Evaluation of the Accuracy of Diagnostic Tests From Repeated Measurements Without a Gold Standard

Bas ENGEL, Jantien BACKER, and Willem BUIST

A model is presented to evaluate the accuracy of diagnostic tests from data from individuals that are repeatedly tested in time. Repeated measurements from three diagnostic tests for foot-and-mouth disease, applied to vaccinated and experimentally infected cattle, were analyzed. At any time the true disease status of the individuals was unknown, i.e., no gold standard was available. The model allows for correlation between repeated test results, in consequence of the underlying structure for the unknown true disease status, but also by the distribution of the test results conditional upon true disease status. The model also allows for dependence between the different diagnostic tests conditional upon true disease status. Prior information about the structure of the prevalence and the specificity of the tests was incorporated in a Bayesian analysis. Posterior inference was carried out with Markov chain Monte Carlo. Simulated data were analyzed to gain insight into the performance of the posterior Bayesian inference. The simulated data are typical for the expensive and, therefore, modestly sized infection experiments that are conducted under controlled conditions.

Key Words: Bayesian analysis; Foot and mouth disease; Infection experiment; Latent class model; Longitudinal data; MCMC; Sensitivity; Specificity.

1. INTRODUCTION

In an infection experiment, animals are experimentally infected and repeatedly tested in time, under controlled conditions. This study was motivated by data from cattle that were experimentally infected with the foot-and-mouth disease (FMD) virus. An animal was considered to be truly positive when it had started to produce specific antibodies against the FMD virus (seroconversion). However, the moment of seroconversion was unknown. All of the diagnostic tests applied to sera of the animals were imperfect and no perfect reference test was available. Consequently, at any time where animals were tested, their true disease

Dr. Bas Engel is Senior Statistician, Dr. Jantien Backer is Research Scientist and Willem Buist is Statistician, Quantitative Veterinary Epidemiology & Risk Analysis, Central Veterinary Institute of Wageningen UR, P.O. Box 65, 8200 AB Lelystad, The Netherlands. Bas Engel is presently Assistant Professor, University of Wageningen, Biometris, P.O. Box 100, 6700 AC Wageningen, The Netherlands (E-mail: Bas.Engel@wur.nl).

© 2009 The Author(s). This article is published with open access at Springerlink.com
Journal of Agricultural, Biological, and Environmental Statistics, Volume 15, Number 1, Pages 83–100
DOI: [10.1007/s13253-009-0013-y](https://doi.org/10.1007/s13253-009-0013-y)

status was unknown. The aim of this study was to evaluate the sensitivity (probability that a truly positive animal tests positive) and specificity (probability that a truly negative animal tests negative) of the diagnostic tests.

The majority of the literature on evaluation of diagnostic tests in the absence of a gold standard involves tests that are applied once to each individual, e.g., Hui and Walter (1980), Yang and Becker (1997), Qu, Tan, and Kutner (1996), Qu and Hadgu (1998), Dendukuri and Joseph (2001), Georgiades et al. (2003), Engel et al. (2006) (and references therein). When many individuals are repeatedly tested over a longer period and possibly some or all are tested at different times, these repeated measurements need to be modeled in a biologically sensible and statistically parsimonious fashion. Cook, Ng, and Meade (2000) modeled repeated diagnostic measurements with a latent Markov model and considered maximum likelihood estimation. Miglioretti (2003), although in a different context, presented a related model, employing a Bayesian analysis. Here, the model for true disease status for the FMD data is motivated by a simple mechanistic structure, employing a threshold concept. The diagnostic tests in the FMD study are based on similar biological principles. Therefore, conditional upon the true disease status (seropositive or not), the model allows for dependence between the tests. In contrast to Cook, Ng, and Meade (2000), we allow for dependence in time between repeated results of the same test, conditional upon true disease status of the individual, as well. We performed a Bayesian analysis, incorporating prior information about the prevalence structure and the specificities of the diagnostic tests, employing the Markov chain Monte Carlo algorithm (Gelman et al. 1995) that is implemented in the WinBUGS program (Spiegelhalter et al. 2003). Properties of parameter estimators (posterior medians) and 95% credible intervals (95CI) were investigated in a simulation study.

The FMD data are introduced in Section 2 and models for this data are derived in Section 3. Prior distributions are discussed in Section 4 and settings for the simulation study are presented in Section 5. Results of the analyses of the FMD data and the simulated data are presented in Section 6. Summarizing remarks about the analyses, possible extensions of the model, and a brief discussion about choice of true state and conditional (in) dependence between diagnostic tests are presented in Section 7.

2. THE FMD DATA

FMD is a contagious disease that occurs in cloven-hoofed cattle. Outbreaks can cause high financial losses, with a considerable negative social impact (Kitching, Hutber, and Thrusfield 2005). The test data that are analyzed in this article (Table 1) are from sera from 261 vaccinated cattle that were experimentally infected with the FMD virus. 75 of the animals were tested at least twice. The data are part of a larger dataset that was collected for a workshop in Brescia, Italy (Brocchi et al. 2006). Six diagnostic tests were performed under the same conditions by the same laboratory in Italy. Engel et al. (2008) performed a Bayesian analysis of the large dataset, which comprised field data as well, but only one data point was used (that was nearest to 28 days post infection) per experimentally infected

Table 1. Summary statistics of the FMD data.

Number of observations (on 3 diagnostic tests) per animal			
Number of observations (time points) per individual		Number of animals	
1		186	
2		16	
3		22	
4		1	
5		14	
6		10	
7		1	
8		10	
9		1	
Total number of animals		261	
Joint test results (summed over time points)			
Joint test results per time point	Number of times		
	Total	≤ 21 dpi	> 21 dpi
— — —	179	105	74
— — +	8	6	2
— + —	44	9	2
— + +	6	4	2
+ — —	12	6	6
+ — +	20	13	7
+ + —	20	2	18
+ + +	258	116	142
Total number of times	514	261	253

animal. Three tests clearly outperformed the other tests with respect to accuracy. In the present study, attention was restricted to these three tests.

The tests are designed to detect specific antibodies against the FMD virus. They are based on the principle that highly purified FMD vaccines contain purified inactivated FMD capsids. The nonstructural proteins are absent or almost absent. Vaccination, therefore, does not induce antibodies against the nonstructural proteins, whereas infection does. True positive state was defined as seroconversion for nonstructural proteins, i.e., presence of (specific) antibodies.

Times (days post infection, or dpi) where tests were performed, varied from 7 to 168 dpi and the 25th, 50th, and 75th percentiles were 11, 21, and 56 dpi, respectively. Most joint test results were all negative (35% of the time) or all positive (50% of the time).

3. THE MODEL

3.1. MODELING PREVALENCE

Once an animal has seroconverted, it will remain true positive, because antibodies remain in the blood during the experimental period. The moment of experimental infection

is the natural origin for time t (dpi). A latent pressure $h(t)$ is assumed that increases with time, such that an individual becomes true positive when $h(t)$ exceeds the individual's threshold u . Hence, prevalence $p(t) = P(u \leq h(t))$ at time t . Cf. Engel et al. (2008), u is assumed to follow a standard Logistic distribution:

$$\text{logit}(p(t)) = h(t) \quad \text{with } h(t) = a + b(\log(T) - \log(21)), b > 0, \quad (3.1)$$

where $T = t$, when $t \leq 21$, and $T = 21$, when $t > 21$. This reflects expert opinion that seroconversion after 21 dpi is very unlikely. Prevalence starts with $p(0) = 0$ and increases with time until a maximum P is reached at $t = 21$ dpi, where $a = \text{logit}(P)$.

An alternative three-parameter model can be motivated by a two-step process: an individual becomes seropositive with probability P , and then seroconverts when $h(t) \geq u$. With a similar choice for pressure $h(t)$, but avoiding the nonsingularity at $t = 21$ dpi:

$$\text{logit}(p(t)/P) = h(t) \quad \text{with } h(t) = a + b(\log(t) - \log(21)), b > 0. \quad (3.2)$$

Here, $a = \text{logit}(p(21)/P)$; the logit of the expected proportion of seropositives that turns seropositive before 21 dpi. We will analyze data with Equations (3.1) and (3.2).

3.2. MODELING JOINT TEST RESULTS CONDITIONAL UPON TRUE DISEASE STATUS

Initially, similar to e.g., Cook, Ng, and Meade (2000) and Schulzer, Anderson, and Drance (1991), conditional upon true status D (0/1 for a true negative/positive) of the individual, results of the same test were assumed to be independent in time. Joint test results y_1, y_2 , and y_3 (0/1 for a negative/positive test) of the three tests at any time t were modeled cf. Engel et al. (2006). For a true negative or true positive animal, respectively:

$$\begin{aligned} P(y_1 = i, y_2 = j, y_3 = k | D = 0) &= \exp(\theta_{-ijk}) / \sum_{l,m,n} \exp(\theta_{-lmn}), \\ P(y_1 = i, y_2 = j, y_3 = k | D = 1) &= \exp(\theta_{+ijk}) / \sum_{l,m,n} \exp(\theta_{+lmn}), \end{aligned} \quad (3.3)$$

where

$$\begin{aligned} \theta_{-ijk} &= -i\delta_{-1} - j\delta_{-2} - k\delta_{-3} + ij\varepsilon_{-12} + ik\varepsilon_{-13} + jk\varepsilon_{-23}, \\ \theta_{+ijk} &= (i-1)\delta_{+1} + (j-1)\delta_{+2} + (k-1)\delta_{+3} + (i-1)(j-1)\varepsilon_{+12} \\ &\quad + (i-1)(k-1)\varepsilon_{+13} + (j-1)(k-1)\varepsilon_{+23}, \quad i, j, k = 0, 1. \end{aligned}$$

Because all tests aim for detection of antibodies, while seroconversion is the true positive state of interest, (conditional) probabilities for test results y are assumed to depend on time through true disease status D only. “Main effects” δ_{-} and δ_{+} are associated with specificities and sensitivities and “interactions” ε_{-} and ε_{+} are associated with dependence between tests, conditional upon true disease status $D = 0$ or 1. For instance:

$$\text{logit}(P(y_2 = 1 | D = 1, y_1 = i, y_3 = k)) = \delta_{+2} + (i-1)\varepsilon_{+12} + (k-1)\varepsilon_{+23}. \quad (3.4)$$

So, δ_{+2} is the logit of the sensitivity of test 2, conditional upon the correct test results $y_1 = y_3 = 1$ of the other tests. Interactions ε are log odds ratios, for instance:

$$\varepsilon_{+23} = \log \left(\frac{P(y_2 = 1|D = 1, y_1 = i, y_3 = 1)}{P(y_2 = 0|D = 1, y_1 = i, y_3 = 1)} \bigg/ \frac{P(y_2 = 1|D = 1, y_1 = i, y_3 = 0)}{P(y_2 = 0|D = 1, y_1 = i, y_3 = 0)} \right). \quad (3.5)$$

This is the logarithm of the cross product ratio of tests 2 and 3, conditional upon result $y_1 = i$ of test 1. Because in Equation (3.3) attention is restricted to two-factor interactions only, Equation (3.5) does not depend on y_1 . Conditional independence between tests corresponds to interactions $\varepsilon = 0$. Positive interaction terms ε correspond to positive dependence between tests, i.e., a tendency for concordant observations.

The sensitivity (α) and specificity (β) of a test can be derived from the conditional joint probabilities, e.g., for test 1 from Equation (3.3) the sensitivity is

$$\alpha_1 = \sum_{j,k} \exp(\theta_{+1jk}) \bigg/ \sum_{l,m,n} \exp(\theta_{+lmn}). \quad (3.6)$$

3.3. MODELING SUBJECT SPECIFIC SENSITIVITY AND SPECIFICITY

To incorporate dependence between repeated measurements, conditional upon true disease status D , a notion of subject specific sensitivity and specificity was introduced by adding normally distributed random animal effects ν to the main effects δ in the multinomial probabilities of Equation (3.3). Conditional upon D , there are two (potential) sources of dependence between test results y : global correlation due to (unobserved) characteristics of the individual that affect a test in a similar fashion at all times, associated with the random ν effects, and local residual correlation between tests at each time, associated with the interaction terms ε . Separate animal effects $\nu_- = (\nu_{-1}, \nu_{-2}, \nu_{-3})'$ and $\nu_+ = (\nu_{+1}, \nu_{+2}, \nu_{+3})'$ were introduced for tests 1, 2, and 3, with common variances σ_0^2 and σ_1^2 for true negatives and true positives, respectively. Both independent and correlated ν effects for tests within individuals, with common correlations ρ_0 and ρ_1 for true negatives and positives, were considered. The model is graphically represented in Figure 1.

Marginal probabilities are obtained by integration over random effects ν . For modestly sized variances σ_0^2 and σ_1^2 , this approximately involves shrinkage, e.g., the following expressions may be compared with Equations (3.4) and (3.5):

$$\begin{aligned} & \text{logit}(P(y_2 = 1|D = 1, y_1 = i, y_3 = k)) \\ & \approx \xi_1(\delta_{+2} + (i - 1)\varepsilon_{+12} + (k - 1)\varepsilon_{+23}) \quad \text{and} \\ & \log \left(\frac{P(y_2 = 1|D = 1, y_1 = i, y_3 = 1)}{P(y_2 = 0|D = 1, y_1 = i, y_3 = 1)} \bigg/ \frac{P(y_2 = 1|D = 1, y_1 = i, y_3 = 0)}{P(y_2 = 0|D = 1, y_1 = i, y_3 = 0)} \right) \\ & \approx \xi_1 \varepsilon_{+23}, \end{aligned} \quad (3.7)$$

where $\xi_1 = 1/\sqrt{1 + \sigma_1^2/c^2}$ and $c = \frac{15}{16}\pi/\sqrt{3}$ (see e.g. Engel, Buist, and Visscher 1995). Equation (3.7) will not be used in the actual calculations, but will be useful to motivate priors for main effects δ and interactions ε in Section 4.

Figure 1. A graphical representation of the model.

3.4. CALCULATING MARGINAL SENSITIVITY AND SPECIFICITY

Let the subject specific sensitivity of a random individual, say for test 1, be $\alpha_1(\mathbf{v}_+)$, where this sensitivity depends on the individual through its random effects $\mathbf{v}_+ = (v_{+1}, v_{+2}, v_{+3})'$. The marginal sensitivity is: $\alpha_1 = E(\alpha_1(\mathbf{v}_+))$ (and similarly the marginal specificity is $\beta_1 = E(\beta_1(\mathbf{v}_-))$). The average of the subject specific sensitivities $\alpha_1(\mathbf{v}_+)$ of the individuals offers an approximation for α_1 that can be simply included in the Markov chain Monte Carlo (MCMC) process:

$$\hat{\alpha}_{simple,1} = \overline{\alpha_1(\mathbf{v}_+)}. \quad (3.8)$$

A more accurate, but computer intensive approximation can be derived by Gauss–Hermite quadrature (Dahlquist and Björck 1974):

$$\hat{\alpha}_{\text{GH},1} = \sum_{i=-n}^n \sum_{j=-n}^n \sum_{k=-n}^n \pi^{-3/2} w_i w_j w_k a_1(\mathbf{v}_{ijk}). \quad (3.9)$$

Here, by spectral decomposition, $\mathbf{v}_{ijk} = \sqrt{2}\mathbf{S}\mathbf{\Lambda}^{1/2}\mathbf{x}_{ijk}$, $\mathbf{\Lambda} = \sigma_1^2 \text{diag}(1 + 2\rho_1, 1 - \rho_1, 1 - \rho_1)$,

$$\mathbf{S} = \begin{pmatrix} \frac{1}{3}\sqrt{3} & \frac{1}{2}\sqrt{2} & \frac{1}{5}\sqrt{5} \\ \frac{1}{3}\sqrt{3} & -\frac{1}{2}\sqrt{2} & \frac{1}{5}\sqrt{5} \\ \frac{1}{3}\sqrt{3} & 0 & -\frac{2}{5}\sqrt{5} \end{pmatrix},$$

$\mathbf{x}_{ijk} = (x_i, x_j, x_k)'$, abscissas x_l and weights w_l , $l = -n, \dots, n$, are derived from Hermite polynomials (Zhang and Jin 1996), and from Equations (3.3) and (3.6):

$$\begin{aligned} \alpha_1(\mathbf{v}_{ijk}) &= \sum_{r,s} \left(\exp(\theta_{+lrs}(\mathbf{v}_{ijk})) / \sum_{l,m,n} \exp(\theta_{+lmn}(\mathbf{v}_{ijk})) \right), \\ \theta_{+lmn}(\mathbf{v}_{ijk}) &= (l-1)(\delta_{+1} + v_{ijk,1}) + (m-1)(\delta_{+2} + v_{ijk,2}) + (n-1)(\delta_{+3} + v_{ijk,3}) \\ &\quad + (l-1)(m-1)\varepsilon_{+12} + (l-1)(n-1)\varepsilon_{+13} + (m-1)(n-1)\varepsilon_{+23}, \\ l, m, n &= 0, 1. \end{aligned}$$

Every fifth value of the δ 's, ε 's, σ_0^2 , σ_1^2 , ρ_0 , and ρ_1 (but not the v 's) was taken from the Markov chain of length 50,000, after a burn in of 5,000. Each set of simulated parameter values, $\hat{\alpha}_{\text{GH},i}$ and $\hat{\beta}_{\text{GH},i}$, $i = 1, 2, 3$, was evaluated by Equation (3.9) with $n = 20$ ($n = 30$ yielded the same results), providing samples of size 10,000 from the posteriors.

4. PRIOR DISTRIBUTIONS

Expert opinion was expressed about the time to seroconversion and about the maximum prevalence, involving parameters a and P (Table 2). In Equation (3.1), when seroconversion occurs, it will occur before 21 dpi and in Equation (3.2), more cautiously, largely before 21 dpi. Engel et al. (2008; table 2, vaccinated animals of group G4), employing Equation (3.1), used an informative prior for the maximum prevalence. However, the prior and posterior were found to be markedly different. Therefore, in the present study diffuse priors were used for a in Equation (3.1) and $\text{logit}(P)$ in Equation (3.2).

A diffuse prior from 0 to 21 dpi was specified for the median time m , where $p(m) = P/2$ (Table 2). The prior for coefficient b followed from $b = (\text{logit}(P/2) - a)/(\log(m) - \log(21))$ for Equation (3.1), and $b = -a/(\log(m) - \log(21))$ for Equation (3.2). For Equation (3.1), the priors for a and m (Table 2) and the induced prior for b were used. For Equation (3.2), to ensure that b was positive valued, the induced prior was approximated by a lognormal distribution (Table 2) and used together with the prior for a .

Priors for σ_0 and σ_1 followed from consideration of log odds ratios between individuals with random effects $\mathbf{u}$ and $\mathbf{u}'$, e.g., from Equation (3.4):

$$\begin{aligned} &\text{logit}(P(y_1 = 1 | D = 1, y_2 = y_3 = 1, \mathbf{v} = \mathbf{u})) \\ &\quad - \text{logit}(P(y_1 = 1 | D = 1, y_2 = y_3 = 1, \mathbf{v} = \mathbf{u}')) = u_1 - u'_1. \end{aligned}$$

Table 2. Prior distributions for the FMD data and for the simulated data. In the simulation $P = 0.9$ or 0.4 . From left to right, for the FMD data and the simulated data, the 5th/50th/95th percentiles of the priors for P are 0.09/0.50/0.91, 0.80/0.90/0.95, and 0.31/0.40/0.50. The 5th/50th/95th percentiles of the prior of m are 1.2/10.5/19.7.

Parameter		FMD data	Simulated data	
			True $P = 0.9$	True $P = 0.4$
a	Eq. (3.1)	$N(0, 0.5)^{(1)}$	$N(2.1654, 4.4584)$	$N(-0.4055, 16.46)$
	Eq. (3.2)	$N(2.1654, 4.4584)$	$N(2.1654, 4.4584)$	$N(2.1654, 4.4584)$
m	Eq. (3.1)	$21 * \text{Beta}(1.1, 1.1)$	$21 * \text{Beta}(1.1, 1.1)$	$21 * \text{Beta}(1.1, 1.1)$
$\text{logit}(P)$	Eq. (3.2)	$N(0, 0.5)$	$N(2.1654, 4.4584)$	$N(-0.4055, 16.46)$
b	Eq. (3.1)	Induced by a and m	Induced by a and m	Induced by a and m
	Eq. (3.2)	$\text{Lognormal}(1.3, 0.67)^{(2)}$	$\text{Lognormal}(1.3, 0.67)$	$\text{Lognormal}(1.3, 0.67)$
δ_{-1}		$N(2.85, 1.13)$	$N(2.85, 1.13)$	$N(2.85, 1.13)$
δ_{-2}		$N(2.85, 1.13)$	$N(2.85, 1.13)$	$N(2.85, 1.13)$
δ_{-3}		$N(2.85, 1.13)$	$N(2.85, 1.13)$	$N(2.85, 1.13)$
δ_{+1}		$N(0, 0.5)$	$N(0, 0.5)$	$N(0, 0.5)$
δ_{+2}		$N(0, 0.5)$	$N(0, 0.5)$	$N(0, 0.5)$
δ_{+3}		$N(0, 0.5)$	$N(0, 0.5)$	$N(0, 0.5)$
all ε		$\text{Gamma}(0.48, 1)^{(3)}$	$\text{Gamma}(0.48, 1)$	$\text{Gamma}(0.48, 1)$
σ_0, σ_1		$\text{Uniform}(0, 0.77)$	—	—
ρ_0, ρ_1		$\text{Uniform}(0, 1)$	—	—

⁽¹⁾ $N(\mu, \tau)$ is a Normal distribution with mean μ and precision τ (variance $1/\tau$).

⁽²⁾ $\text{Lognormal}(\mu, \tau)$ is $N(\mu, \tau)$ after log transformation.

⁽³⁾ $\text{Gamma}(a, b)$ is a Gamma distribution with mean a/b and variance a/b^2 .

Choosing $\pm \log(6)$ for the fifth and 95th percentiles $\pm 1.645\sqrt{2}\sigma_1$, an upper bound for a Uniform prior (Gelman 2006) for σ_1 (and σ_0) of 0.77 was obtained. Because shrinkage factors, such as ξ_1 in Equation (3.7), will be larger than 0.91, priors for main effects δ (Table 2) were chosen cf. Engel et al. (2006, 2008): diffuse priors for effects δ_+ and more informative priors for effects δ_- (because producers of tests aim for high specificity). Diffuse priors were used for interactions ε and correlations ρ (Table 2).

5. SIMULATION

Data were simulated with Mathematica (Wolfram Research, Inc. 2007) and analyzed with WinBUGS (Spiegelhalter et al. 2003). Data were mainly simulated for the prevalence structure in Equation (3.2), for three diagnostic tests, for 20 individuals with 10 times per individual (4, 6, ..., 22 dpi), for 10 individuals with 20 time points (3, 4, ..., 22 dpi) and for 10 individuals with 10 time points (4, 6, ..., 22 dpi) (Table 3). Subject specific ν effects were not included in the simulation. In the simulation, test 3 was independent of tests 1 and 2, but in the analysis possible dependence between all tests was assumed. Values $P = 0.4$ and 0.9 were considered typical for maximum prevalence for vaccinated and nonvaccinated animals, respectively. Relatively high (0.95, 0.75, 0.85) and modest to low (0.70, 0.40, 0.60) sensitivities were considered. Specificities were chosen equal to 0.97.

Table 3. Parameter values in the simulation.

Parameter	Value		
P maximum prevalence	0.9	0.9	0.4
$a = \text{logit}(p(21)/P)$	2.197	2.197	2.197
m time (dpi) where prevalence is $P/2$	10	10	10
b	2.962	2.962	2.962
α_1 sensitivity of test 1	0.95	0.70	0.70
α_2 sensitivity of test 2	0.75	0.40	0.40
α_3 sensitivity of test 3	0.85	0.60	0.60
β_1, β_2 , and β_3 specificity of tests 1, 2, and 3	0.97	0.97	0.97
$\delta_{+1} = \text{logit}(P(y_1 = 1 D = 1, y_2 = 1))$	3.540	1.865	1.865
$\delta_{+2} = \text{logit}(P(y_2 = 1 D = 1, y_1 = 1))$	1.193	-0.0207	-0.0207
$\delta_{+3} = \text{logit}(\alpha_3)$	1.735	0.405	0.405
$\delta_{-1} = \text{logit}(P(y_1 = 0 D = 0, y_2 = 0))$	3.505	3.505	3.505
$\delta_{-2} = \text{logit}(P(y_2 = 0 D = 0, y_1 = 0))$	3.505	3.505	3.505
$\delta_{-3} = \text{logit}(\beta_3)$	3.476	3.476	3.476
$\varepsilon_{+12} = \log\left(\frac{P(y_1=1, y_2=1 D=1)P(y_1=0, y_2=0 D=1)}{P(y_1=0, y_2=1 D=1)P(y_1=1, y_2=0 D=1)}\right)$	$1.504 = \log(4.5)$	1.504	1.504
$\varepsilon_{-12} = \log\left(\frac{P(y_1=1, y_2=1 D=0)P(y_1=0, y_2=0 D=0)}{P(y_1=0, y_2=1 D=0)P(y_1=1, y_2=0 D=0)}\right)$	$0.693 = \log(2)$	0.693	0.693
$\varepsilon_{-13}, \varepsilon_{+13}, \varepsilon_{-23}, \varepsilon_{+23}$	0	0	0
Simulation results presented in Table:	5(a)	5(b)	5(c), 6(a), 6(b)

6. RESULTS

The FMD data were analyzed with prevalence structures using Equations (3.1) and (3.2). Models fitted to the data comprised different choices for $\sigma_0^2, \sigma_1^2, \rho_0$, and ρ_1 for the random effects ν , a model without interaction terms ε and a model ignoring the repeated measurements (as if the data were from as many individuals as there were time points). Chains of length 50,000, after a burn in of 5,000, on average required some eight hours computing time on a desktop computer with Intel Xeon processors, 3.60 GHz CPU and 3Gb RAM.

In the simulation study, simulations were initially run with prevalence structures using Equations (3.1) and (3.2), with similar results. Results after 50,000 and 100,000 MCMC iterations were basically the same. Subsequently, more extensive simulations were performed with Equation (3.2), employing chains of length 50,000, after a burn in of 5,000. Results of these simulations will be presented below. Although the simulated datasets were smaller and more regular than the FMD data, computing time was still substantial: 55,000 iterations required some 25 minutes or, occasionally, several hours. Some 2%–5% of the simulations terminated with an error message from WinBUGS. In an analysis of real data this is often resolved by a judicious choice of starting values. In the simulation study, simulations with error messages were discarded. Label switching (Redner and Walker 1984) produces parameter values with the same likelihood (but lower prior probability). The algorithm may get “stuck” in the neighborhood of such a “mirror image” (see also Engel et al. 2006), yielding unacceptably low specificities (and sensitivities) after 55,000 iterations. Simulations that did not comply with the condition $\alpha + \beta > 1$ for all tests were

marked as potential cases of label switching and discarded. This occurred infrequently, except when prevalence and sensitivities were relatively low (Table 5(c)) and some 3% of the simulations had to be discarded.

6.1. RESULTS FOR THE FMD DATA

Results from the Gauss–Hermite integration of Equation (3.9) are reported in Table 4. Similar results were found with the simple average of Equation (3.8). Results for various models fitted to the data, employing the prevalence structure found in Equation (3.1) (Table 4(a)), were similar, except for the model without interaction ($\varepsilon = 0$) and the model where the repeated measurements structure was ignored. Random effects v for additional dependence, conditional upon true disease status, had no apparent effect on the results. Only when interactions ε were assumed to be 0, was the correlation ρ_1 noticeably increased, thus also accounting for conditional dependence between tests for true positives.

With Equation (3.2) (Table 4(b)), for some of the models considered, starting values had to be chosen with care, to avoid label switching. Although results using the prevalence structures of Equations (3.1) and (3.2) were similar (considering the width of the associated 95CI), sensitivities were noticeably higher for Equation (3.2). Omission of interaction terms ε for Equation (3.2) had a less pronounced effect on correlation ρ_1 than for Equation (3.1). When the repeated measurement structure was ignored, practically the same results were found for Equation (3.2) as before. This is consistent with the simulation results in Section 6.2.

We could argue that in Equation (3.2), if we have high confidence in expert opinion, the prior for a should have been focused on higher values, in order to make prior information for Equations (3.1) and (3.2) more comparable. However, a $N(3.4, 2.67)$ distribution for a that focuses on higher values for $p(21)/P$ around 0.95, produced nearly the same results as in Table 4(b).

Inspection of the sum of the test results, per individual and per time of measurement, suggested that 7 individuals might have seroconverted after 21 dpi. Without these individuals, estimated sensitivities (and specificities) with Equations (3.1) and (3.2) (not shown) were quite similar. When the repeated measurement structure is ignored, it will no longer be apparent that after 21 dpi some animals have three negative test results followed by three positive test results. Indeed, in the latter instance, results under Equation (3.1) or Equation (3.2), as reported in the last column of Table 4, were practically the same (and similar to the results obtained without the 7 suspect individuals). When times $t > 21$ were replaced by 21 dpi in the analysis with Equation (3.2), effectively bringing the “offending” observations of the 7 animals forward in time, prevalence at 21 dpi was increased, sensitivity was decreased and specificity was increased, and results were virtually identical to the results with Equation (3.1). When time t was truncated at 28 dpi, results with Equation (3.2) hardly changed from the results without truncation as reported in Table 4(b).

Apparently, the greater flexibility of Equation (3.2) has some impact on the results in connection with the 7 animals that possibly did not seroconvert before 21 days. It is hard to discriminate between Equations (3.1) and (3.2) in terms of goodness of fit. Considering the width of the credible intervals, the differences between Equations (3.1) and (3.2) are

Table 4. Results of analyses of the FMD data. Estimates (posterior medians) and 95% credible intervals (in parentheses) for different models. The full model comprises the threshold concept and ν effects.

Parameter	Full model	Full model, but: $\rho_0 = \rho_1 = 0$	Full model, but without ν effects, i.e. $\sigma_0^2 = \sigma_1^2 = 0$	Full model, but without interaction: $\varepsilon = 0$	Independence between time points (ignoring repeated measurements)
(a) Results with prevalence structure (1)					
α_1	0.93 (0.88, 0.96)	0.92 (0.88, 0.96)	0.93 (0.89, 0.96)	0.95 (0.92, 0.97)	0.96 (0.93, 0.98)
α_2	0.87 (0.82, 0.91)	0.87 (0.81, 0.91)	0.87 (0.83, 0.91)	0.89 (0.85, 0.92)	0.91 (0.87, 0.94)
α_3	0.87 (0.82, 0.91)	0.87 (0.82, 0.91)	0.86 (0.82, 0.90)	0.90 (0.86, 0.93)	0.91 (0.87, 0.94)
β_1	0.98 (0.94, 0.99)	0.98 (0.94, 0.99)	0.97 (0.94, 0.99)	0.97 (0.94, 0.99)	0.95 (0.91, 0.99)
β_2	0.96 (0.91, 0.98)	0.96 (0.92, 0.98)	0.96 (0.92, 0.98)	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)
β_3	0.96 (0.92, 0.99)	0.96 (0.91, 0.99)	0.96 (0.92, 0.98)	0.96 (0.92, 0.98)	0.96 (0.92, 0.98)
P	0.74 (0.67, 0.80)	0.75 (0.68, 0.82)	0.74 (0.64, 0.79)	0.73 (0.67, 0.79)	0.64 (0.59, 0.69)
σ_0^2	0.20 (0.00, 0.57)	0.17 (0.00, 0.56)	—	0.29 (0.01, 0.58)	—
σ_1^2	0.54 (0.34, 0.59)	0.54 (0.35, 0.59)	—	0.56 (0.42, 0.59)	—
ρ_0	0.47 (0.02, 0.94)	—	—	0.52 (0.04, 0.98)	—
ρ_1	0.45 (0.03, 0.94)	—	—	0.85 (0.49, 0.99)	—
(b) Results with prevalence structure (2)					
α_1	0.96 (0.93, 0.98)	0.96 (0.92, 0.98)	0.97 (0.94, 0.99)	0.97 (0.94, 0.99)	0.96 (0.93, 0.98)
α_2	0.91 (0.87, 0.95)	0.90 (0.86, 0.94)	0.92 (0.88, 0.95)	0.91 (0.87, 0.94)	0.92 (0.87, 0.95)
α_3	0.91 (0.87, 0.94)	0.90 (0.86, 0.94)	0.91 (0.87, 0.94)	0.92 (0.88, 0.95)	0.92 (0.87, 0.95)
β_1	0.96 (0.90, 0.98)	0.96 (0.92, 0.99)	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)
β_2	0.95 (0.90, 0.98)	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)	0.95 (0.90, 0.97)	0.95 (0.91, 0.98)
β_3	0.95 (0.89, 0.98)	0.95 (0.91, 0.98)	0.95 (0.91, 0.98)	0.95 (0.92, 0.98)	0.96 (0.92, 0.98)
P	0.79 (0.70, 0.90)	0.77 (0.69, 0.86)	0.77 (0.70, 0.85)	0.78 (0.69, 0.86)	0.69 (0.62, 0.78)
σ_0^2	0.30 (0.01, 0.58)	0.25 (0.01, 0.58)	—	0.33 (0.01, 0.58)	—
σ_1^2	0.54 (0.36, 0.59)	0.54 (0.33, 0.59)	—	0.54 (0.37, 0.59)	—
ρ_0	0.60 (0.06, 0.98)	—	—	0.55 (0.01, 0.95)	—
ρ_1	0.51 (0.04, 0.94)	—	—	0.69 (0.13, 0.95)	—

not marked. However, the difference between e.g., 0.97 and 0.93 for sensitivity is practically important and estimated sensitivities tend to get more emphasis than their associated credible intervals. We tend to favor Equation (3.2), because it allows for greater flexibility in incorporating prior information about the prevalence structure.

6.2. RESULTS FOR THE SIMULATED DATA

All results are for the prevalence structure found in Equation (3.2) and with specificity $\beta_1 = \beta_2 = \beta_3 = 0.97$.

High Prevalence and High Sensitivity ($P = 0.9$, $\alpha_1 = 0.95$, $\alpha_2 = 0.75$, $\alpha_3 = 0.85$)

In the left-hand side of Table 5(a) results are shown for the repeated measurements model employing the threshold concept. Clearly, the posterior medians as estimates for sensitivity, specificity, and maximum prevalence are performing well: bias and root mean squared error (RMSE) are small. The coverage probability and average width of the 95CI

Table 5. Results of analyses of simulated data with prevalence structure of Equation (3.2). 20 animals with 10 time points each (4, 6, ..., 22 dpi). Bias (% bias), width of 95% credible interval (% CI) and root mean squared error (% RMSE), as a percentage of the true parameter value, and percentage coverage probability (% CP) of the 95% credible interval. Results with the repeated measurements structure for prevalence (by the threshold concept) and results ignoring the repeated measurements structure.

Parameter	True value	With repeated measurements structure				Ignoring repeated measurements			
		% bias	% CI	% CP	% RMSE	% bias	% CI	% CP	% RMSE
(a) Relatively high prevalence and high sensitivity (1225 simulated datasets)									
α_1	0.95	-1.3	10	94	3	-0.8	11	94	2
α_2	0.75	-0.5	22	96	6	-0.3	24	96	6
α_3	0.85	-0.8	16	95	4	-1.1	18	95	4
β_1	0.97	-0.6	7	97	2	-0.8	9	97	2
β_2	0.97	-0.5	7	98	2	-0.5	8	98	2
β_3	0.97	-0.6	7	98	2	-1.3	9	98	2
a	2.197	-0.7	70	100	10	0.5	63	97	15
b	2.962	2.6	71	97	16	4.0	62	93	17
P	0.90	-0.3	18	100	17	-0.3	16	100	3
(b) Relatively high prevalence and moderate to low sensitivity (1294 simulated datasets)									
α_1	0.70	-0.4	27	97	6	1.3	37	96	9
α_2	0.40	0.2	48	95	12	3.0	57	95	15
α_3	0.60	-0.3	32	96	8	-0.8	39	96	9
a	2.197	-0.6	72	100	9	-0.7	72	100	11
b	2.962	3.4	74	100	17	8.0	80	95	20
P	0.90	-0.2	18	100	2	-0.4	18	100	2
(c) Relatively low prevalence and moderate to low sensitivity (1410 simulated datasets)									
α_1	0.70	-2.0	40	95	11	2.1	61	96	16
α_2	0.40	0.0	71	95	19	8.4	98	94	25
α_3	0.60	-1.7	48	95	13	-5.4	67	96	17
a	2.197	-1.0	80	100	6	-2.6	81	95	6
b	2.962	6.1	113	96	32	15.9	157	91	48
P	0.40	0.4	52	100	5	-1.6	50	100	7

are satisfactory, except for the intervals for parameters a (or equivalently p/P) and P , that were too conservative.

Results for the interaction terms ε (not shown) were poor. For instance, for tests 1 and 2, for true positives, bias and RMSE of ε_{+12} , expressed as percentages of the true parameter value, were -73.9% and 76.5% , respectively, and the coverage probability of the 95CI was only 58%. Because these experiments are quite expensive, they are modestly sized and apparently offer little information about conditional dependence between tests. In particular it was hardly possible to establish that test 3 is independent of tests 1 and 2. The large negative bias for the interaction terms is probably a consequence of the prior gamma distribution that favors small values. Another prior might possibly produce smaller bias, but RMSEs would still be poor.

In the right-hand side of Table 5(a), results are shown for the same datasets, but now the repeated measurements structure was completely ignored in the analysis. Results are similar to the left-hand side of the table. For this parameter configuration, there was little gain in accounting for the repeated measurements structure of the data.

Because similar results for specificity were found for the other parameter configurations, in the following we will focus on sensitivity.

High Prevalence and Modest Sensitivity ($P = 0.9, \alpha_1 = 0.70, \alpha_2 = 0.40, \alpha_3 = 0.60$)

Posterior medians in Table 5(b) are practically unbiased. RMSEs are somewhat increased and 95CI are wider, but still quite acceptable. Although lower sensitivities offer more opportunity for detection of dependence between tests for true positives, results for interaction terms (not shown) were still poor: for ε_{+12} percentage bias and RMSE were -33.9% and 45.5% , respectively and the 95CI showed 83% coverage. Again results in the left- and right-hand sides of the table are quite similar, although 95CI from the analysis ignoring the repeated measurements are perceptibly wider.

Modest Prevalence and Modest Sensitivity ($P = 0.4, \alpha_1 = 0.70, \alpha_2 = 0.40, \alpha_3 = 0.60$)

Posterior medians in Table 5(c) are still practically unbiased. The average width of the 95CI has markedly increased, but generally coverage probabilities are acceptable. The RMSE has increased, but estimated values (posterior medians) will e.g., still show that test 3 has poor sensitivity. There is a marked difference with the analysis ignoring the repeated measurements, the latter showing distinctly wider 95CI and higher RMSE.

For interaction ε_{+12} , bias, RMSE and coverage of the 95CI were -57.2% , 63.4% , and 76% , respectively.

For each parameter with true value θ and estimates $\hat{\theta}$ and $\hat{\theta}'$ with and without the repeated measurements structure, differences $(\hat{\theta} - \theta)^2 - (\hat{\theta}' - \theta)^2$ were analyzed with Wilcoxon's signed rank test. All RMSEs for the corresponding parameters on the left- and right-hand sides of Table 5(c) were significantly ($P < 0.05$) different, except for specificity β_2 . This was also true for Table 5(b), although differences were less marked. Even in Table 5(a) differences were mostly significant, albeit of little importance.

In the simulation, test 3 was conditionally independent of tests 1 and 2, but data were analyzed assuming possible conditional dependence between all tests. Suppose that, in practice, test 3 relates to a different biological principle than tests 1 and 2. We might assume that all interactions with test 3 are negligible. The simulated data of Table 5(c) were reanalyzed with $\varepsilon_{-13} = \varepsilon_{-23} = \varepsilon_{+13} = \varepsilon_{+23} = 0$. Although computing time was reduced by some 50% and hardly any simulations had to be discarded, results (not shown) were similar to the results in Table 5(c). So, the extra information about test 3 did not result in any appreciable gain in accuracy of estimated sensitivities (or specificities).

Additional Simulations With Varying Numbers of Animals and Times per Animal

Table 6 offers a modest (because of the excessive computer time) impression of the effects of varying numbers of animals and time points per animal. For 10 animals with 20 time points each (Table 6(a)), RMSE is higher than in Table 5(c) for 20 individuals with 10 time points each. Results on the right-hand side of Table 6(a), where the repeated measurement structure is ignored, are relatively poorer than results in the right-hand side of Table 5(c). This confirms that for this configuration it is worthwhile to properly account for the repeated measurements (at the least in the structure for true disease status).

Comparing results with 10 animals with 10 times each (Table 6(b)), with 10 animals and 20 times each (Table 6(a)), and 20 animals and 10 times each (Table 5(c)), we conclude that employing more animals is more profitable than testing on more time points. With the exception of sensitivity α_2 , results in Table 6(a) and 6(b), particularly for parameters a , b , and P , are not markedly different. When the main interest is in studying the development

Table 6. More results of analyses of simulated data with prevalence structure in Equation (3.2) and varying numbers of time points and numbers of animals.

Parameter	True value	With repeated measurements structure				Ignoring repeated measurements			
		% bias	% CI	% CP	% RMSE	% bias	% CI	% CP	% RMSE
(a) Relatively low prevalence and moderate to low sensitivity (977 simulated datasets)									
10 animals with 20 time points each (3, 4, . . . , 22 dpi)									
α_1	0.70	−2.8	41	94	14	−1.2	62	90	23
α_2	0.40	−0.7	74	96	20	5.3	98	92	29
α_3	0.60	−1.8	50	95	15	−8.1	68	93	23
a	2.197	−0.9	81	100	5	−0.5	81	100	8
b	2.962	12.4	166	96	48	26.5	208	86	67
P	0.40	0.5	54	95	4	0.8	50	98	10
(b) Relatively low prevalence and moderate to low sensitivity (957 simulated datasets)									
10 animals with 10 time points each (4, 6, . . . , 22 dpi)									
α_1	0.70	−5.0	55	96	16	−2.8	75	96	22
α_2	0.40	0.7	97	96	26	7.6	127	94	34
α_3	0.60	−3.3	66	96	18	−8.2	87	95	25
a	2.197	−0.7	82	100	4	−1.5	82	100	5
b	2.962	11.1	176	95	48	27.0	293	95	67
P	0.40	0.6	54	95	4	−0.2	53	100	7

of the disease in time, there is hardly any gain in efficiency when a grid of more densely spaced time points is used.

7. DISCUSSION

The model in this article, with the prevalence structure found in Equations (3.1) or (3.2), can be used for analysis of test data from infection experiments with a disease with an irreversible true positive status. We have shown in a simulation study that even for the expensive, and, therefore, modestly sized, infection experiments that are conducted under controlled conditions, proper modeling of the repeated measurements is worthwhile for moderate prevalence and modest sensitivity.

When animals may recover, pressure $h(t)$ may be modified to include e.g., a quadratic time component. Intermittent behavior, e.g., in shedding virus when swabs of saliva or feces are collected for testing, may be modeled by adding a residual term e to threshold u at each time point. In that case it would be expedient to use (independent) normal distributions for u and e , with an associated probit link function. When some animals may turn seropositive after considerable time, an asymmetric link function, like the complementary log–log link function (associated with the extreme value distribution), may be more appropriate. The distribution of u implicitly defines a distribution for the moment of seroconversion. The latter distribution can also be specified directly, maybe motivated by some likely hazard function.

The prevalence structure may be generated by a latent Markov model, as discussed by Cook, Ng, and Meade (2000). The time dependent transition probabilities, from times t' to t , from $D = 1$ to 0, and from $D = 0$ to 1 are 0 and $P(h(t') < u \leq h(t))/P(u > h(t'))$, respectively. For known true status D , it would be natural to consider possible dependence between repeated measurements, see e.g., Hedeker (2003). There is no obvious reason, when true disease status is unknown, to ignore this dependence. Therefore, the random v effects, associated with subject specific sensitivity and specificity, were introduced in the model. For the FMD data there was no indication that the extra conditional dependence was of any marked importance. This may at least partly be due to the high values of sensitivity (and specificity) that offer little opportunity to detect additional dependence in time on top of the (trivial) dependence in time that follows from the unknown true state. In the simulation, v effects had to be omitted, to reduce computing time, which was still quite substantial.

Simulated data were also analyzed assuming that test 3 was conditionally independent of tests 1 and 2. This assumption may not be valid, because even when tests relate to different biological principles (Gardner et al. 2000), conditional dependence may occur when animals may be in varying stages of a disease. The highly structured prevalence for repeated test results mitigates potential problems with identifiability of parameters that is often associated with latent class models for diagnostic data. Here, there was no apparent gain in accuracy of sensitivity and specificity of the tests when the independence of test 3 was imposed upon the model, although in theory identifiability should be considerably enhanced by conditional independence.

Sera with known true status, such as true negative sera from animals from a well protected and carefully monitored environment, were not included in the present analysis. However, extension of the analysis with known true negatives (or positives) is straightforward. Data with known true status increase the possibility to study and motivate possible conditional correlation between repeated test results.

The present model could be the basis for an analysis of sera obtained from repeated visits to farms during an outbreak. Test results may have been obtained much later from the stored serum samples. Suppose that a farm entered surveillance when it was located within a certain radius of another farm that was confirmed positive. Possibly, at that moment, the farm was already positive itself. Suppose that the date of collection of a sample and the associated farm are known. Also assume that true state of farms was eventually known, based on reliable end screening. For farms that were confirmed positive, there is no gold standard, because generally the moment the disease was introduced at a farm is not known. The moment of introduction can be included as an extra unknown parameter, with a prior based on expert opinion, possibly utilizing additional information from data collected when herds were culled, see Engel et al. (2005) for an example with Classical Swine Fever. In principle, information about possible transmission of the disease between farms could be included as well, although this would complicate the analysis considerably. Accounting for the selective way farms enter into the dataset would be one of the problems. Excessive computer time is likely to be another.

Generally, the choice of true positive state is a difficult issue in latent class models for diagnostic data. For example, when some tests aim for detection of antibodies, such as ELISA tests, and others for detection of genetic material, such as PCR tests, the mixture that dominates the data may not necessarily correspond to the preferred true positive state that is practically relevant. For instance, during an outbreak, the true state of interest would be infectiousness of an animal. In that case, credible intervals may cover sensitivities and specificities corresponding to several possible mixtures, such as seroconversion or not, and absence or presence of genetic material (including that of dead virus or bacteria). When posteriors are not multimodal, the posterior median, as an estimator, is a compromise between these mixtures. Unless additional structure can be added to the model, based on information about the relation between different potential mixtures, this fundamental bias with respect to a preferred true positive state may remain whatever the size of the dataset. For the FMD data, with regard to export, first and foremost was detection of animals that had been in contact with the virus and seroconversion was an obvious choice for true state in this respect. This was also the natural choice regarding the background of the tests, since all were ELISA tests that aimed for detection of antibodies.

ACKNOWLEDGEMENTS

This research has been performed within the KB-8 Theme Animal Health and Welfare, "Improving Statistical Methods for Diagnostic Test Evaluation," KB-08-003 01.14, financed by the Dutch Ministry of Agriculture, Nature, and Food Quality.

[Received March 2008. Revised November 2008. Published Online February 2010.]

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

REFERENCES

- Brocchi, E., Bergmann, I. E., Dekker, A., Paton, D. J., Sammin, D. J., Greiner, M., Grazioli, S., De Simone, F., Yadin, H., Haas, B., Bulut, N., Malirat, V., Neitzert, E., Goris, N., Parida, S., Sorensen, K., and De Clercq, K. (2006), "Comparative Evaluation of Six ELISAs for the Detection of Antibodies to the Non-Structural Proteins of Foot-and-Mouth Disease Virus," *Vaccine*, 24, 6966–6979.
- Cook, R. J., Ng, E. T. M., and Meade, M. O. (2000), "Estimation of Operating Characteristics for Dependent Diagnostic Tests Based on Latent Markov Models," *Biometrics*, 56, 1109–1117.
- Dahlquist, G. Björck, Å. (1974), *Numerical Methods*, New Jersey: Prentice-Hall.
- Dendukuri, N., and Joseph, L. (2001), "Bayesian Approaches to Modelling the Conditional Dependence Between Multiple Diagnostic Tests," *Biometrics*, 57, 158–167.
- Engel, B., Bouma, A., Stegeman, A., Buist, W., Elbers, A., Kogut, J., Döpfer, D., and de Jong, M. C. M. (2005), "When Can a Veterinarian Be Expected to Detect Classical Swine Fever Among Breeding Sows in a Herd During an Outbreak?" *Preventive Veterinary Medicine*, 67, 195–212.
- Engel, B., Buist, W., Orsel, K., Dekker, A., de Clercq, K., Grazioli, S., and van Roermund, H. (2008), "A Bayesian Evaluation of 6 Diagnostic Tests for Foot-and-Mouth Disease for Vaccinated and Non-Vaccinated Cattle," *Preventive Veterinary Medicine*, 86, 124–138.
- Engel, B., Buist, W., and Visscher, A. (1995), "Inference for Threshold Models With Variance Components From the Generalized Linear Model Perspective," *Genetics Selection Evolution*, 27, 15–32.
- Engel, B., Swildens, B., Stegeman, A., Buist, W., and de Jong, M. (2006), "Estimation of Sensitivity and Specificity of Three Conditionally Dependent Diagnostic Tests in the Absence of a Gold Standard," *Journal of Agricultural, Biological and Environmental Statistics*, 11, 360–380.
- Gardner, I. A., Stryhn, H., Lind, P., and Collins, M. T. (2000), "Conditional Dependence Between Tests Affects the Diagnosis and Surveillance of Animal Diseases," *Preventive Veterinary Medicine*, 45, 107–122.
- Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 3, 515–533.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman & Hall.
- Georgiades, M. P., Johnson, W. O., Gardner, I. A., and Singh, R. (2003), "Correlation-Adjusted Estimation of Sensitivity and Specificity of Two Diagnostic Tests," *Applied Statistics*, 52, 63–76.
- Hedeker, D. (2003), "A Mixed-Effects Multinomial Logistic Regression Model," *Statistics in Medicine*, 22, 1433–1446.
- Hui, S. L., and Walter, S. D. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167–171.
- Kitching, R. P., Hutber, A. M., and Thrusfield, M. V. (2005), "A Review of Foot and Mouth Disease Virus With Special Consideration for the Clinical and Epidemiological Factors Relevant to Predictive Modelling of the Disease," *Veterinary Journal*, 169, 197–209.
- Miglioretti, D. L. (2003), "Latent Transition Regression for Mixed Outcomes," *Biometrics*, 59, 710–720.
- Qu, Y., and Hadgu, A. (1998), "A Model for Evaluating Sensitivity and Specificity for Correlated Diagnostic Tests in Efficacy Studies With an Imperfect Reference Test," *Journal of the American Statistical Association*, 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.
- Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM-Algorithm," *SIAM Review*, 26, 195–239.

- Schulzer, M., Anderson, D. R., and Drance, S. M. (1991), "Sensitivity and Specificity of a Diagnostic Test Determined by Repeated Observations in the Absence of an External Standard," *Journal of Clinical Epidemiology*, 44, 1167–1991.
- Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003) *WinBUGS User Manual. Version 1.4*, available at <http://www.mrc-bsu.cam.ac.uk/bugs>.
- Wolfram Research, Inc. (2007), *Mathematica, Version 6.0*, Champaign, IL.
- Yang, I., and Becker, M. P. (1997), "Latent Variable Modelling of Diagnostic Accuracy," *Biometrics*, 53, 948–958
- Zhang, S., and Jin, J. (1996), *Computation of Special Functions*, New York: Wiley.