the log likelihood under the assumed model M when evaluated at $\boldsymbol{\theta}^*$. By the invariance property of maximum likelihood, the maximum likelihood estimator of the AUC converges to AUC^* (as $I \to \infty$), where $AUC^* = g(\boldsymbol{\theta}^*)$ and g relates the parameters $\boldsymbol{\theta}$ to the AUC by (10).

Initially, I will examine the asymptotic bias of the AUC (defined as $AUC^* - AUC$) for a model which assumes conditional independence when the true model is a GRE model. Table 1 presents AUC^* for different model parameters in which A and B lead to an AUC of 0.9, while C and D lead to an AUC of 0.8. The parameter values are constructed so that A and C have the weaker dependence ($\sigma_0 = \sigma_1 = 1$), while B and D have stronger dependence ($\sigma_0 = \sigma_1 = 2$), between tests. Results are presented for true prevalences of 20% and 50%. The results show that ignoring the dependence between tests leads to asymptotically biased estimates of AUC.

I also examined the effect of misspecifying the dependence structure between tests on the AUC. Table 2 shows the asymptotic bias of the AUC for the GRE model when the true random effects distribution is a combined GRE/FM model (8) and (9). The results show that there can be sizable asymptotic bias as the random effects distribution departs from a normal distribution. For example, the first row shows a case in which only 10% of disease-free or diseased

Table 1

Asymptotic bias ($AUC^* - AUC$) in estimating the AUC under an independence model when the true model is a GRE model. These calculations assume that $J = 4$, $K = 5$, and there are no rater-specific differences in the ROC curves.

Parameter values[a]	True prevalence	True AUC	AUC*	Bias	$E_{\text{GRE}}\{\log L_{\text{GRE}}\}$	$E_{\text{GRE}}\{\log L_{\text{Ind}}\}$
A. with $\sigma_0 = \sigma_1 = 1$	0.2	0.90	0.933	0.033	−3.1829	−3.3714
	0.5	0.90	0.932	0.032	−4.0387	−4.2053
B. with $\sigma_0 = \sigma_1 = 2$	0.2	0.90	0.960	0.060	−2.4721	−2.8142
	0.5	0.90	0.968	0.068	−3.1708	−3.5052
C. with $\sigma_0 = \sigma_1 = 1$	0.2	0.80	0.862	0.062	−3.8941	−3.9661
	0.5	0.80	0.888	0.088	−4.5877	−4.7286
D. with $\sigma_0 = \sigma_1 = 2$	0.2	0.80	0.939	0.139	−3.3962	−3.7715
	0.5	0.80	0.955	0.155	−4.1361	−4.5843

[a]All scenarios are based on model (2) where A: $C_{0,1} = 1.0$, $C_{0,2} = 2.9$, $C_{0,3} = 4.8$, $C_{0,4} = 6.7$, $C_{1,1} = -1.5$, $C_{1,2} = -1.1$, $C_{1,3} = -0.6$, and $C_{1,4} = -0.10$; B: $C_{0,1} = 2.25$, $C_{0,2} = 3.75$, $C_{0,3} = 5.25$, $C_{0,4} = 6.75$, $C_{1,1} = -2.25$, $C_{1,2} = -1.75$, $C_{1,3} = -1.25$, and $C_{1,4} = -.75$; C: $C_{0,1} = 1.0$, $C_{0,2} = 0.5$, $C_{0,3} = 0$, $C_{0,4} = -0.5$, $C_{1,1} = -1.5$, $C_{1,2} = -0.1$, $C_{1,3} = -1.6$, and $C_{1,4} = -3.0$; D: $C_{0,1} = 1.1$, $C_{0,2} = 2.7$, $C_{0,2} = 4.3$, $C_{0,4} = 5.9$, $C_{1,1} = -1.65$, $C_{1,2} = -0.65$, $C_{1,3} = 1.65$, and $C_{1,4} = 2.65$.

Table 2

Asymptotic bias ($AUC^* - AUC$) in estimating the AUC with a GRE when the true random effects distribution is a combined GRE/FM model (T) given by equations (8) and (9) with varying η_0 and η_1. These calculations assume that $\sigma_0 = \sigma_1 = 2$, $J = 4$, $K = 5$, there is a prevalence of 0.2, and there are no rater-specific differences in the ROC curves. For all scenarios, $C_{0,1} = 0.5$, $C_{0,2} = 1.5$, $C_{0,3} = 2.0$, $C_{0,4} = 3.0$, $C_{1,1} = -4.0$, $C_{1,2} = -2$, $C_{1,3} = -1$, and $C_{1,4} = 0$.

η_0	η_1	True AUC	AUC^*	Bias	$E_T\{\log L_T\}$	$E_T\{\log L_{\text{GRE}}\}$
0.1	0.1	0.848	0.707	−0.141	−4.3301	−4.3604
0.01	0.01	0.822	0.802	−0.200	−4.2689	−4.2710
0.4	0.4	0.920	0.802	−0.118	−3.9189	−3.9543
0.01	0.2	0.848	0.621	−0.227	−4.2875	−4.2925
0.2	0.01	0.853	0.727	−0.126	−4.2736	−4.3155
0.2	0.2	0.874	0.677	−0.197	−4.2814	−4.3218

patients are subject to no diagnostic error and the remainder follow a GRE dependence structure. In this case, there is sizable asymptotic bias for estimating AUC ($0.707 - 0.848 = -0.141$). Even for a very small departure from a GRE model (second row in table), in which only 1% of individuals have no diagnostic error and the remainder follow a GRE model, there is a nonnegligible asymptotic bias when we incorrectly assume a GRE model ($0.802 - 0.822 = -0.20$). Fortunately, even for a small departure from a GRE model, we are able to distinguish between a true GRE/FM model and a misspecified GRE model based on likelihood comparisons. In Table 2, the expectations of the individual contribution to the log likelihood for the true model, $E_T(\log L_T)$, were nonnegligibly larger than these expectations for the misspecified GRE model, $E_T(\log L_{\text{GRE}})$. These results are in contrast to the results for binary tests (Albert and Dodd, 2004), where the expected log likelihoods under the correct and incorrect models were often indistinguishable from each other, suggesting difficulty in distinguishing between these two conditional dependence models.

Using simulation studies, the ability to distinguish between models is discussed in the next section.

4. Simulation

Simulations were performed to examine the robustness of AUC estimation to the dependence between tests and to examine the difficulty in choosing the correct dependence structure. I simulated data according to either the GRE or FM model, and I fit both the GRE and FM models to realizations from each simulation scenario. Table 3 shows the results of simulations under four scenarios. Scenarios A and B are GRE models with AUC = 0.90 and 0.80, respectively. Scenarios C and D are FM models with AUC = 0.90 and 0.80, respectively. I present results for sample sizes of $I = 150$ and $I = 500$. For the larger sample size ($I = 500$), estimates of the AUC are approximately unbiased under the correct dependence structure. However, consistent with the asymptotic results in the previous section, estimates are biased under a misspecified dependence structure. For example, for $I = 500$ with scenario A, estimates of the AUC under the correctly specified GRE model were nearly unbiased, while estimates

Table 3

Simulation study: estimating AUC under a correct and a misspecified random effects distribution. Prevalence is 0.20, $J = 7$, and $K = 5$ in all simulations. $I = 150$ and 500. Models were fit under FM and GRE random effects distributions and were generated under both of these distributions. For each scenario, 1000 simulated datasets were generated.

				Avg. est. AUC		
True random effects distribution	Scenario[a]	I	True AUC	Avg. est. GRE	AUC FM	% ($L_{\text{GRE}} > L_{\text{FM}}$)
GRE	*A*	150	0.90	0.90 (0.028)	0.69 (0.27)	99.8
	A	500	0.90	0.90 (0.011)	0.70 (0.26)	100
GRE	*B*	150	0.80	0.80 (0.04)	0.77 (0.22)	100
	B	500	0.80	0.80 (0.02)	0.78 (0.21)	100
FM	*C*	150	0.90	0.72 (0.16)	0.90 (0.02)	2.2
	C	500	0.90	0.73 (0.17)	0.90 (0.01)	0.1
FM	*D*	150	0.80	0.58 (0.06)	0.77 (0.10)	9.8
	D	500	0.80	0.55 (0.04)	0.80 (0.02)	0.4

[a]Scenario A is based on (2) with GRE random effects and $C_{0,1} = 1.0$, $C_{0,2} = 2.9$, $C_{0,3} = 4.8$, $C_{0,4} = 6.7$, $C_{1,1} = -1.5$, $C_{1,2} = -1.1$, $C_{1,3} = -0.6$, and $C_{1,4} = -0.1$, and $\sigma_0 = \sigma_1 = 1$. Scenario B is based on (2) with GRE random effects and $C_{0,1} = 1.0$, $C_{0,2} = 0.5$, $C_{0,3} = 0$, $C_{0,4} = -0.5$, $C_{1,1} = -1.5$, $C_{1,2} = -0.1$, $C_{1,3} = -1.6$, and $C_{1,4} = -3.0$, and $\sigma_0 = \sigma_1 = 1$. Scenario C is based on (2) with FM random effects and $C_{0,1} = 0$, $C_{0,2} = 0.5$, $C_{0,3} = 1.0$, $C_{0,4} = 1.5$, $C_{1,1} = -1.5$, $C_{1,2} = -1.1$, $C_{1,3} = -0.6$, and $C_{1,4} = -0.1$, and $\eta_0 = \eta_1 = 0.2$. Scenario D is based on (2) with FM random effects and $C_{0,1} = -0.75$, $C_{0,2} = -0.25$, $C_{0,3} = 0.25$, $C_{0,4} = 0.75$, $C_{1,1} = -1.5$, $C_{1,2} = -1.1$, $C_{1,3} = -0.6$, and $C_{1,4} = -0.1$, and $\eta_0 = \eta_1 = 0.2$.

of AUC were biased and highly variable under the misspecified FM model. Fortunately, unlike the case of binary tests where distinguishing between the FM and GRE models is very difficult (Albert and Dodd, 2004), we are able to distinguish between models for the dependence structure using likelihood comparisons. Because the GRE and FM models have the same number of parameters, by the likelihood principle, the model with the largest likelihood is most supported by the data. Based on simply ranking the likelihoods, there was over a 99% probability of choosing the correct model with both sample sizes for scenario A. For scenario D, the correct FM model was chosen in over 90% of the cases with $I = 150$ and in over 99% of the simulated realizations with the larger sample size of $I = 500$. This model selection is based on choosing between two models, one of which is known to be correct. In practice, of course, this is not feasible because many more models are possible, and in any case, we never know the true model. The simulation is useful in demonstrating the ability to distinguish between competing models for the dependence structure. In practice, model diagnostics will need to be performed in order to assume that the dependence structure is appropriately modeled. This will be discussed in Section 5.

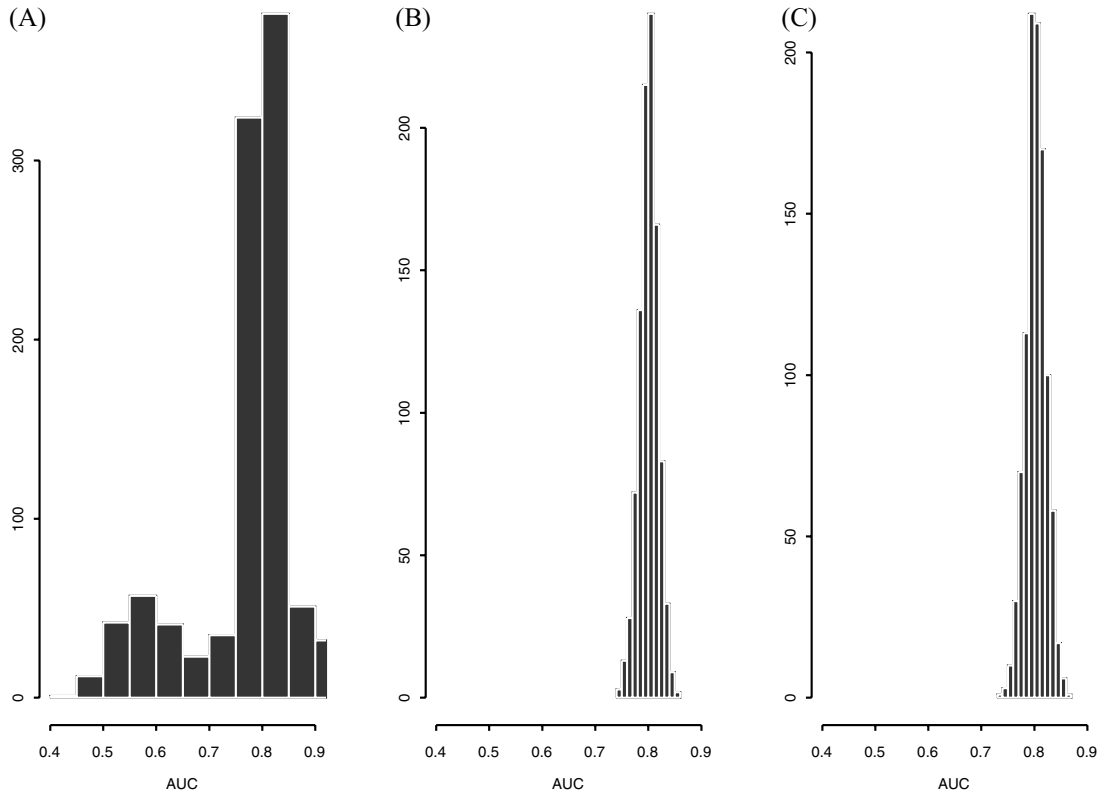

Figure 1. Histogram of estimated AUC under scenario D in Table 3. (A) $I = 150$ and prevalence = 0.2. (B) $I = 500$ and prevalence = 0.2. (C) $I = 150$ and prevalence = 0.5.

For scenarios A, B, and C, estimates of AUC were nearly unbiased under the correct dependence structure for the smaller sample size. However, for scenario D, there was sizable small-sample bias under the correct FM model. A histogram of parameter estimates under scenario D is presented in Figure 1. For the smaller sample size ($I = 150$) the estimates appear to be bimodal, demonstrating the poor small-sample properties in this case (Figure 1A). Estimates were nearly unbiased and approximately normally distributed for scenario D for the larger sample size (Figure 1B). A simulation with a larger prevalence of 50% for the sample size of $I = 150$ was conducted. Although there was sizable bias for a prevalence of 20% (Figure 1A; Table 3), there was little bias for the larger prevalence of 50%. Specifically, the mean estimates of AUC under the correctly specified FM model were 0.80 (SE = 0.02). Further, in this case, the AUC estimates appear to be normally distributed (Figure 1C).

5. An Analysis of the Uterine Cancer Data

I applied the random effects approaches to a study in which seven pathologists ($J = 7$) rated 118 slides ($I = 118$) with potential carcinoma in situ of the uterine cervix. Each pathologist independently rated each of the 118 slides in a random order, classifying lesions on each of the slides on a five-category ordinal scale ranging from 1 (negative) to 5 (invasive carcinoma) ($K = 5$). In this study, there was a clinical definition of carcinoma which ideally could be used as a gold standard. However, as mentioned by Zhou et al. (2005), diagnosis based on this clinical definition was not available due to technological limitations. More details on this study are in Holmquist, McMahan, and Williams (1967). Landis and Koch (1977) applied their methodology to these data in order to assess agreement between raters. However, assessing agreement between raters is a different problem from estimating diagnostic accuracy. Zhou et al. (2005) estimated rater-specific ROC curves on these data with their nonparametric conditional independence approach. I will illustrate the importance of accounting for dependence between tests using this dataset.

I fit models (12) and (13) under independence and GRE and FM dependence structures. The maximum likelihood estimates of rater-specific AUC, along with the log likelihoods for each model, are presented in Table 4. Standard errors of the estimated rater-specific AUCs were estimated using the bootstrap (with 800 bootstrap replications). Models A, B, and C show estimates of rater-specific AUC for a cumulative probit model with additive rater-specific effects (12) under conditional independence and GRE and FM dependence structures, respectively. Comparisons between the conditional independence, GRE, and FM models were made using the Akaike information criterion (AIC; Akaike, 1974). The AIC is a penalized likelihood technique whereby different potentially nonnested models are ranked according to the value $AIC = -2\log L + 2p$, where p is the number of model parameters. Ranking the AIC values is not an explicit test of whether one model fits better than another. Rather, the AIC is a method for choosing a model, among numerous potentially nonnested competing models, which is most supported by the data. Models with the smallest AIC are those most supported

Table 4
Analysis of the cervical cancer pathology data. Estimates (standard errors) of the rater-specific AUC are obtained under models (12) and (13) and under different random effects distributions.

		Estimated rater-specific AUC (rater)								
Model	Description	1	2	3	4	5	6	7	logL	AIC
A	Model (12), indep	0.96 (0.01)	0.94 (0.02)	0.91 (0.02)	0.92 (0.02)	0.92 (0.02)	0.91 (0.02)	0.96 (0.01)	−836.02	1714.0
B	Model (12), GRE	0.74 (0.07)	0.76 (0.05)	0.72 (0.06)	0.84 (0.06)	0.68 (0.05)	0.76 (0.08)	0.83 (0.04)	−711.37	1468.7
C	Model (12), FM	0.96 (0.01)	0.92 (0.03)	0.91 (0.02)	0.92 (0.02)	0.91 (0.02)	0.91 (0.02)	0.95 (0.02)	−798.68	1643.4
D	Model (13), indep	0.94 (0.02)	0.92 (0.03)	0.90 (0.04)	0.93 (0.03)	0.95 (0.02)	0.86 (0.03)	0.99 (0.01)	−779.23	1672.5
E	Model (13), GRE	0.87 (0.04)	0.83 (0.05)	0.89 (0.04)	0.99 (0.01)	0.86 (0.05)	0.79 (0.05)	0.95 (0.04)	−660.02	1438.0
F	Model (13), FM	0.94 (0.02)	0.89 (0.03)	0.90 (0.04)	0.94 (0.03)	0.93 (0.02)	0.85 (0.03)	1.00 (0.01)	−740.48	1599.0

by the data. The AIC is substantially smaller for the GRE than for the FM model, and both are substantially smaller than that for the conditional independence model. Models *D*, *E*, and *F* show estimates for the cumulative probit model (13), which allows for a more flexible mean structure than the rater-specific constant shifts in the cutpoints assumed by model (12). Models *E* and *F* correspond to the GRE and FM models, while model *D* corresponds to the conditional independence model. Zhou et al.'s (2005) nonparametric conditional independence model is identical to model *D* (e.g., log likelihood as well as rater-specific estimates of AUC are identically estimated to round-off error). Note that for each dependence structure, the cumulative probit models with additive rater effects (*A*, *B*, and *C*) have a larger AIC (i.e., worse fit) than the models with nonadditive rater effects (*D*, *E*, and *F*). Further, a likelihood ratio test comparing model *F* with model *C* was highly significant, favoring the more complex model (difference in 2× log likelihood was 116.4 with 36 degrees of freedom).

Estimates of model (13) with the GRE and the FM dependence structures (models *E* and *F*) resulted in substantially reduced AICs relative to the conditional independence model (model *D*) of Zhou et al. The dependence parameters for the GRE model were estimated as $\hat{\sigma}_0 = 1.95$ and $\hat{\sigma}_1 = 0.76$, while the dependence parameters for the FM model were estimated as $\hat{\eta}_0 = 0.20$ and $\hat{\eta}_1 = 0.02$. Thus, both the GRE and FM models suggest sizable conditional dependence. The GRE model had an AIC value which was substantially smaller than that obtained with the independence or the FM model, suggesting that of the three models, the GRE describes the uterine cancer data best. Further, estimates of AUC (along with their bootstrap estimates of standard error) were substantially different between the different models, with the GRE model, for all but one rater, showing rater-specific AUC estimates lower than those obtained with either the independence or the FM model. In addition, the ordering of the estimated rater-specific AUCs differs between the GRE and independence models. For example, the independence model estimates rater 7 as having the largest AUC among all raters, while the GRE model estimates rater 4 as having the largest AUC. Figure 2 shows the estimated rater-specific ROC curves derived under model (13) with a GRE dependence structure (i.e., best-fitting model). These estimates are substantially different from those presented by Zhou et al., which were derived under a conditional independence structure. Overall, the diagnostic accuracy is not as good under the better-fitting GRE model as compared with the conditional independence model.

I examined the fit of the independence model in more detail. Specifically, I developed a diagnostic for dependence, whereby I plotted the observed minus the expected correlation for each pairwise combination of tests. This is a generalization of a diagnostic for the correlation structure developed by Qu et al. (1996) for the random effects model with binary test results. Specifically, I plot $\{Empirical\ Correlation\}_{ij} - \widehat{Corr}_{ij}$ for each pairwise comparison of tests i and j, where $i < j$ and where $\widehat{Corr}_{ij}$ is the model-based correlation under model (13) with a conditional independence assumption (e.g., $b_{d_i,i}$ all equal to zero). The expected correlation is derived in the Appendix. Figure 3 shows the diagnostic plot for the uterine cancer data. The difference appears to be relatively constant and substantially larger than zero, reflecting the inadequacy

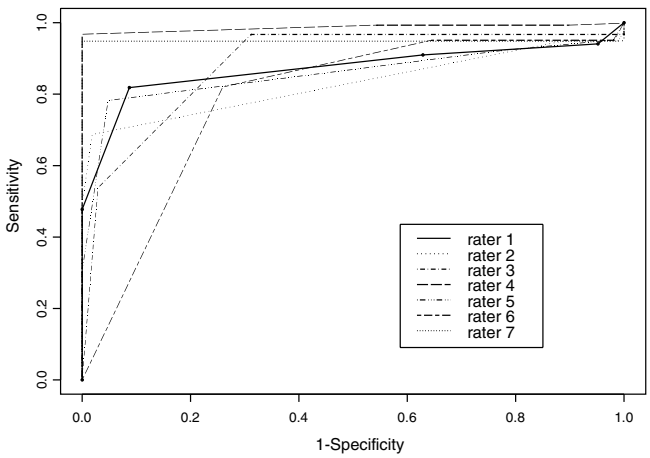

Figure 2. Estimated rater-specific ROC curves for model (2) with a Gaussian random effect.

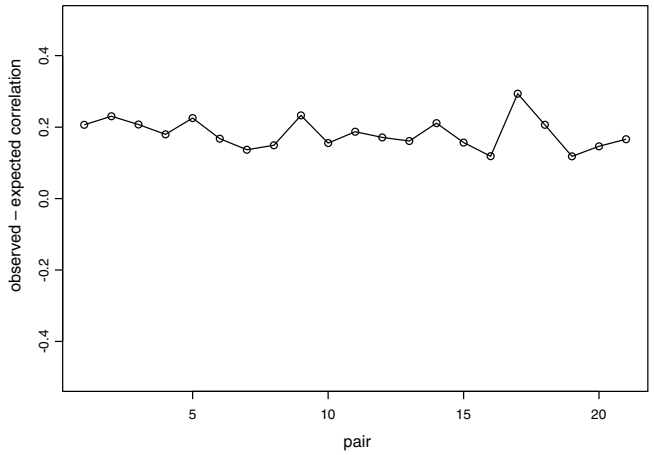

Figure 3. Observed minus expected correlation under a conditional independence assumption.

of the conditional independence model to explain the large correlation between tests. A similar plot is presented for the GRE model (Figure 4). Specifically, I plot the difference of the observed correlation and the expected correlation under model (13) with a GRE dependence structure. Note that these differences are substantially closer to zero, demonstrating that the GRE model provides a much better description of the correlation between tests than the conditional independence model.

I fit the combined GRE/FM model with (13) to the uterine cancer data. The resulting log likelihood for the combined model was -659.0, compared with -660.0 for the GRE model, suggesting that there is little improvement in incorporating the extra FM component for describing the dependence structure (the AIC for the GRE/FM model was 1440.0, compared with a value of 1438.0 for the GRE model). Estimates of the combined model parameters were $\hat{\sigma}_0 = 1.82, \hat{\sigma}_1 = 0.76, \hat{\eta}_0 = 0$, and $\hat{\eta}_1 = 0.01$, which nearly reduces to those for the GRE model (where $\eta_0 = \eta_1 = 0$, $\hat{\sigma}_0 = 1.95$, and $\hat{\sigma}_1 = 0.76$). Further estimates of the rater-specific AUCs under the combined model were identical, within round-off error, to those reported for the GRE model (under (13)) in Table 4.

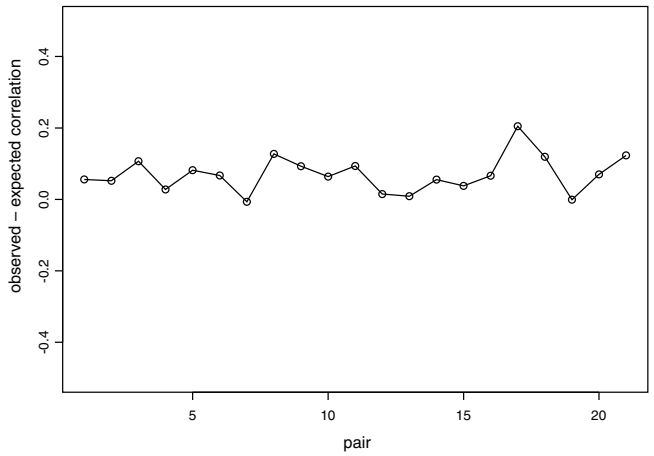

Figure 4. Observed minus expected correlation under a Gaussian random effects assumption.

6. Discussion

Methodology which accounts for conditional dependence between tests was proposed to estimate ROC curves without a gold standard. This work extends the GRE model of Qu et al. (1996) from the estimation of sensitivity and specificity for binary tests to the estimation of the ROC curve for ordinal tests. The methodology also extends the recent work of Zhou et al. (2005) to allow for conditional dependence between tests. Asymptotic results and simulations show the importance of correctly modeling the dependence structure between tests. Further, I illustrated the methodology with cervical cancer testing data, which was used to illustrate the nonparametric conditional independence approach proposed by Zhou et al. The analysis of these data illustrated the importance of accounting for conditional dependence. Recall that the GRE model fits substantially better than the independence model (by AIC comparisons as well as by using a graphical diagnostic) and that estimates of the ROC curve under the GRE model, in general, showed substantially poorer diagnostic accuracy than the independence model. The GRE model also fits better than the FM model, and there was no advantage in fitting a combined GRE/FM model over the GRE model in describing these data. Software for fitting the different models is written in `GAUSS` (Aptec Systems, 1992) and is available from the author.

The proposed random effects models fit into a larger literature in modeling repeated ordinal data. For example, Kottas et al. (2005) proposed a general class of random effects models in which the random effects distribution is a mixture of normals. Although this approach is similar to the proposed GRE model, there are notable differences. First, in this article, the GRE model is used to account for conditional dependence between tests while estimating the ROC curve without a gold standard. In contrast, Kottas et al. focused on using the model to estimate agreement between raters. Second, a maximum likelihood approach is taken in this article, while Kottas et al. proposed a Bayesian approach for inference.

Previous results (Albert and Dodd, 2004) showed that for a limited number of binary tests it is very difficult to distinguish between competing models for the dependence between tests, yet the choice of a dependence structure has a large effect on model-based estimates of sensitivity and specificity. In this article, I investigated the problem of estimating ROC curves from ordinal tests without a gold standard. Similar to the results for binary tests, inferences on the ROC curve for ordinal tests were not robust to the choice of dependence structure between tests. However, unlike the case for binary tests, where it is difficult to distinguish between competing models for the dependence between tests, it is easier to distinguish between models with ordinal tests. These results highlight the importance of developing model diagnostics for correctly specifying the dependence between ordinal tests. I

proposed a simple technique for examining this dependence structure by comparing the expected versus the observed correlation between ordinal tests. More research in this area is warranted.

There are potential problematic issues with estimating ROC curves without a gold standard. Particularly when the diagnostic accuracy is low (e.g., when the AUC is below 80%), I found that there was a tendency for the likelihood to have multiple local maximums, making estimation problematic. At the very least, practitioners should perform estimation using different starting values for the parameters in order to be assured that a global maximum has been achieved.

A major concern with using latent class modeling approaches for estimating diagnostic accuracy without a gold standard has been that the latent state may not relate to an actual clinical state or gold standard test (Begg and Metz, 1990; Hadgu and Miller, 2001; Pepe and Alonzo, 2001). For example, the clinical state or gold standard, if it exists at all, may truly be a continuous measure and, therefore, be misspecified as a dichotomous gold standard. Zhou et al. (2005) addressed this concern by assuming the existence of an unobserved binary gold standard test.

Albert and Dodd (2006) discussed the advantages of observing even a small percentage of gold standard information in order to improve the statistical properties of estimators of diagnostic accuracy for binary tests. Specifically, they showed that estimates of sensitivity and specificity are insensitive to the conditional dependence between binary tests when the gold standard test is available for a small fraction of randomly selected individuals. Examining this problem for estimating the ROC curve for ordinal tests is an area of future research.

Acknowledgements

I thank the associate editor and two reviewers for their thoughtful comments, which led to an improved manuscript. I thank Dr Lori Dodd for discussions on this subject. I thank Sara Joslyn for editing this manuscript. I thank the Center for Information Technology, NIH, for providing access to the high-performance computational capabilities of the Biowulf cluster computer system.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **AC-19,** 716–723.

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60,** 427–435.

Albert, P. S. and Dodd, L. E. (2006). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. Biometric Research Branch/NCI Technical Report 40. Available at `http://linus.nci.nih.gov/~brb/TechReport.htm`).

Albert, P. S., McShane, L. M., Shih, J. H., and the NCI Bladder Tumor Network. (2001). Latent modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57**, 610–619.

Aptech Systems. (1992). *Gauss Systems,* Version 3.0. Kent, Washington: Aptech Systems.

Begg, C. B. and Metz, C. E. (1990). Commentary: Consensus and "gold standards." *Medical Decision Making* **10,** 24–29.

Booth, J. G. and Sarkar, S. (1998). Monte Carlo approximation of bootstrap variances. *American Statistician* **52,** 354–357.

Choi, Y. K., Johnson, W. O., Collins, M. T., and Gardner, I. A. (2006). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics* **11,** 210–229.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Espeland, M. A. and Handelman, S. L. (1989). Using latent class models to characterize and assess relative-error in discrete measurements. *Biometrics* **45,** 587–599.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* **87,** 476–486.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Hadgu, A. and Miller, W. (2001). Comment on: "Using a combination of reference tests to assess the accuracy of a diagnostic test." *Statistics in Medicine* **20,** 656–658.

Harville, D. A. and Mee, R. W. (1984). A mixed-model procedure for analyzing ordinal categorical data. *Biometrics* **40,** 393–408.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50,** 933–944.

Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990). Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* **10,** 24–29.

Holmquist, N. D., McMahan, C. A., and Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology* **84,** 334–345.

Ishwaran, H. (2000). Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *Canadian Journal of Statistics* **28,** 715–730.

Ishwaran, H. and Gatsonis, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics* **28,** 731–750.

Kottas, A., Muller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics* **14,** 610–625.

Landis, J. R. and Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple raters. *Biometrics* **33,** 363–374.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. New York: Chapman and Hall.

Moulton, L. H. and Zeger, S. L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap. *Biometrics* **45,** 381–394.

Pepe, M. S. and Alonzo, T. A. (2001). Reply to comment on: "Using a combination of reference tests to assess the accuracy of a diagnostic test" (1999, **18**, 2987–3003). *Statistics in Medicine* **20,** 658–660.

Qiu, Z., Song, P. X. K., and Tan, M. (2002). Bayesian hierarchical models for multi-level repeated data using WinBUGS. *Journal of Biopharmaceutical Statistics* **12,** 121–135.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **53,** 797–810.

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis* **22,** 537–557.

Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurements* **23,** 283–297.

Zhou, X. H., Castelluccio, P., and Zhou, C. (2005). Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* **61,** 600–609.

Received January 2006. *Revised June* 2006. *Accepted September* 2006.

Appendix

Evaluating $E_T\{\log L_M\}$ for $J = 4$ When T Is the True Model and M Is the Misspecified Model

Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)'$ be a random vector reflecting the four ordinal ratings on a given patient. Denote $\boldsymbol{i} = (i_1, i_2, i_3, i_4)$. The $E_T\{\log L_M\}$ can be written as

$$E_T\{\log L_M\} = \sum_{\boldsymbol{i}} P_T\{\boldsymbol{Y} = \boldsymbol{i}\} \times \log\left\{\sum_{l=0}^{1} P_M(\boldsymbol{Y} = \boldsymbol{i} \mid d = l) P_d^l (1 - P_d)^{1-l}\right\},$$

where $P_T(\boldsymbol{Y})$ is the joint distribution of $\boldsymbol{Y}$ under the true model, evaluated using equation (1). In addition, $P_M(\boldsymbol{Y} \mid d_i)$ is the conditional distribution of $\boldsymbol{Y}$ given true disease status d under the misspecified model, and P_d is the probability that $d = 1$ or the disease prevalence. Further, the summation is over all possible combinations of i_1, i_2, i_3, and i_4 (5^4 terms for $K = 5$).

Correlation between Two Tests under an Independence Model

Denote Y_1 and Y_2 as any two tests on a given person. The correlation between these two ordinal tests can be expressed as $corr(Y_1, Y_2) = \text{cov}(Y_1, Y_2)/\sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}$, where $\text{cov}(Y_1, Y_2) = P_d(1 - P_d)\{E(Y_1 \mid d = 1)E(Y_2 \mid d = 1) + E(Y_1 \mid d = 0)E(Y_2 \mid d = 0) - E(Y_1 \mid d = 1)E(Y_2 \mid d = 0) - E(Y_1 \mid d = 0)E(Y_2 \mid d = 1)\}$, and where $\text{Var}(Y_j) = P_d\{E(Y_j^2 \mid d = 1) - E(Y_j \mid d = 1)^2\}\ (1 - P_d)\{E(Y_j^2 \mid d = 0) - E(Y_j \mid d = 0)^2\} + P_d(1 - P_d)\{E(Y_j \mid d = 0) - E(Y_j \mid d = 1)\}$. Further, $E(Y_j \mid d) = \sum_{l=1}^{K} lP(Y_j = l \mid d)$ and $E(Y_j^2 \mid d) = \sum_{l=1}^{K} l^2 P(Y_j = l \mid d)$, where $P(Y_j \mid d)$ is given by equation (11).