Appl. Statist. (2003)
52, *Part* 1, *pp.* 63–76

Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests

Marios P. Georgiadis, Wesley O. Johnson, Ian A. Gardner and Ramanpreet Singh

University of California, Davis, USA

[Received June 2001. Revised August 2002]

Summary. Models for multiple-test screening data generally require the assumption that the tests are independent conditional on disease state. This assumption may be unreasonable, especially when the biological basis of the tests is the same. We propose a model that allows for correlation between two diagnostic test results. Since models that incorporate test correlation involve more parameters than can be estimated with the available data, posterior inferences will depend more heavily on prior distributions, even with large sample sizes. If we have reasonably accurate information about one of the two screening tests (perhaps the standard currently used test) or the prevalences of the populations tested, accurate inferences about all the parameters, including the test correlation, are possible. We present a model for evaluating dependent diagnostic tests and analyse real and simulated data sets. Our analysis shows that, when the tests are correlated, a model that assumes conditional independence can perform very poorly. We recommend that, if the tests are only moderately accurate and measure the same biological responses, researchers use the dependence model for their analyses.

Keywords: Bayesian approach; Data augmentation approach; Dependent screening tests; Gibbs sampling; Prevalence

1. Introduction

New screening tests for disease are often introduced because they are cheaper, less labour intensive or yield results more rapidly than existing methods. However, a frequent prerequisite is that the new test should be at least as accurate as the existing test that it will replace. A critical assessment of the accuracy of a test requires a panel of reference samples of known status (so-called gold standard samples) or the use of alternative approaches such as latent class analysis that do not require the true disease of each individual to be known. For the problem to be identifiable in the two-test two-population case when there is no reference test (see Hui and Walter (1980) and Johnson *et al.* (2001)) existing methods require that the test results be independent, conditional on disease status. The assumption of conditional independence can be tenuous, however, especially if the two tests have a similar biological basis. For example, two serological tests that detect antibodies to the same disease or two skin tests that detect cellular immune responses would be expected to be moderately to strongly conditionally dependent. In contrast, tests that measure different biological phenomena, e.g. a serum antibody detection test and virus isolation from tissues, would probably be conditionally independent or only weakly dependent. When test dependence exists, the use of a conditional independence model might result in biased estimates of sensitivity and specificity (Vacek, 1985; Torrance-Rynard and Walter, 1997).

Address for correspondence: Wesley O. Johnson, Department of Statistics, University of California, Davis, CA 95616, USA.
E-mail: wojohnson@ucdavis.edu

 0035–9254/03/52063

Several researchers have developed models to allow the estimation of the accuracy of dependent tests, including Sinclair and Gastwirth (1996), Qu *et al.* (1996), Qu and Hadgu (1998) and Hadgu and Qu (1998). Recently, Dendukuri and Joseph (2001) have developed two Bayesian models for single-population data: a fixed effects model and a random-effects model that was a modification of previously developed models. Both models lack identifiability unless results based on at least four tests are available. With single-population data, there are only 3 degrees of freedom for estimating seven and nine parameters in fixed and random-effects models respectively (Dendukuri and Joseph, 2001). The random-effects model of Dendukuri and Joseph (2001) involves the specification of priors on the prevalence, and on the 'intercept–slope' combinations for diseased and non-diseased individuals, which is somewhat complicated. In contrast, their prior specification for the fixed effects model is relatively straightforward. For both models, most of the full conditionals are unrecognizable and require a general purpose simulation mechanism.

We develop a simple Bayesian approach to inferences for prevalence and accuracy of two screening tests under dependence when K populations are tested. We were motivated to develop such an approach because, in animal health surveillance, most screening of animal populations is done by serum antibody tests, which are likely to be conditionally dependent for the same infectious agent. In addition, animals are naturally aggregated into herds or flocks, often resulting in multiple, $K > 1$, populations for analysis. The addition of a second population in the context of a fixed effects model increases the degrees of freedom to 6 for estimating eight parameters and, hence, decreases the reliance on additional (prior) information. In our experience, the use of $K > 2$ populations does not further reduce the identifiability problem. However, there is no difficulty in increasing the number of populations, thus increasing the available information for those parameters that are estimable from the data alone. Finally, all the full conditionals in our Gibbs sampler involve simple, independent beta and Bernoulli sampling.

Since our model is not identifiable, results necessarily depend heavily on the prior input. Often, substantive prior information is available for the characteristics of one test and/or the prevalences of the sampled populations with limited information about the remaining parameters. Frequently, when a standard reference test has been used for a long time, sensitivity and specificity estimates of that test are available (see for example Jarrett *et al.* (1982), Zhou (1998) and Alonzo and Pepe (1999)). In other situations, prevalence data might be available for some populations, thus allowing for an adequate prior specification of those parameters. We present an analysis of the effect of prior selection on inferences and give guidance about how much prior information is necessary to make reasonable inferences.

In Section 2, we discuss serum antibody detection tests for toxoplasmosis in pigs and Section 4 presents the analysis of these data. Our model and methods are presented in Section 3. Section 5 gives results of a simulation study based on known values of the parameters and presents an evaluation of the effect of various 'correct' and 'incorrect' prior specifications on the resulting inferences. We present our conclusions and recommendations in Section 6.

2. Screening for toxoplasmosis

For many animal diseases, serologic screening of blood samples from live animals or animals at slaughter is the main surveillance method for monitoring temporal and spatial patterns of disease of economic or public health importance. In this paper, we analyse data from a test evaluation study of five serum-based screening tests (Dubey *et al.*, 1995) for toxoplasmosis (*Toxoplasma gondii*) in pigs to demonstrate an application of our model to real data. We chose these data for three reasons. First, humans are often believed to become infected with

Toxoplasma gondii through the consumption of undercooked infected pork and rapid screening tests are needed that can detect infected pigs, and hence reduce the risk of infection of humans. Second, Dubey *et al.* (1995) used a highly credible reference test (the isolation of the parasite from heart muscle in cats or mice) which ensured accurate estimation of test accuracy for the five serum-based tests. Third, a subsequent evaluation of these data indicated moderate to strong conditional dependence between all five tests (Gardner *et al.*, 2000).

In the study of Dubey *et al.* (1995), samples were collected in two batches (group 1, samples 1–463; group 2, samples 464–1000) and we used this grouping to define the two populations for our analysis because data on the source herds were not available. Because *Toxoplasma gondii* infection is subclinical in adult pigs and samples were collected randomly from representative herds, it was reasonable to assume initially that the sensitivities and specificities were constant in the two populations.

We used the data on the microscopic agglutination test (MAT) and enzyme-linked immunosorbent assay (ELISA) for the present analysis because they had the greatest sensitivity and specificity of all the tests evaluated. In addition, the ELISA test is currently being considered for widespread screening of pigs in the USA as a replacement for the MAT, which has been the most widely used test for toxoplasmosis in animals. The primary advantage of the ELISA test over the MAT is that the ELISA test can be automated, enabling the rapid screening of large numbers of samples. We used the same decision cut-offs reported by Dubey *et al.* (1995) (positive if titer $\geqslant$ 1:20 for the MAT and an optical density value greater than 0.36 for the ELISA). We excluded one record that had a missing ELISA result. The cross-classified data are presented in Table 1.

We analysed these data with three models: a *conditional independence* model (the expectation–maximization (EM) algorithm (see Dempster *et al.* (1977)) and the SEM algorithm (see Meng and Rubin (1991)), the Bayesian independence model of Johnson *et al.* (2001) and the Bayesian dependence model as described in the following section.

3. Model and methods

We denote a positive or negative test result for test i ($i = 1, 2$) by T_i^+ and T_i^- respectively, whereas the presence or absence of the condition that the test detects is symbolized as D and $\bar{D}$. The sensitivity of test i is $\eta_i = P(T_i^+|D)$ and its specificity is $\theta_i = P(T_i^-|\bar{D})$. The prevalence of the condition in population K is $\pi_k = P(D|k)$.

Table 1. Cross-classified test results for the MAT and ELISA presented by Dubey *et al.* (1995)

MAT	*ELISA*	
	T^+	T^-
T^+	67	25
T^-	41	329
T^+	97	33
T^-	36	371

The results of the two tests performed on independent samples from $K \geqslant 1$ populations can be cross-classified in K 2×2 tables; see Table 1 when $K = 2$. Prevalences must be distinct. The data are defined as $\{x_{ijk}\}$, where for example x_{11k} corresponds to the number of individuals out of n_k sampled from population k that tested positively on both tests. Let $\{p_{ijk}\}$ denote the corresponding cell probabilities. These data are assumed to constitute K independent multinomial samples.

Define the conditional probabilities

$$\begin{aligned}
\eta_{11} &= P(T_1^+, T_2^+|D),\\
\eta_{12} &= P(T_1^+, T_2^-|D),\\
\eta_{21} &= P(T_1^-, T_2^+|D),\\
\eta_{22} &= P(T_1^-, T_2^-|D),\\
\theta_{11} &= P(T_1^+, T_2^+|\bar{D}),\\
\theta_{12} &= P(T_1^+, T_2^-|\bar{D}),\\
\theta_{21} &= P(T_1^-, T_2^+|\bar{D}),\\
\theta_{22} &= P(T_1^-, T_2^-|\bar{D}),
\end{aligned}$$

which are assumed to be the same for all populations (Hui and Walter (1980), Georgiadis *et al.* (1998) and Johnson *et al.* (2001); see Enøe *et al.* (2000) for references to many others). By the law of total probability, the sensitivity and specificity for, say, test 1 are respectively $\eta_1 = P(T_1^+, T_2^+|D) + P(T_1^+, T_2^-|D) = \eta_{11} + \eta_{12}$ and $\theta_1 = P(T_1^-, T_2^+|\bar{D}) + P(T_1^-, T_2^-|\bar{D}) = \theta_{21} + \theta_{22}$ etc. The cell probabilities are given by $p_{ijk} = \pi_k \eta_{ij} + (1-\pi_k)\theta_{ij}$. Our model for two populations is thus completely specified. There are a total of $6 + K$ parameters since the η_{ij}s and the θ_{ij}s each sum to 1.

Then define the conditional correlations between test outcomes as

$$\begin{aligned}
\rho_D &= \frac{\delta_D}{\sqrt{\{\eta_1(1-\eta_1)\eta_2(1-\eta_2)\}}}, \qquad & \delta_D &= \eta_{11} - \eta_1\eta_2,\\
\rho_{\bar{D}} &= \frac{\delta_{\bar{D}}}{\sqrt{\{\theta_1(1-\theta_1)\theta_2(1-\theta_2)\}}}, \qquad & \delta_{\bar{D}} &= \theta_{22} - \theta_1\theta_2.
\end{aligned} \tag{3.1}$$

These are 0 if and only if the tests are conditionally independent, e.g. with $\eta_{11} = \eta_1\eta_2, \eta_{12} = \eta_1(1-\eta_2)$, etc.

The likelihood function for the observed multinomial data, $\{x_{ijk}\}$, is

$$L = \prod_{ijk} \{\pi_k \eta_{ij} + (1-\pi_k)\theta_{ij}\}^{x_{ijk}}. \tag{3.2}$$

Since the model lacks identifiability, we require informative prior distributions for at least some component parameters. We initially focused on $(\pi_1, \ldots, \pi_K, \eta_1, \theta_1)$ since we emphasize the situation where information for test 1 characteristics is moderate to strong, and information for the prevalences may be readily available. We choose independent beta priors for these parameters, as in Johnson and Gastwirth (1991), Gastwirth *et al.* (1991), Joseph *et al.* (1995), Mendoza-Blanco *et al.* (1996) and Johnson *et al.* (2001). Hanson *et al.* (2003) considered the case where the number of populations is large and the prevalences are considered to be exchangeable. Here we focus on a limited number of populations.

In general, we expect less information to be available for the correlations and for the accuracy of the new test (test 2). Thus, we consider the following reparameterization for the remaining

parameters, which facilitates the prior specification and implementation of the Gibbs sampler. Define

$$\left.\begin{aligned} \lambda_D &= P(T_2^+|T_1^+, D) = \eta_{11}/\eta_1, \\ \gamma_D &= P(T_2^+|T_1^-, D) = \eta_{21}/(1-\eta_1), \\ \lambda_{\bar{D}} &= P(T_2^-|T_1^-, \bar{D}) = \theta_{22}/\theta_1, \\ \gamma_{\bar{D}} &= P(T_2^-|T_1^+, \bar{D}) = \theta_{12}/(1-\theta_1). \end{aligned}\right\} \tag{3.3}$$

Thus equations (3.3), in conjunction with (π, η_1, θ_1), define a one-to-one transformation of the parameters. If the tests are conditionally independent, $\lambda_D = \gamma_D = \eta_2$ and $\lambda_{\bar{D}} = \gamma_{\bar{D}} = \theta_2$, whereas, if the tests are positively correlated, $\lambda_D > \eta_2 > \gamma_D$ and $\lambda_{\bar{D}} > \theta_2 > \gamma_{\bar{D}}$.

A simple and effective approach to inference involves the choice of independent beta prior distributions for the $6 + K$ parameters $\omega \equiv (\pi_1, \ldots, \pi_K, \eta_1, \theta_1, \lambda_D, \lambda_{\bar{D}}, \gamma_D, \gamma_{\bar{D}})$. We generally expect to choose informative priors for the η_1, θ_1 and some or all of the prevalences and non-informative priors for the remaining four. However, in Appendices A and B, we present a method that can be used to obtain informative beta priors for the λs and γs.

For our model, the likelihood (3.2) is complicated because the disease status is unknown. Let $\{z_{ijk}\}$ denote the collection of counts corresponding to individuals that are D. For example, z_{11k} counts the unobserved number of individuals who are D out of the x_{11k} from population k that tested positively on both tests. If the latent data $\{z_{ijk}\}$ were known, the 'augmented data' likelihood based on the new parameterization factorizes into terms that resemble independent binomial contributions (see Appendix A). Thus, the corresponding 'data augmentation' posterior is in the form of the product of eight independent beta posteriors, which are easily sampled. We develop a Gibbs sampling approach (Tanner, 1996); details are given in Appendix A.

We checked for convergence of the Gibbs sampler by plotting output for each sampled variate. The plots stabilized consistently within a few hundred iterations over all analyses. Repeated analyses based on different starting values resulted in virtually identical inferences with relatively small Monte Carlo (MC) sample sizes; runs with very large MC sample sizes verified that the stability of our algorithm was uniformly excellent.

4. Analysis of toxoplasmosis data

Sensitivity and specificity estimates for the MAT and ELISA test reported by Dubey *et al.* (1995) for both groups were $\eta_{\mathrm{M}} = 0.829$, $\eta_{\mathrm{E}} = 0.73$, $\theta_{\mathrm{M}} = 0.903$ and $\theta_{\mathrm{E}} = 0.859$. Since these estimates were based on an analysis that involved the isolation of the parasite from heart muscle in cats or mice as a highly credible test, the resulting estimates are taken as 'true' for our analysis so that we can see how well our method does under the more standard situation where parasite isolation results would be unavailable. The estimate of the prevalence of *Toxoplasma gondii* for the entire data set was 0.17. No true prevalence data were presented for the two groups but, because the apparent prevalences by using isolation methods from mice were 0.069 and 0.142 for groups 1 and 2 respectively, it is likely that the prevalences were truly distinct. Since the grouping of the subsamples was not based on biological characteristics of the samples, inferences about the prevalences for each subsample were not of primary interest.

The EM estimates and corresponding large sample 95% confidence intervals based on the conditional independence model are given in Table 2. Observe that the interval for η_{M} includes 1, raising questions about the appropriateness of large sample inference.

For the Bayesian analyses, we require *a priori* distributions for the parameters. However, because information for the accuracies of the test was based on the gold standard version of the same data used for analysis, this requirement was violated to illustrate our procedure. To mimic

Table 2. Estimates of test accuracies and population prevalences and corresponding 95% intervals based on maximum likelihood and Bayes methods using the conditional independence model and based on the Bayes dependence model†

Model	*Maximum likelihood conditional independence*	*Bayes dependence*	*Bayes conditional independence*
η_M	0.999 (0.99,1)	0.806 (0.628,0.923)	0.827 (0.698,0.93)
η_E	0.811 (0.794,0.828)	0.715 (0.393,0.929)	0.911 (0.776,0.996)
θ_M	0.972 (0.966,0.977)	0.895 (0.811,0.969)	0.941 (0.912,0.978)
θ_E	0.901 (0.899,0.903)	0.855 (0.768,0.939)	0.940 (0.904,0.989)
π_1	0.176 (0.173,0.179)	0.143 (0.028,0.255)	0.197 (0.14,0.268)
π_2	0.220 (0.217,0.223)	0.192 (0.072,0.306)	0.232 (0.179,0.295)

†The true values are $\eta_M = 0.829$, $\eta_E = 0.73$, $\theta_M = 0.903$ and $\theta_E = 0.859$.

a situation with good information about one test, we constructed prior distributions for MAT sensitivity and specificity that were centred on its true values but were relatively diffuse. The prior for η_M had a mode of 0.83 and a fifth percentile 0.15 below the mode, Be(24.09, 5.73), and the prior for θ_M had a mode equal to 0.9 and fifth percentile equal to 0.75, Be(23.05, 3.45). The priors for π_1 and π_2 were arbitrarily centred at 0.07 and 0.20 respectively and were made very diffuse (95th percentiles 0.50 and 0.70 respectively). The beta priors had parameters (1.3,5) for π_1 and (1.5,3) for π_2.

Estimates from our dependence model with the priors described above, and uniform priors for λ_D, $\lambda_{\bar{D}}$, γ_D and $\gamma_{\bar{D}}$, are given in Table 2. These estimates were very close to the true values given in Dubey *et al.* (1995). Priors that gave greater weight to the information for MAT accuracy resulted in greater posterior precision for all parameters. Inferences for the correlations are ρ_D, 0.33 (-0.28, 0.83), and $\rho_{\bar{D}}$, 0.49 (0.003, 0.74).

In contrast, an analysis with the Bayesian independence model using the same priors for η_M, θ_M, π_1 and π_2, and uniform priors for η_E and θ_E, overestimated the accuracy of the test. Three of the four 95% probability intervals did not include the true value (see Table 2).

To check the validity of the assumption of equal accuracy across populations we considered separate analyses of the two populations. For each population, the model and prior information considered were identical with the corresponding model and prior in the two-population case. Results for the two one-population analyses were consistent with the two-population analysis, indicating that our assumption was valid. The point estimates were nearly identical and intervals were slightly wider in all except one instance. In this instance, the intervals for η_M were (0.15, 0.92) in population 1 whereas they were (0.32, 0.93) for population 2. Hence, the corresponding interval for the combined data (0.63, 0.92) seemed to us to be an appropriate synthesis of the information from the two populations.

5. Simulation study

Here we attempt to assess appropriate conditions for obtaining valid inferences for the accuracy of tests when the tests are correlated, and to compare inferences from the independence and dependence models. We generated data sets for two populations by using combinations of known values of the parameters with varying sample sizes. The values x_{ijk} were generated on the basis of these known values. An analysis of the simulated data set should yield parameter estimates that are equal to the values that were used to construct the data. Selected results are presented in Tables 3–5. For example, data for simulation 1 were obtained as follows: we specified

Table 3. Selected simulation results for π_1 and π_2 using both the correlation and the conditional independence models†

Simulation	π_1 *from the following models:*			π_2 *from the following models:*		
	True	*Dependence*	*Conditional independence*	*True*	*Dependence*	*Conditional independence*
1	0.01	0.02 (0.07)	0.09 (0.08)‡	0.99	0.98 (0.06)	0.95 (0.06)‡
	0.03	0.04 (0.08)	0.10 (0.08)‡	0.97	0.97 (0.07)	0.93 (0.07)‡
	0.1	0.10 (0.12)	0.16 (0.10)‡	0.9	0.91 (0.11)	0.87 (0.09)
	0.3	0.29 (0.15)	0.33 (0.12)	0.7	0.71 (0.15)	0.69 (0.12)
2	0.01	0.01 (0.02)	0.02 (0.005)‡	0.99	0.99 (0.02)	0.98 (0.005)‡
	0.03	0.02 (0.03)	0.04 (0.008)‡	0.97	0.98 (0.03)	0.96 (0.007)‡
	0.1	0.09 (0.05)	0.11 (0.012)‡	0.9	0.91 (0.05)	0.89 (0.012)
3	0.01	0.01 (0.04)	0.02 (0.03)	0.99	0.98 (0.06)	0.94 (0.06)‡
	0.03	0.03 (0.05)	0.04 (0.04)	0.97	0.96 (0.07)	0.92 (0.08)‡
	0.1	0.10 (0.08)	0.10 (0.07)	0.9	0.90 (0.11)	0.86 (0.09)
	0.3	0.30 (0.12)	0.29 (0.11)	0.7	0.70 (0.13)	0.67 (0.12)
4	0.01	0.03 (0.10)		0.99	0.98 (0.07)	

†The sample size was 200 for simulations 1, 3 and 4, and 2000 for simulation 2. All simulations have $\rho_D = 0.5$ and $\rho_{\bar{D}} = 0.6$. Simulation 2 has priors for η_2 and θ_2 that were somewhat diffuse (fifth percentile = mode − 0.3). Simulation 4 has uniform priors for the λ- and γ-parameters and the prior for η_1 is centred at 0.85 (fifth percentile = 0.50) and the prior for θ_1 is centred at 0.80 (fifth percentile = 0.45).
‡Interval fails to cover the true value.

$(\pi_1, \pi_2, \eta_1, \eta_2, \theta_1, \theta_2, \rho_D, \rho_{\bar{D}})$ as (0.01, 0.99, 0.85, 0.95, 0.80, 0.90, 0.5, 0.6) and $n_1 = n_2 = 200$. We then calculated $(\{\eta_{ij}\}, \{\theta_{ij}\})$, e.g. $\eta_{11} = \{\rho_D\sqrt{\{\eta_1(1-\eta_1)\}}\sqrt{\{\eta_2(1-\eta_2)\}} + \eta_1\eta_2\}$, which in this case equals 0.85. Next, the expected values were calculated for the eight cells, e.g. $E(x_{111}) = n_1\{\pi_1\eta_{11} + (1-\pi_1)\theta_{11}\}$. For simulation 1, this was 19.9, which was rounded to 20. The resulting data for simulation 1 were $x_{111} = 20, x_{121} = 21, x_{211} = 2, x_{221} = 157, x_{112} = 168, x_{122} = 1, x_{212} = 21$ and $x_{222} = 11$. Most simulations used $n_1 = n_2 = 200$, unless otherwise noted.

The prevalence values used to generate the data included $\pi_1 = 0.01, 0.03, 0.1, 0.3$ with $\pi_2 = 1 - \pi_1$ in each situation. The conditional correlation values ρ_D and $\rho_{\bar{D}}$ were 0.5 and 0.6 respectively for the data in all the simulations, except where otherwise noted. Test sensitivities and specificities were moderate in simulations 1 and 4–12 ($\eta_1 = 0.85, \eta_2 = 0.95, \theta_1 = 0.8$ and $\theta_2 = 0.9$) and very high in simulation 2 ($\eta_1 = 0.98, \eta_2 = 0.99, \theta_1 = 0.98$ and $\theta_2 = 0.99$). Simulation 3 was based on moderate sensitivity and very high specificity values ($\eta_1 = 0.9, \eta_2 = 0.9, \theta_1 = 0.99$ and $\theta_2 = 0.99$).

Simulations 1–3, discussed in Section 5.1 and presented in Tables 3 and 4, involved comparisons between dependence and independence models when the priors were accurately specified. Uncertainty about test 1 characteristics was modelled with a mode equal to the value used to generate the data set and the fifth percentile equal to the mode minus 0.05. The prior distributions for the second test were centred at the true values with the fifth percentile equal to either the mode minus 0.1 (moderately peaked; simulations 1 and 3) or the mode minus 0.3 (diffuse; simulation 2). The prior distributions for the prevalences, where not explicitly described, were centred at the true value with 95th percentile for π_1 equal to the mode plus 0.1 and the fifth percentile of the π_2 prior equal to the mode minus 0.1. Results when prior distributions were inaccurate are presented in Section 5.2 and are based on simulations 5–12, which are presented in Table 5.

Table 4. Selected simulation results for η_1, η_2, θ_1 and θ_2 using both the correlation and the conditional independence models†

Simulation	*Results for the following parameters and models:*							
	$\eta_1 = 0.85$		$\eta_2 = 0.95$		$\theta_1 = 0.80$		$\theta_2 = 0.90$	
	Dependence	*Conditional independence*	*Dependence*	*Conditional independence*	*Dependence*	*Conditional independence*	*Dependence*	*Conditional independence*
1	0.85 (0.07)	0.87 (0.06)	0.95 (0.07)	0.98 (0.04)‡	0.81 (0.08)	0.84 (0.07)	0.90 (0.09)	0.97 (0.06)‡
	0.85 (0.08)	0.88 (0.07)	0.94 (0.08)	0.98 (0.04)‡	0.80 (0.08)	0.84 (0.07)	0.89 (0.10)	0.97 (0.06)‡
	0.85 (0.09)	0.88 (0.07)	0.93 (0.10)	0.98 (0.05)‡	0.80 (0.09)	0.84 (0.07)	0.89 (0.12)	0.96 (0.06)‡
	0.85 (0.09)	0.88 (0.07)	0.91 (0.12)	0.98 (0.05)	0.80 (0.10)	0.84 (0.08)‡	0.87 (0.13)	0.95 (0.07)‡
	$\eta_1 = 0.98$		$\eta_2 = 0.99$		$\theta_1 = 0.98$		$\theta_2 = 0.99$	
2	0.98 (0.02)	0.99 (0.004)‡	0.99 (0.01)	0.997 (0.002)‡	0.98 (0.02)	0.99 (0.004)‡	0.99 (0.02)	0.999 (0.001)‡
	0.97 (0.02)	0.99 (0.004)‡	0.98 (0.02)	0.997 (0.002)‡	0.97 (0.03)	0.99 (0.004)‡	0.98 (0.03)	0.999 (0.002)‡
	0.97 (0.04)	0.99 (0.005)‡	0.98 (0.04)	0.997 (0.002)‡	0.97 (0.05)	0.99 (0.004)‡	0.98 (0.05)	0.999 (0.002)‡
	$\eta_1 = 0.90$		$\eta_2 = 0.90$		$\theta_1 = 0.99$		$\theta_2 = 0.99$	
3	0.90 (0.08)	0.92 (0.06)	0.90 (0.08)	0.94 (0.06)	0.99 (0.03)	0.99 (0.02)	0.98 (0.03)	0.99 (0.03)
	0.90 (0.07)	0.92 (0.06)	0.90 (0.09)	0.94 (0.06)	0.99 (0.04)	0.99 (0.02)	0.98 (0.03)	0.99 (0.02)
	0.90 (0.08)	0.93 (0.06)	0.89 (0.10)	0.94 (0.07)	0.98 (0.05)	0.99 (0.03)	0.98 (0.05)	0.99 (0.03)
	0.90 (0.09)	0.93 (0.06)	0.89 (0.11)	0.94 (0.07)	0.98 (0.06)	0.99 (0.03)	0.97 (0.05)	0.99 (0.03)
	$\eta_1 = 0.85$		$\eta_2 = 0.95$		$\theta_1 = 0.80$		$\theta_2 = 0.90$	
4	0.86 (0.09)		0.95 (0.08)		0.82 (0.13)		0.91 (0.13)	

†The sample size was 200 for simulations 1, 3 and 4, and 2000 for simulation 2. All simulations have $\rho_D = 0.5$ and $\rho_{\tilde{D}} = 0.6$. Simulation 2 has priors for η_2 and θ_2 that were somewhat diffuse (fifth percentile = mode − 0.3). Simulation 4 has uniform priors for the λ- and γ-parameters and the prior for η_1 is centred at 0.85 (fifth percentile = 0.50) and the prior for θ_1 is centred at 0.80 (fifth percentile = 0.45).
‡Interval fails to cover the true value.

Table 5. Results of simulations with misspecified prior distributions for the prevalences†

Simulation	π_1		π_2		*Estimates (widths of 95% probability intervals) for the following parameters:*			
	True	*Estimate (width of 95% probability interval)*	*True*	*Estimate (width of 95% probability interval)*	$\eta_1 = 0.85$	$\eta_2 = 0.95$	$\theta_1 = 0.80$	$\theta_2 = 0.90$
5‡	0.01	0.03 (0.11)	0.99	0.97 (0.08)	0.85 (0.08)	0.95 (0.08)	0.81 (0.08)	0.91 (0.11)
6§	0.01	0.05 (0.09)	0.99	0.96 (0.07)	0.86 (0.07)	0.96 (0.07)	0.81 (0.08)	0.92 (0.10)
7§§	0.01	0.04 (0.10)	0.99	0.96 (0.07)	0.86 (0.07)	0.96 (0.07)	0.81 (0.08)	0.92 (0.11)
8*	0.01	0.06 (0.11)	0.99	0.96 (0.08)	0.86 (0.07)	0.97 (0.07)	0.82 (0.08)	0.92 (0.11)
9**	0.01	0.05 (0.11)	0.99	0.96 (0.08)	0.86 (0.07)	0.96 (0.07)	0.81 (0.08)	0.92 (0.11)
10††	0.01	0.004 (0.08)	0.99	0.998 (0.06)	0.85 (0.07)	0.94 (0.07)	0.80 (0.08)	0.90 (0.10)
11††	0.03	0.006 (0.10)	0.97	0.997 (0.08)	0.84 (0.08)	0.93 (0.09)	0.80 (0.11)	0.89 (0.11)
12††	0.10	0.04 (0.18)	0.90	0.97 (0.16)	0.83 (0.10)	0.90 (0.13)	0.78 (0.11)	0.86 (0.15)

†All the simulations have uniform priors for the λ- and γ-parameters and have modes for η_1 and θ_1 equal to the true parameter value and fifth percentiles equal to the mode minus 0.05. The interval fails to cover the true values for π_1 and π_2 in simulations 6 and 8.
‡ $\tilde{\pi}_1 = 0.01$; 95th percentile 0.30; $\tilde{\pi}_2 = 0.99$; fifth percentile 0.70.
§ $\tilde{\pi}_1 = 0.05$; 95th percentile 0.15; $\tilde{\pi}_2 = 0.95$; fifth percentile 0.85.
§§ $\tilde{\pi}_1 = 0.05$; 95th percentile 0.25; $\tilde{\pi}_2 = 0.95$; fifth percentile 0.75.
* $\tilde{\pi}_1 = 0.10$; 95th percentile 0.30; $\tilde{\pi}_2 = 0.90$; fifth percentile 0.70.
** $\tilde{\pi}_1 = 0.10$; 95th percentile 0.40; $\tilde{\pi}_2 = 0.90$; fifth percentile 0.60.
†† The prior for $\tilde{\pi}_1$ is beta(0.2, 0.9) and the prior for $\tilde{\pi}_2$ is beta(0.9, 0.2).

5.1. Comparison of the dependence and independence models

On the basis of the results in Tables 3 and 4, estimates obtained by using the dependence model were better than those for the independence model, especially when the true conditional correlation values were 'moderate' (ρ_D and $\rho_{\bar{D}}$ of 0.5 and 0.6 respectively). For example, see simulation 1 where many of the 95% intervals obtained with the independence model did not include the true parameter value. This finding was more pronounced in simulation 2 where the priors for η_2 and θ_2 were more diffuse (fifth percentile 0.3 below the mode) and the sample size was large ($n_i = 10000$). When the conditional correlations were low ($\rho_D = 0.1$ and $\rho_{\bar{D}} = 0.2$) the dependence model gave almost the same (accurate) results as in the moderate correlation case, whereas the independence model results were also accurate: point estimates were within 0.02 of the true parameter values for the accuracies of the test, and within 0.03 for the prevalence parameters. All 95% probability intervals, except for one, included the true parameter values (the results are not shown).

The discrepancy in results for the conditional independence and dependence models was less pronounced in the high test accuracy situations. For example when the true parameter values were $\eta_1 = 0.98, \eta_2 = 0.99, \theta_1 = 0.98, \theta_2 = 0.99, \rho_D = 0.5$ and $\rho_{\bar{D}} = 0.6$, and with $n_1 = n_2 = 200$, the results from both models were practically equivalent. In this instance, all the 95% intervals based on the independence model included the true parameter values. However, when $n_1 = n_2 = 10000$, the interval estimates based on the independence model excluded the true parameter values (see simulation 2). An 'intermediate' situation, where the true sensitivity values were moderate and the true specificity values were very high, is illustrated by simulation 3.

The dependence model estimates yielded wider 95% probability intervals than those based on the independence model. This reflected the extra uncertainty that is associated with estimating more parameters with essentially the same information. Greater precision was not necessarily a desirable trait of the independence model, because it often yielded inaccurate results when the tests were correlated.

The precision of estimates tended to decrease with increasing π_1 (and decreasing π_2) in most situations considered, and this was more pronounced for the correlation model (see, for example, simulations 1–3). However, the difference in precision of the estimates in the two models tended to be smaller when the conditional correlations were closer to 0 (the results are not shown).

5.2. Sensitivity analysis

5.2.1. Uniform distributions for λ_D, $\lambda_{\bar{D}}$, γ_D and $\gamma_{\bar{D}}$

We performed simulations (not shown) to assess the effect of using uniform informative prior distributions for (λ, γ). We found that the inferences with our simulated data sets were, for all practical purposes, identical. The loss in precision of the estimates was also minimal.

5.2.2. Diffuse but correctly specified prior distributions for η_is and θ_is

We performed several simulations (not shown) in which the prior distributions for the sensitivity and specificity parameters for test 1 were made progressively more diffuse, and with the prior distributions on the prevalences specified as indicated at the beginning of Section 5. The inferences were minimally affected, even when the fifth percentile of the respective prior distributions for η_1 and θ_1 was 0.35 below the prior mode, which was set equal to the true value, and with uniform priors for the λ- and γ-parameters (see simulation 4, Table 4).

5.2.3. Misspecified prior distributions for η_is and θ_is

We performed several simulations in which the prior distributions for η_1 had a mode, say $\tilde{\eta}_1$, that was different from the true value. For those simulations, the prior distributions for π_1, π_2 and θ_1 were still peaked at the correct values ($\pi_1 = 0.01, \pi_2 = 0.99$ and $\theta_1 = 0.80$). Uniform priors were used for (λ, γ). The true values for η_1, η_2 and θ_2 were 0.85, 0.95 and 0.90 respectively. Several priors for η_1 were used where $0.7 \leqslant \tilde{\eta}_1 \leqslant 0.9$ and the fifth percentiles were between $\tilde{\eta}_1 - 0.15$ and $\tilde{\eta}_1 - 0.05$. The 95% intervals for all the parameters included the true values (except for one interval in one simulation), and the prevalence and specificity parameters were minimally affected by the 'misspecified' priors for η_1 (the ranges of the modes of the posterior distributions were $\theta_1, 0.81, \theta_2$, 0.90–0.91, π_1, 0.02–0.03, and $\pi_2, 0.96-0.98$). Point estimates for (η_1, η_2) in the two worst case simulations considered were (0.88, 0.97) and (0.90, 0.97). In these simulations, $\tilde{\eta}_1$ was 0.90 and 0.95 and the fifth percentile for the priors on η_1 were 0.85 and 0.90. Other simulations gave better results with more diffuse priors on η_1.

5.2.4. Misspecified prior distributions for πs

Table 5 gives results of simulations where the priors used for the prevalences were misspecified, whereas those for η_1 and θ_1 were correctly specified. Misspecified priors for the prevalences substantially affected prevalence inferences when the sample size was small. For example, when the prior distributions for prevalences were not peaked at the true value, some 95% intervals excluded the true value (e.g. simulations 6 and 8). In those cases, making the prevalence priors more diffuse was a good remedy. 'Excessively' diffuse prior distributions on the prevalences sometimes resulted in inferences about sensitivity and specificity that were inaccurate (Table 5, simulations 10–12).

5.2.5. Misspecified prior distributions for η_1, θ_1 and πs

Consider a situation with true values for $(\pi_1, \pi_2, \eta_1, \eta_2, \theta_1, \theta_2, \rho_D, \rho_{\bar{D}}) = (0.05, 0.95, 0.85, 0.95, 0.80, 0.90, 0.5, 0.6)$. Prior distributions for η_1 and θ_1 were peaked at 0.03 below the true parameter value: 0.82 (fifth percentile 0.67) for η_1 and at 0.77 (fifth percentile 0.62) for θ_1; the π_is were peaked at 0.15 (95th percentile 0.25) and 0.85 (fifth percentile 0.75), and with uniform priors for (λ, γ) parameters. With $n_1 = n_2 = 200$, the estimates (and 95% intervals) were η_1, 0.87 (0.81, 0.91), η_2, 0.97 (0.92, 0.99), θ_1, 0.83 (0.77, 0.89), and θ_2, 0.94 (0.87, 0.99), π_1, 0.11 (0.06, 0.17), and π_2, 0.9 (0.85, 0.95). With $n_1 = n_2 = 2000$, the respective estimates were η_1, 0.87 (0.84, 0.89), η_2, 0.98 (0.95, 0.996), θ_1, 0.84 (0.80, 0.88), and θ_2, 0.95 (0.91, 0.99), π_1, 0.11 (0.06, 0.15), and π_2, 0.92 (0.90, 0.95). Increasing the sample size made the intervals narrower. Since the model was non-identifiable and the prior distributions were focused on incorrect values, increasing the sample size resulted in some intervals that excluded the true parameter values. In contrast, the use of the independence model yielded biased estimates and all except one of the intervals excluded the true values (the results are not shown).

6. Conclusions

We have presented a model and statistical methodology for estimating the sensitivity and specificity of two correlated diagnostic tests when there is no gold standard. The maximum-likelihood-based analysis of the toxoplasmosis data that assumed conditional independence overestimated the accuracies of the tests. When we assumed that we had very good prior information for the existing MAT, the Bayesian conditional independence model still overestimated

the parameters, whereas our conditional dependence model gave excellent results that were similar to those of the gold standard analysis. The assumption of a constant accuracy of the test in the two populations was well justified both biologically and empirically and analysis based on combined data resulted in more precise inferences.

Findings from the simulation study indicated that, when we had good prior information for either the two prevalences or the characteristics of one of the tests (and moderately good prior information for the other two parameters), the posterior inferences were very good for all model parameters, even if we had no prior information on the characteristics of the second test and the conditional correlation values. The conditional dependence model performed much better than the conditional independence model especially when the dependence was moderate. Many 95% intervals obtained by using the independence model did not include the true values, even when all the prior distributions were correctly specified. In our final simulations, where all the prior distributions for $(\pi_1, \pi_2, \eta_1, \theta_1)$ were misspecified, the dependence model yielded superior inferences. Because the model is non-identifiable, the construction of reasonable prior distributions is important. We recommend that a sensitivity analysis using different priors should be performed for all analyses using the dependence model.

In situations where the accuracies of both tests were high (near 1) the independence model seemed adequate, even with conditionally dependent tests. Consistent with our findings in the simulation study, analyses of other data (the results are not shown), when dependent tests were either highly accurate but measuring the same biological response (e.g. Hui and Walter (1980)) or were likely to be independent or weakly dependent (Georgiadis *et al.*, 1998), have shown that a conditional independence model yields results that are similar to those from the dependence model.

In conclusion, we recommend that, if the tests are only moderately accurate and measure the same biological responses, researchers should use the dependence model for their analyses. The more accurate the prior information that is available, and the larger the sample size, the better will be the posterior inferences. For analyses involving at least two populations, we recommend that the assumption of equal accuracy is checked by performing separate analyses for each population. In other situations (e.g. low dependence with correlations of less than 0.20 and high test accuracy with moderate to high correlation), the use of a conditional independence model will be preferable because the model is identified, thus reducing the dependence on the prior distribution when large samples are available. Furthermore, the use of the conditional independence model makes the specification of the prior easier because, in this case, independent beta priors can be placed directly on all parameters.

Acknowledgements

We thank Dr P. Lind and Dr J. P. Dubey for permitting the use of the *Toxoplasma gondii* data set and Dr L. Pearson for the EM code. We also thank two referees for suggestions that greatly improved the presentation. The study was funded in part by National Research Initiative Competitive Grants Program–US Department of Agriculture award 98-35204-6535.

Appendix A: The Gibbs sampler

Consider the tables of counts $\{x_{ijk}\}$ and $\{z_{ijk}\}$, as described in Section 3. The augmented likelihood is

$$L(\pi, \eta, \theta|x, z) \propto \prod_k \{\pi_k^{z_{\cdot\cdot k}}(1-\pi_k)^{x_{\cdot\cdot k}-z_{\cdot\cdot k}}\}\lambda_D^{z_{11\cdot}}(1-\lambda_D)^{z_{12\cdot}}\gamma_D^{z_{21\cdot}}(1-\gamma_D)^{z_{22\cdot}}\eta_1^{z_{1\cdot\cdot}}(1-\eta_1)^{z_{2\cdot\cdot}}$$
$$\times \lambda_{\bar{D}}^{x_{22\cdot}-z_{22\cdot}}(1-\lambda_{\bar{D}})^{x_{21\cdot}-z_{21\cdot}}\gamma_{\bar{D}}^{x_{12\cdot}-z_{12\cdot}}(1-\gamma_{\bar{D}})^{x_{11\cdot}-z_{11\cdot}}\theta_1^{x_{2\cdot\cdot}-z_{2\cdot\cdot}}(1-\theta_1)^{x_{1\cdot\cdot}-z_{1\cdot\cdot}}. \quad \text{(A.1)}$$

Define the parameter set as ω and the data as x. Then:

$$\left.\begin{aligned} z_{11k}|x,\omega &\sim \text{Bin}\left\{x_{11k}, \frac{\pi_k\eta_1\lambda_D}{\pi_k\eta_1\lambda_D + (1-\pi_k)(1-\theta_1)(1-\gamma_{\bar{D}})}\right\},\\ z_{12k}|x,\omega &\sim \text{Bin}\left\{x_{12k}, \frac{\pi_k\eta_1(1-\lambda_D)}{\pi_k\eta_1(1-\lambda_D) + (1-\pi_k)(1-\theta_1)\gamma_{\bar{D}}}\right\},\\ z_{21k}|x,\omega &\sim \text{Bin}\left\{x_{21k}, \frac{\pi_k(1-\eta_1)\gamma_D}{\pi_k(1-\eta_1)\gamma_D + (1-\pi_k)\theta_1(1-\lambda_{\bar{D}})}\right\},\\ z_{22k}|x,\omega &\sim \text{Bin}\left\{x_{22k}, \frac{\pi_k(1-\eta_1)(1-\gamma_D)}{\pi_k(1-\eta_1)(1-\gamma_D) + (1-\pi_k)\theta_1\lambda_{\bar{D}}}\right\}. \end{aligned}\right\} \tag{A.2}$$

The augmented data likelihood and the independent beta priors (using an obvious notation for the hyperparameters of the priors) yield the augmented data posterior as a product of independent beta distributions:

$$\left.\begin{aligned} \pi_k &\sim \text{beta}(a_{\pi_k} + z_{..k}, b_{\pi_k} + x_{..k} - z_{..k}),\\ \eta_1 &\sim \text{beta}(a_{\eta_1} + z_{1..}, b_{\eta_1} + z_{2..}),\\ \theta_1 &\sim \text{beta}(a_{\theta_1} + x_{2..} - z_{2..}, b_{\theta_1} + x_{1..} - z_{1..}),\\ \lambda_D &\sim \text{beta}(a_{\lambda_D} + z_{11.}, b_{\lambda_D} + z_{12.}),\\ \lambda_{\bar{D}} &\sim \text{beta}(a_{\lambda_{\bar{D}}} + x_{22.} - z_{22.}, b_{\lambda_{\bar{D}}} + x_{21.} - z_{21.}),\\ \gamma_D &\sim \text{beta}(a_{\gamma_D} + z_{21.}, b_{\gamma_D} + z_{22.}),\\ \gamma_{\bar{D}} &\sim \text{beta}(a_{\gamma_{\bar{D}}} + x_{12.} - z_{12.}, b_{\gamma_{\bar{D}}} + x_{11.} - z_{11.}). \end{aligned}\right\} \tag{A.3}$$

We now set up the Gibbs sampler as follows.

Step 1: select starting values for the parameters.
Step 2: using the initial value for ω, sample from the conditional distributions for $z_{ijk}|x,\omega$ specified in expression (A.2).
Step 3: using the newly sampled augmented data, sample from the data augmentation posteriors (A.3), to obtain a new ω.
Step 4: repeat steps 2 and 3 iteratively many times to obtain, after a 'burn-in' period, an MC sample, which can be regarded as a dependent sample from the eight-dimensional joint posterior (Tanner, 1996).

The MC sample is used to make posterior inferences as described in Tanner (1996). Using the iterates of ω we can obtain respective iterates for η_2 and θ_2 as well as the correlation parameters ρ_D and $\rho_{\bar{D}}$ from $\eta_2 = \lambda_D\eta_1 + \gamma_D(1-\eta_1)$, $\theta_2 = \gamma_{\bar{D}}(1-\theta_1) + \lambda_{\bar{D}}\theta_1$ and expression (3.1).

Appendix B: Induced informative prior distributions for $\lambda_D, \lambda_{\bar{D}}, \gamma_D, \gamma_{\bar{D}}$

Our approach is first to construct independent prior distributions for the sensitivity and specificity of both tests, the prevalences and the correlation parameters. This is essentially the approach taken by Dendukuri and Joseph (2001). We take this specification and to induce a distribution on the parameters in expression (3.3) by using the relationships

$$\left.\begin{aligned} \lambda_D &= (\delta_D + \eta_1\eta_2)/\eta_1,\\ \gamma_D &= (\eta_2 - \delta_D - \eta_1\eta_2)/(1-\eta_1),\\ \lambda_{\bar{D}} &= (\delta_{\bar{D}} + \theta_1\theta_2)/\theta_1,\\ \gamma_{\bar{D}} &= (\theta_2 - \delta_{\bar{D}} - \theta_1\theta_2)/(1-\theta_1), \end{aligned}\right\} \tag{B.1}$$

in conjunction with those in expression (3.1). This is accomplished by simulating vectors from the specified prior, and then transforming via expressions (3.1) and (B.1) to obtain an MC sample for the parameters defined in expression (3.3). We used the difference of two beta distributions as our prior for the correlations. For example, subtracting two uniform distributions yields a triangular distribution on $(-1, 1)$. By experimenting, one can obtain many possible shapes.

We sample the above prior distributions and use each set of sample values to calculate $\delta_D = \rho_D \sqrt{\{\eta_1(1-\eta_1)\}}\sqrt{\{\eta_2(1-\eta_2)\}}$ and $\delta_{\bar{D}} = \rho_{\bar{D}}\sqrt{\{\theta_1(1-\theta_1)\}}\sqrt{\{\theta_2(1-\theta_2)\}}$. The sampled values are used to obtain values for λ_D, $\lambda_{\bar{D}}$, γ_D and $\gamma_{\bar{D}}$ by using expression (B.1). The histogram of say γ_D-values is taken as the marginal prior for γ_D. We then find a beta distribution that reasonably matches the mode and the fifth or 95th percentiles of the induced histogram, which can then be used as the prior for γ_D.

References

Alonzo, T. A. and Pepe, M. (1999) Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statist. Med.*, **18**, 2987–3003.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Dendukuri, N. and Joseph, L. (2001) Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, **57**, 158–167.

Dubey, J. P., Thulliez, P., Weigel, R. M., Andrews, C. D., Lind, P. and Powell, E. C. (1995) Sensitivity and specificity of various serologic tests for detection of *Toxoplasma gondii* infection in naturally infected sows. *Am. J. Veter. Res.*, **56**, 1030–1036.

Enøe, C., Georgiadis, M. P. and Johnson, W. O. (2000) Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Prev. Vet. Med.*, **45**, 61–81.

Gardner, I. A., Stryhn, H., Lind, P. and Collins, M. T. (2000) Conditional dependence affects the diagnosis and surveillance of animal diseases. *Prev. Vet. Med.*, **45**, 107–122.

Gastwirth, J. L., Johnson, W. O. and Reneau, D. M. (1991) Bayesian analysis of screening data: application to AIDS in blood donors. *Can. J. Statist.*, **19**, 135–150.

Georgiadis, M. P., Gardner, I. A. and Hedrick, R. P. (1998) Field evaluation of sensitivity and specificity of a polymerase chain reaction (PCR) for detection of Nucleospora salmonis in rainbow trout. *J. Aquat. Anim. Hlth*, **10**, 372–380.

Hadgu, A. and Qu, Y. (1998) A biomedical application of latent class models with random effects. *Appl. Statist.*, **47**, 603–616.

Hanson, T., Johnson, W. O. and Gardner, I. A. (2003) Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold-standard. *J. Agric. Biol. Environ. Statist.*, to be published.

Hui, S. L. and Walter, S. D. (1980) Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.

Jarrett, O., Golder, M. C. and Wiejer, K. (1982) A comparison of three methods of feline leukaemia virus diagnosis. *Veter. Rec.*, **110**, 325–328.

Johnson, W. O. and Gastwirth, J. L. (1991) Bayesian inference for medical screening tests: approximations useful for the analysis of acquired immune deficiency syndrome. *J. R. Statist. Soc.* B, **53**, 427–439.

Johnson, W. O., Gastwirth, J. L. and Pearson, L. M. (2001) Screening without a gold standard: the Hui-Walter paradigm revisited. *Am. J. Epidem.*, **9**, 921–924.

Joseph, L., Gyorkos, T. W. and Coupal, L. (1995) Bayesian estimation of disease prevalence and parameters for diagnostic tests in the absence of a gold standard. *Am. J. Epidem.*, **141**, 263–272.

Mendoza-Blanco, J. R., Tu, X. M. and Iyengar, S. (1996) Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: applications to HIV screening. *Statist. Med.*, **15**, 2161–2167.

Meng, X. L. and Rubin, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Ass.*, **86**, 899–909.

Qu, Y. and Hadgu, A. (1998) A model for evaluating the sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *J. Am. Statist. Ass.*, **93**, 920–928.

Qu, Y., Tan, M. and Kutner, M. H. (1996) Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**, 797–810.

Sinclair, M. D. and Gastwirth, J. L. (1996) On procedures for evaluating the effectiveness of reinterview survey methods: application to labour force data. *J. Am. Statist. Ass.*, **91**, 961–969.

Tanner, M. A. (1996) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd edn. New York: Springer.

Torrance-Rynard, V. L. and Walter, S. L. (1997) Effects of dependent errors in the assessment of diagnostic test performance. *Statist. Med.*, **16**, 2157–2175.

Vacek, P. M. (1985) The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, **41**, 959–968.

Zhou, X.-H. (1998) Comparing accuracies of two screening tests in a two-phase study for dementia. *Appl. Statist.*, **47**, 135–147.