Hierarchical model for certification of a country as "free" from an animal pathogen

E.A. Suess [a,*] , I.A. Gardner [b] , W.O. Johnson [c]

[a] *Department of Statistics, California State University, Hayward, CA 94542*

[b] *Department of Medicine and Epidemiology, School of Veterinary Medicine University of California, Davis, CA 95616*

[c] *Department of Statistics, University of California, Davis, CA 95616*

Abstract

Certification that a country, region or state is "free" from a disease has implications for trade in animals and animal products. We develop a Bayesian model for assessment of (i) the probability that a country is disease "free" (or infected), (ii) the proportion of infected herds in an infected country, and (iii) the within-herd prevalence in infected herds. The model uses test results from animals sampled in a two-stage cluster sample of herds within a country. Model parameters are estimated using modern Markov-chain Monte Carlo methods. We demonstrate our approach using published data from surveys of Newcastle disease and porcine reproductive and respiratory syndrome in Switzerland, and for three simulated data sets.

Key words: Herd-level testing; Disease freedom; Bayesian modeling; Markov-chain Monte Carlo; Gibbs sampler; Adaptive Rejection Sampling; Newcastle disease; Porcine reproductive and respiratory syndrome

* Corresponding author. Tel.: 510-885-3879; fax: 510-885-4714; email address: es-

1. Introduction

To facilitate animal and animal-products trade, veterinary authorities in a country (region, etc.) might try to provide evidence that livestock populations are free from important infectious agents. Countries might have "always been free" of a pathogen based on years of negative disease surveillance data or might have eradicated the agent recently. Historic evidence of freedom from infection might be based on criteria such as lack of clinical disease for a specified period of time, cessation of use of vaccines that might disguise the condition, no positive diagnoses at local diagnostic laboratories, and (often) some test-based survey or surveillance data. Formal incorporation of this evidence into the analysis would be useful for making inferences about a country's status regarding to a particular pathogen. In addition, the risk of pathogen introduction can vary geographically depending on the extent of animal contact and/or movement of animal or animal products within and between neighboring regions or countries. This factor might also warrant consideration when data are analyzed.

To provide the necessary assurance of freedom from infection (or a prevalence below a defined threshold), most countries will conduct a national survey using internationally recognized diagnostic tests on a large sample of animals. These surveys could be based on samples collected at slaughter or on testing of live animals in herds. In the latter case, the testing generally would be performed using a two-stage cluster-sampling scheme with the selection of k herds and then a random sample of n animals (the selection could be age-specific or focused on high-risk groups) from each herd. The sample size (n) is often the

uess@csuhayward.edu

same from herd to herd, but it could vary based on formulas designed to adjust for the total herd size.

Serologic tests typically are used in national surveys because they are inexpensive, and are rapid and easy to perform. However, such tests will always have imperfect sensitivity and specificity. Thus, a survey that resulted in only a few reactors (positive test results) does not imply infection.

Criteria for assessment of disease freedom have been suggested by Baldock (1998) and a frequentist approach to the analysis of two-stage cluster sampling designs incorporating imperfect test sensitivity and specificity has been developed by Cameron and Baldock (1998). As an alternative analytic approach, Audigé and Beckett (1999) developed a stochastic simulation model that allowed for the incorporation of uncertainty in input parameters through the use of probability distributions. They used the magnitude of the likelihood ratio as an indicator of country-level infection. Recently, Audigé et al. (2001) updated the model to incorporate uncertainty in the likelihood ratio and prior probability of country-level infection.

In this paper, we use a Bayesian approach to model test results from a two-stage cluster sample. Our main objective is to extend the work of Audigé and Beckett (1999) and Audigéet al. (2001) to an all-encompassing model for diagnostic-test data from herd-level testing that will be useful for making inferences about infection status at three levels – the country, the herd, and within the herd. The model has been implemented in Fortran90 (Digital Equipment Corporation, 110 Spit Brook Road, mail stop ZKO2-3/N30, Nashua, New Hampshire, 03062-2698) and the prior and posterior analyses are performed in R (Free Software Foundation, Temple Place - Suite 330, Boston, MA 02111-1307, USA). We illustrate this modeling approach using survey data from Switzerland for Newcastle disease (ND) virus and porcine reproductive and respiratory syndrome (PRRS), and for three simulated data sets.

We begin by discussing the formulation of our model for disease freedom in Section 2. In our model we assume individual-test results are available for each animal from a two-stage cluster sample and assume an equal sample size (n) within herds. In Section 3, we explain the Bayesian approach to inference. In Section 4, we present results for real survey data and for simulated data examples. Finally, we give our conclusions in Section 5.

2. Model

Our model assumes that the diagnostic-test results from a herd-level cluster sample are available. This sampling scheme is used to produce data by randomly selecting k herds (clusters) from the population of herds in a country, and then, within herd i, n_i animals are selected randomly and tested. We assume the herd size is large relative to n_i. We also assume the diagnostic test used to detect the pathogen in question is not perfect; either the test sensitivity ($\eta = P(+|I)$) or specificity ($\theta = P(-|\bar{I})$) or both are < 1 (where I indicates a truly infected animal and $+$ indicates a positive test result).

Our primary goal is to assess the probability that an animal population in a country is "free" from a specific pathogen (i.e. that the prevalence is either zero or so small that it is of no practical relevance). To model the country-level infection status, we let $Z = 1$ if the animal population is infected and $Z = 0$

otherwise. Let $\gamma = P(Z = 1)$, so that $(1-\gamma)$ is the probability the country is "free" of the pathogen.

The prevalence of infection within each herd is also of interest – as is the average prevalence among the infected herds. The prevalence of infection within the i^{th} herd (λ_i) is defined as the prevalence of infection in the population from which the i^{th} herd was sampled. We assume two conditions – either *infected* or *non-infected* – exist under which the i^{th} herd can be selected; and hence, each of the k herds is drawn randomly from one of these two populations. If the i^{th} herd is sampled from the *infected* population, we define π_i as the within-herd prevalence of the infection. We also assume that the prevalences (π_i) in infected herds vary. The actual prevalence for herd i (π_i) is assumed to be drawn from a $beta(\alpha, \beta)$ distribution where the unknown parameters (α, β) determine the average prevalence of infection $(\mu = \alpha/(\alpha+\beta))$ among infected herds and also how variable these prevalences are about the mean. The proportion of infected herds (i.e. or herd-level prevalence) is assumed to be τ and thus the i^{th} herd is assumed either to have prevalence π_i with probability τ or prevalence zero with probability $(1-\tau)$. Thus, we define λ_i to be the prevalence of infection in the population from which the i^{th} herd is sampled, as:

$$
\begin{aligned}
\lambda_i &= \pi_i \qquad && \textit{with probability } \tau \\
&= 0 && \textit{otherwise.}
\end{aligned}
$$

It follows that λ_i and λ_j, $i \neq j$, are independent conditional on the vector of prevalences $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_k]'$, and τ.

At the herd level, we model the true infection status using latent data $\{t_i : i = 1, ..., k\}$, where each t_i is an indicator of the i^{th} herd's true infection status.

Thus each t_i is a Bernoulli random variable with probability τ,

$$t_i|\tau \sim Ber(\tau).$$

This leads to the equality ($\lambda_i = \pi_i t_i$) which implies that if the i^{th} herd is infected ($t_i = 1$), then $\lambda_i = \pi_i$. Finally, we note if a herd is infected (i.e. $t_i = 1$), then the country is infected also; thus, $Z = 1$.

We define additional latent data for the i^{th} herd that identifies the true infection-status of each animal tested. The latent data $\{v_{ij} : i = 1, ..., k, j = 1, ..., n\}$ are a group of indicator variables where $v_{ij} = 1$ if the j^{th} animal in the i^{th} herd is infected and $v_{ij} = 0$ if it is not infected. The conditional distributions of the v_{ij}'s are independent and distributed as Bernoulli,

$$v_{ij}|Z = 1, t_i = 1, \pi_i \sim Ber(\pi_i).$$

If a herd is not infected, then each individual animal in the herd is not infected by definition; i.e. $v_{ij} = 0|t_i = 0$, with probability one. Also, for a non-infected country, each animal is not infected; $v_{ij} = 0|Z = 0$, with probability 1. Finally, an important feature of this model is that its structure incorporates correlation in the true infection status between animals (j and j', $j \neq j'$) within the i^{th} herd:

$$Corr(v_{ij}, v_{ij'}) = \frac{(1-\tau)\pi_i}{1-\tau\pi_i} \geq 0.$$

However, the model leaves the true infection status of different animals in separate herds (i and i') independent.

The data available for our model are the individual-animal test results from a two-stage cluster sample of herds within a country. The data are represented as $\{X_{ij} : i = 1, ..., k, j = 1, ..., n_i\}$, where $X_{ij} = 1$ if the j^{th} animal within the i^{th} herd tests positive and $X_{ij} = 0$ if the animal tests negative. For each herd, we can collectively represent the data as $\{(X_i, n_i) : i = 1, ..., k\}$, where $X_i = \sum_j X_{ij}$. Therefore, the X_{ij}'s denote the outcomes for each individual-animal's test result, and the X_i's give the total number of animals within each herd that test positive for the pathogen.

The individual-test results for each animal within a herd are assumed to be Bernoulli random variables with probability of testing positive that depends on the country's infection status, the within-herd level prevalence, λ_i, and the test parameters η and θ. The conditional distribution of X_i – given $Z = 1$, λ_i, η, and θ – is binomial because each animal is assumed to be selected randomly within a herd. Specifically, we model the results for the i^{th} herd in an infected country as

$$X_i | Z = 1, \lambda_i, \eta, \theta \sim Bin\left(n_i, \lambda_i \eta + (1 - \lambda_i)(1 - \theta)\right)$$

and

$$X_i | Z = 0, \theta \sim Bin\left(n_i, (1 - \theta)\right).$$

Regarding the individual-test results, we assume that the outcomes are independent and Bernoulli, conditional on the infection status of the country, the infection status of the animal tested, and the test parameters, namely

$$X_{ij} | Z = 1, \{v_{ij}\}, \eta, \theta \sim Ber(\eta^{v_{ij}} (1 - \theta)^{(1 - v_{ij})}) \tag{1}$$

and

$$X_{ij} | Z = 0, \theta \sim Ber(1 - \theta). \tag{2}$$

The model parameters are summarized in Table 1. A flow chart showing the levels that are modeled with latent data and their relationship to the data collected is presented Figure 1.

3. The Bayesian Approach

We require prior distributions for the unknown model parameters and the joint distribution of these parameters in conjunction with the latent data used in the model. Independent beta priors are assumed for the model parameters μ, γ, τ, η, and θ, and an independent gamma prior for $\alpha + \beta$. In general, for a generic parameter ν we use the notation $\nu \sim beta(a_\nu, b_\nu)$ to specify its beta prior distribution. Suppose ν is a model parameter for which a $beta(a_\nu, b_\nu)$ is to be selected. For the parameter ν, three questions are asked of an expert familiar with the country's animal population and the specific animal pathogen:

(1) What do you believe is the most-likely value of ν? This value is chosen to be the mode of the corresponding beta prior.

(2) What is your 5^{th} percentile of the possible values of ν? (e.g. you are 95% certain that ν exceeds what value?)

(3) What is your 95^{th} percentile of the possible values of ν?

Answers to these questions are used to obtain the prior density for ν. If the most-likely value of ν is < 0.50, the mode and the 95^{th} percentile are used (and if it is > 0.50, the mode and the 5^{th} percentile are used) to determine the parameters (a_ν, b_ν). Only one of the percentiles is used because the beta density is not symmetric in these two cases. If the mode is chosen to be 0.50, we use both percentiles because the prior will be symmetric. Finally, for the selected prior density, we calculate 95% prior intervals for comparison with the 95% posterior intervals that are calculated from the posterior densities.

The prior distributions on the parameters (α, β) are derived using information from an expert about the average within-herd prevalence in the infected herds ($\mu = E[\pi_i | t_i = 1] = \alpha/(\alpha + \beta)$) and the standard deviation ($\sigma = \sqrt{\mu(1-\mu)/(\alpha+\beta+1)}$) of the possible prevalences among infected herds. In the simplest example, a prior guess for μ might be 0.30; that is, the expert's best guess for the average prevalence among herds is 30%. The expert might also be confident that 95% of all prevalences are within 10% of his/her best guess for μ, which implies a guess that $\sigma \approx 0.05$ (assuming the distribution of prevalences is approximately symmetric) one then could solve the equations $\mu = 0.30$ and $\sigma = 0.05$ for the expert guesses for α and β. Our approach is more complicated than this, but the basic idea is conveyed by this illustration. The exact derivation of the priors is given in Appendix A.

The distributional assumptions for the three sets of latent data ($\{v_{ij}\}$, $\{t_i\}$, and $\{\pi_i\}$) were presented in the previous section. These – in conjunction with the distributional assumptions for the data in (1) and (2) and the prior distributional assumptions given here – result in the joint distribution (given in Appendix B) of all of the quantities. This is used to obtain the so-called "full conditional posterior distributions," which are sampled iteratively in a process called "Gibbs sampling" (Appendix C). The Gibbs sampler is used to simulate a Monte Carlo (MC) sample of values from the joint posterior distribution (Gelfand and Smith, 1990; Casella and George, 1992; Gelman et al., 1995; Tanner, 1996). The sampling is conditional, because each sampled value depends on the data ($X = \{X_{ij}\}$), and on the previously sampled values of the parameters and latent data.

The Gibbs sampler for our model results in the MC sample $\{v_{ij}^{(h)}\}$, $\{t_i^{(h)}\}$, $\{\pi_i^{(h)}\}$, $\alpha^{(h)}$, $\beta^{(h)}$, $Z^{(h)}$, $\gamma^{(h)}$, $\tau^{(h)}$, $\eta^{(h)}$, and $\theta^{(h)}$: $h = 1, ..., N$, which is obtained in this order. Initial values are chosen for all parameters and the latent data. (We use the means of the corresponding prior distributions as initial values of the parameters.) We specify the initial values for the latent data as the test-result values as if sensitivity and specificity were perfect. For each iteration (h) the sampling within the iteration depends on the previously sampled values. For example, $v_{ij}^{(h+1)}$ is sampled conditional on the data X and on the most-recent values of the other variables. Next, $t_i^{(h+1)}$ is sampled conditional on the same collection and now also on $\{v_{ij}^{(h+1)}\}$. During each iteration, all of the values are sampled – producing a dependent chain of MC samples for each parameter and for each latent data value.

The Bayesian posterior analysis is performed by plotting smoothed histograms of the sampled values for particular parameters as estimates of the posterior densities. MC sample modes and means are used as point estimates and posterior (credible) intervals are calculated as the 0.025 and 0.975 percentiles of the corresponding MC sample values for each parameter. In all of the analyses presented in this paper, 200,000 iterations of the Gibbs sampler were generated and the last 50,000 iterations were used to estimate the posterior distributions.

4. Illustrations

In this section, we present results of our Bayesian analysis for two data sets previously evaluated by other authors. We analyze survey data that were collected in Switzerland to assess freedom from Newcastle disease virus in poultry (Gohm et al., 1999) and porcine reproductive and respiratory syndrome in swine (Audigé et al., 1997; Canon et al., 1998). We also present the results of three simulated data sets.

4.1. Survey data

4.1.1. Newcastle disease

During 1996, blood samples were collected from a central poultry slaughterhouse in Switzerland to assess the status of Newcastle-disease (ND) virus in the country. Samples were collected from $k = 260$ flocks with $n = 30$ birds per flock. Sample sizes were based on the assumption that at least 1% of flocks were infected and within-flock prevalence was at least 10%. Serum samples were tested for ND antibodies by ELISA. Upon initial inspection of ELISA-test results, four flocks had test results consistent with a high likelihood of infection (22, 14, 10, and 10 positive samples out of 30). The remaining 256 flocks had 3 or fewer positive test results (3 flocks with 3 positives, 9 flocks with 2 positives, 50 flocks with 1 positive, and 194 flocks with 0 positives).

The priors were elicited from Dr. Laurent Audigé (a coauthor of the original study) to reflect his uncertainty in the model parameters for ND virus, and are given in Table 2. The parameters selected for the prior distributions were chosen to best fit the prior information given in terms of the prior mode, the lower 5^{th} percentile value and the upper 95^{th} percentile. A prior on γ was chosen having mode 0.20 and the prior mode for the herd-level prevalence (τ), was chosen to be 0.01. These modes along with input for the 5^{th} and 95^{th} percentiles were used to find the best-fitting beta priors. The independent gamma priors for the parameters α and β were derived from the prior information about μ and σ. A beta prior on μ was chosen with a mode of 0.30. Also given in the Table 2 are the 95% prior intervals, means, and standard deviations for comparison with the posterior analysis.

The posterior analysis based on the simulated MC samples produced by the Gibbs sampler for model parameters is presented in Table 2. The survey results provide evidence that the country's poultry population was infected with ND virus. The indicator of the infection in the country resulted in a posterior probability of $P(Z = 1|data) = 1$. Because the results indicate that the country is infected, inferences about the flock prevalence and within-flock prevalences are presented. E.g., the posterior estimate of the flock-level prevalence (τ) is 0.018 with a 95% posterior interval $(0.007, 0.035)$. The posterior estimate of the probability that the country is infected (γ) is 0.301 with corresponding interval (0.100, 0.611) Two plots are included to show the updating of the prior distributions. In Fig. 2, the prior and posterior distributions of the mean prevalence of infection within the diseased flocks (μ) is presented. It should be noticed that the posterior estimate of μ is increased from the prior estimate and that the posterior is more concentrated than the prior; this is because ... . In Fig. 3, the prior and posterior prevalence distributions for the infected flocks are presented. Here, the mean of the posterior prevalence distribution of the π_i is increased because of the many test-positive results in multiple herds. Our findings are consistent with those of Gohm et al. (1999) who found a likelihood ratio of 56.3 and concluded that 4 flocks were likely infected and that the Swiss poultry population was not free of ND at the time of the survey.

4.1.2. Porcine reproductive and respiratory syndrome

In July 1996, a survey for porcine reproductive and respiratory syndrome (PRRS) was done to verify that Switzerland was free of the pathogen. Sera were collected from $k = 108$ herds with $n = 5$ pigs tested per herd. All samples were seronegative to PRRS by ELISA (Canon et al 1998).

The priors chosen to reflect the uncertainty in the model parameters were derived from the literature related to PRRS and the expert opionion of Dr. Laurent Audigé. The prior on γ was selected with a mode of 0.01 and a 95% prior interval (0.002, 0.059). The beta prior on the mean within-herd level prevalence (μ) in the infected herds was chosen with a mode of 0.475 and interval (0.283, 0.675). The beta prior on the herd level prevalence τ was chosen with a mode of 0.05 and interval of (0.022, 0.110). The beta prior on the sensitivity η was chose with mode 0.980 and interval (0.883, 0.995). The specificity θ was chosen with mode 0.995 and interval (0.965, 0.999).

The posterior analysis indicated that the country was not infected. The posterior probability that the country was infected was $P(Z = 1|data) = 0.0012$ – indicating that the country was not likely to be infected with PRRS at the time of the survey. The only model parameters that can be estimated in this case are the specificity of the test ($\theta = P(-|\bar{D})$) and an updated estimate of the initial probability the the country is infected (γ) (Table 3). Note that the posterior mode for specificity (0.999) is somewhat larger than the prior mode (0.995). This is to be expected since ... .

4.2. Simulated data examples

In this section we present the results from the analysis of three simulated data examples to demonstrate that our method unequivocally can identify a clearly infected country – but also to show that indeterminate results ($0 < P(Z = 1|data) < 1$) are possible. Two data sets were created assuming an infected country. The first data set was produced with clear evidence that the animal population was infected. The second was produced having few infected herds and few infected animals within the infected herds (to show the potential for indeterminate results). We used informative prior distributions to analyze both data sets (Sections 4.2.1 and 4.2.2). For the mean within herd-level prevalence in infected herds (μ) we assumed a beta prior with mode 0.30 and a 95% prior interval of (0.16, 0.50). For the herd level prevalence (τ) we used a beta prior with mode 0.05 and interval (0.02, 0.11). The prior modes and intervals for the sensitivity (η) and specificity (θ) were 0.995 (0.989, 0.998). Finally, for the probability that the country is infected (γ) a prior mode and interval of 0.30 (0.14, 0.54) was used to derive the prior distribution.

We also present a re-analysis of the ND data that were described in Section 4.1.1. The ND data were modified by removing the 4 likely infected flocks that had at least 10 reactors, leaving the other 256 flocks with 3 or fewer test-positive birds. This analysis shows a realistic case where indeterminate results are produced.

4.2.1. With infected herds

In this scenario, the true herd-level prevalence was $\tau = 0.25$ (which is extremely implausible under the prior with mode 0.05, which might be considered miss-specified in this case) and the true mean within herd-level prevalence for infected herds was $\mu = 0.60$. We produced a data set with $k = 100$ herds having exactly 25 truly infected herds, each having 12 infected animals out of $n = 20$ sampled from each herd. We added false-positive and false-negative test results by assuming that the test sensitivity and specificity were both 0.99. Accordingly, there were 3 false-negatives added and 14 false-positive test results added to the data set. The false positives were randomly placed in 14 of the 75 truly non-infected herds.

Using these data and the informative priors described in Section 4.2, our analysis correctly identified the country as infected (the posterior probability of $P(Z = 1|data) = 1$). The posterior mode for the mean within herd-level prevalence for the infected herds was $\mu = 0.57$ with a 95% posterior interval of (0.52, 0.62) and for the herd level prevalence the posterior mode was $\tau = 0.15$ with a 95% posterior interval of (0.11, 0.20). Note that a strong prior for τ, that is focused well below the true value, has resulted in a posterior interval for τ that excludes the true value.

The results of this example are representative of the many simulated data sets we analyzed. When the proportion of infected herds is moderate to high, and the proportion of test-positive animals is moderate to high within infected herds, the country is consistently determined to be infected and the model parameters are well estimated by the posterior analysis.

As noted in the above illustration, we also found that when an informative prior on the herd-level prevalence is used and if the true τ is outside the plausible range of the prior, the posterior might not concentrate on the true value. This is because the posterior estimate is generally a weighted average of the prior guess and a purely data-based estimate. Thus, we recommend the use of less informative (more dispersed) prior distributions on τ. For example, if a uniform prior for τ is used in the above analysis we obtained a posterior mode of 0.25 and the 95% posterior interval was $(0.18, 0.34)$.

4.2.2. Indeterminate case

For the indeterminate scenario, we generated data for an infected country in which $\tau = 0.02$ and $\mu = 0.15$. We consider data with $k = 100$, with two herds selected to be infected. Within these two herds, 3 of the 20 animals samples were truly infected. The test sensitivity and specificity were both assumed to be 0.98. No false-negatives and 38 false-positive test results were included in the data set. One of the false positives was added to an infected herd (making four test positives for that herd).

The results of the Bayesian analysis were indeterminate. The posterior probability $P(Z = 1|data) = 0.55$, and only one herd was detected as infected in 54% of the iterations of the Gibbs sampler. The posterior mode of the mean within herd-level prevalence for the infected herds (μ) was 0.20 with a 95% posterior interval of (0.12, 0.35) – which shows a large decrease from the prior values 0.30 (0.16, 0.50). Note that the true value (0.15) used to create the data set is included in the posterior interval. The posterior mode of the herd-level prevalence was 0.03 (0.01, 0.07), which included the true value of $\tau = 0.02$ compared with the prior value of 0.05 (0.02, 0.11).

Hence, the results of this analysis are equivocal. The country is not clearly identified as infected and only one herd which was suspected of being infected is also not clearly identified as being infected. For these types of situations, increasing the number of herds sampled is recommended – which will produce data from more infected herds (if they exist) and increase the likelihood of correctly defining the true status of the country.

4.2.3. Newcastle disease without clearly infected flocks

With the intention of creating a realistic data set that would produce indeterminate results, we modified the ND data by removing the test results for the 4 clearly infected flocks. The data set that remained contained 3 flocks with 3 positives, 9 flocks with 2 positives, and 50 flocks with 1 positive test result and 198 flocks with 0 positives.

To analyze the data, we assumed the same prior information and therefore the same prior distributions are used as before to analyze the full data set. The results of the analysis were indeterminate: the $P(Z = 1|data) = 0.0835$.

5. Conclusions

We have developed and presented a purely Bayesian model that is potentially useful for evaluating the status of a country or region with respect to freedom from an animal pathogen. The Bayesian approach incorporates prior knowledge along with the observed data to produce updated posterior inferences. It is the calculation of the posterior distributions that is the main advancement over previous work in this area of research. Specifically, posterior distributions for both the proportion of infected herds and the within-herd prevalence are considered to be of greater utility for risk analysts involved in animal trade than knowledge of a country's infection status alone (R. Fite, pers. comm.).

Our model allows for 3 levels of inference when the country is infected. If the country is not likely to be infected, we report the updated estimates of the initial probability that the country is infected (γ) and the specificity (θ) and $P(Z = 1|data)$. If the country has a probability of being infected – i.e. if $P(Z = 1|data)$ is greater than a specified threshold – our analysis produces a posterior inference for each parameter in the model. We report the country-level inference $P(Z = 1|data)$, the herd-level prevalence, and the within herd-level prevalence distribution for the infected herds. In contrast to the models of Audigé and Beckett (1999) and Audigé et al. (2001), our model does not require the specification of a cutoff value for the number of reactors to define a herd as infected. We model the status of the herd with latent data (t_i) and ultimately determine $P(t_i|data)$ as a method of assessing the status of each herd. For the ND and PRRS examples, we make the same conclusions as the authors of those studies. One of the outputs from our model is an updated estimate of γ; however, this value changes minimally because it is modified only by Z = 0 or 1. Thus, prior and posterior inferences for this parameter generally will be quite similar.

We believe that this model (in conjuction with the software we have developed) will be a valuable tool for making decisions about the infection status of the animal populations within countries, and for monitoring changes in prevalence within infected countries.

Acknowledgments

This study was funded in part by USDA Formula Funds and the USDA NRI Competitive Grants Program. We thank Dr. Laurent Audigé for providing values used to construct prior distributions for the ND and PRRS analyses.

References

[1] Audigé, L., Beckett, S., 1999, A quantitative assessment of the validity of animal-health surveys using stochastic modelling. Preventive Veterinary Medicine 38, 259-276.

[2] Audigé, L., Beckett, S., Canon, N., Hofmann, M., Groit, C., 1997, Freedom from infection: A stochastic simulation model for evaluating the efficacy of national or regional surveys applied to PRRS in Switzerland. Epidémiol. santé anim., 31-32.

[3] Audigé, L., Doherr, M.G., Hauser, R., Salman, M.D., 2001, Stochastic modelling as a tool for planning animal-health surveys and interpreting screening-test results. Preventive Veterinary Medicine 49, 1-17.

[4] Baldock, F.C., 1998, What constitutes freedom from disease in livestock. Australian Veterinary Journal 76, 544-545.

[5] Cameron, A.R., Baldock, F.C., 1998, Two stage sampling in surveys to substantiate freedom from disease. Preventive Veterinary Medicine 34, 19-30.

[6] Canon, L., Audigé, L., Denac, H., Griot, C., 1998, Evidence of freedom from porcine reproductive and respiratory syndrome virus infection in Switzerland. Veterinary Record 143, 142-143.

[7] Casella, G., George, E.I., 1992, Explaining the Gibbs sampler. American Statistician 46, 167-174.

[8] Gelfand, A.E., Smith, A.F.M., 1990, Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398-409.

[9] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995, Bayesian Data Analysis. Chapman Hall, New York, 526pp.

[10] Gilks, W.R., Wild, P., 1992, Adaptive rejection sampling for Gibbs sampling. Applied Statistics, Journal of the Royal Statistical Society 14, 337-348.

[11] Gohm, D., Thür, B., Audigé, L., Hofmann, 1999, A survey of Newcastle disease in Swiss laying-hen flocks using serological testing and simulation modelling. Preventive Veterinary Medicine 38, 277-288.

[12] Ross, S.,1997, A First Course in Probability, 5th edition. MacMillan, New York, 514pp.

[13] Tanner, M.A.,1996, Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition. Springer-Verlag, New York, 207pp.

Appendix A. Prior specification for α and β.

We select a $beta(a_\mu, b_\mu)$ prior on $\mu = \alpha/(\alpha+\beta)$, the mean value of the beta distribution on the within-herd level prevalence in the infected herds, $\{\pi_i\}$. We further define $\psi = \alpha + \beta$, which is functionally related to the standard deviation $\sigma = \sqrt{\mu(1-\mu)/(\psi+1)}$. We select a $gamma(r, s)$ for the parameter ψ with this given specification it is possible to obtain the induced distribution for (α, β) using the usual transformation technique, Ross (1997). Then under the condition that $r = a_\mu + b_\mu$, it follows that $\alpha \sim gamma(a_\mu, s)$ and $\beta \sim gamma(b_\mu, s)$, independently. This makes the Gibbs sampler discussed in Appendix C particularly easy to develop.

The values a_μ and b_μ used to determine the prior on μ are determined as before using expert opinion for the mode, 5^{th} and 95^{th} percentiles for μ. Given the mode, $\tilde{\mu}$, we know $(a_\mu - 1)/(a_\mu + b_\mu - 2) = \tilde{\mu}$, solving for b_μ, we obtain

$$b_\mu = \frac{a_\mu(1-\tilde{\mu}) - 1 + 2\tilde{\mu}}{\tilde{\mu}}.$$

In order to select a value of s for the prior on ψ, a best guess for ψ is necessary. If $\tilde{\psi}$ is given as the mode of the gamma prior, and $a_\mu + b_\mu = r$, then $\tilde{\psi} = (r-1)/s$ and thus

$$s = (r-1)/\tilde{\psi} = (a_\mu + b_\mu + 1)/\tilde{\psi}. \tag{3}$$

The selection of $\tilde{\psi}$ is derived using a normal approximation. We consider c^*, the median of the ψ density given $\mu = \tilde{\mu}$, e.g.

$$P(\psi \leq c^* \mid \mu = \tilde{\mu}) = 0.5,$$

which is the same as

$$P\left(\sqrt{(\alpha+\beta+1)/[\mu(1-\mu)]} \leq \sqrt{(c^*+1)/[\mu(1-\mu)]} \mid \mu = \tilde{\mu}\right) = 0.5,$$

since $\psi = \alpha + \beta$. Now recall that the standard deviation of the gamma distribution is $\sigma = \sqrt{[\mu(1-\mu)]/(\alpha+\beta+1)}$. Thus

$$P\left(\sigma \geq \sqrt{[\tilde{\mu}(1-\tilde{\mu})]/(c^*+1)} \mid \mu = \tilde{\mu}\right) = 0.5.$$

Let $\tilde{k} = \sqrt{[\tilde{\mu}(1-\tilde{\mu})]/(c^*+1)}$. Now let q_α be the $(1-\alpha)$ percentile of the prevalence distribution, e.g., $100(1-\alpha)\%$ of the prevalences in infected herds are smaller than q_α. This assumes that $q_\alpha = \mu + z_\alpha\sigma$, which would be the case if the prevalence distribution were approximately normal. Then the above is approximately equivalent to

$$P(q_\alpha \geq \tilde{\mu} + z_\alpha\tilde{k} \mid \mu = \tilde{\mu}) = 0.5.$$

Finally if the expert gives his or her best guess for q_α, say $\hat{q}_\alpha$, we set $\hat{q}_\alpha = \mu + z_\alpha\tilde{k}$ and solve for c^*, namely

$$c^* = \frac{z_\alpha^2\tilde{\mu}(1-\tilde{\mu})}{(q_{c^*}-\tilde{\mu})^2} - 1.$$

Using the median of the prior on ψ as the best guess of ψ, $\tilde{\psi} = c^*$, which is substituted into (3).

Appendix B. Joint Distribution of all varaibles.

From the model specification and the choice of priors, the joint distribution of the model parameters and the latent data, given the country is infected, $Z = 1$, is

$$p(\{v_{ij}\},\{t_i\},\{\pi_i\},(\alpha,\beta),\gamma,\tau,\eta,\theta|Z=1)$$

$$\begin{aligned}&= p(\{v_{ij}\}|Z=1,\{t_i\},\{\pi_i\})p(\{t_i\}|Z=1,\tau)\\ &\qquad \times p(\{\pi_i\}|Z=1,\{t_i\},(\alpha,\beta))p((\alpha,\beta)|\{t_i\})\\ &\qquad \times p(\gamma)p(\tau)p(\eta)p(\theta)\end{aligned}$$

$$= \prod_{ij}\left[\lambda_i^{v_{ij}}(1-\lambda_i)^{(1-v_{ij})}\right]^{t_i}\left[I_{\{0\}}(v_{ij})\right]^{1-t_i}$$

$$
\begin{aligned}
&\times \left[I_{\{0\}}(\lambda_i)\right]^{1-t_i} \prod_{i=1}^{k} \tau^{t_i}(1-\tau)^{1-t_i} \\
&\times \prod_{i=1}^{k} \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}\right]^{t_i} \\
&\times p((\alpha,\beta)|\{t_i\})p(\gamma)p(\tau)p(\eta)p(\theta). \qquad (4)
\end{aligned}
$$

The joint distribution of the parameters and latent data given the country is non-infected, $Z = 0$, simplifies to

$$
\begin{aligned}
&p(\{v_{ij}\}, \{t_i\}, \{\pi_i\}, (\alpha,\beta), \gamma, \tau, \eta, \theta | Z = 0) \\
&\quad = p(\gamma)p(\tau)p(\eta)p(\theta). \qquad (5)
\end{aligned}
$$

The joint distribution of the data, latent data, all of the parameters is obtained from (4) and (5) as

$$
\begin{aligned}
&p(\{X_{ij}\}, \{v_{ij}\}, \{t_i\}, \{\lambda_i\}, (\alpha,\beta), \gamma, \tau, \eta, \theta | Z = 1) \\
&\propto \eta^{\sum X_{ij}v_{ij}}(1-\eta)^{\sum(1-X_{ij})v_{ij}} \\
&\quad \times \theta^{\sum(1-X_{ij})(1-v_{ij})}(1-\theta)^{\sum X_{ij}(1-v_{ij})} \\
&\quad \times \prod_{ij} \left[\lambda_i^{v_{ij}}(1-\lambda_i)^{(1-v_{ij})}\right]^{t_i} \left[I_{\{0\}}(v_{ij})\right]^{1-t_i} \\
&\quad \times \left[I_{\{0\}}(\lambda_i)\right]^{1-t_i} \prod_{i=1}^{k} \tau^{t_i}(1-\tau)^{1-t_i} \\
&\quad \times \prod_{i=1}^{k} \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}\right]^{t_i} \\
&\quad \times p((\alpha,\beta)|\{t_i\})p(\gamma)p(\tau)p(\eta)p(\theta). \qquad (6)
\end{aligned}
$$

Similarly, using (2) and (5), the joint distribution conditional on $Z = 0$ can be written as

$$
\begin{aligned}
&p(\{X_{ij}\}, \{v_{ij} = 0\}, \{t_i = 0\}, \{\lambda_i = 0\}, (\alpha,\beta), \gamma, \tau, \eta, \theta | Z = 0) \\
&\quad \propto \left[\prod_{ij} \theta^{(1-X_{ij})}(1-\theta)^{X_{ij}}\right](1-\tau)^k p(\gamma)p(\tau)p(\theta). \qquad (7)
\end{aligned}
$$

Appendix C. Full conditional distributions.

The conditional distribution of the infection status of the j^{th} animal within i^{th} herd, given the country is infected $Z = 1$, the i^{th} herd is infected, $t_i = 1$, and the animal has a positive test result, $X_{ij} = 1$, is

$$v_{ij}|Z = 1, X_{ij} = 1, t_i = 1, \pi_i, \eta, \theta \sim Ber(P(I|X_{ij} = 1)),$$

where

$$P(I|X_{ij} = 1) = \frac{\eta\pi_i}{\eta\pi_i + (1-\theta)(1-\pi_i)}.$$

The distribution of v_{ij}, given the country is infected, the i^{th} herd is infected and the animal has a negative test result, $X_{ij} = 0$, is

$$v_{ij}|Z = 1, X_{ij} = 0, t_i = 1, \pi_i, \eta, \theta \sim Ber(P(I|X_{ij} = 0)),$$

where

$$P(I|X_{ij} = 0) = \frac{(1-\eta)\pi_i}{(1-\eta)\pi_i + \theta(1-\pi_i)}.$$

And the distribution of v_{ij} for a negative herd is $v_{ij}|Z = 1, t_i = 0 \sim Ber(0)$. And finally, the distribution of v_{ij} for a non-infected country is $v_{ij}|Z = 0 \sim Ber(0)$.

Next we present the distribution for the infection status of each herd, conditional on the related parameters in the model. The distribution of each t_i, given the country is infected, $Z = 1$, and when $v_{ij} = 1$, is

$$t_i|Z = 1, \sum_j v_{ij} > 0 \sim Ber(1)$$

or if $v_{ij} = 0$ for all j within herd i and $Z = 1$, we have

$$t_i|Z=1,\sum_j v_{ij}=0,\{\pi_i\},\tau \sim Ber\left(\frac{(1-\pi_i)^n\tau}{(1-\pi_i)^{n_i}\tau+1(1-\tau)}\right).$$

And finally for a non-infected country, $t_i|Z=0 \sim Ber(0)$.

The distribution of the within-herd level prevalence, π_i, in a infected herd, is

$$\pi_i|t_i=1,\{v_{ij}\},(\alpha,\beta) \sim Beta\left(\alpha+\sum_{j=1}^{n_i} v_{ij},\beta+n_i-\sum_{j=1}^{n_i} v_{ij}\right).$$

For the case when the i^{th} herd is not infected, $t_i=0$, there is no information about π_i, therefore π_i is not sampled. The parameters (α,β), of the beta distribution on the within-herd prevalence in the infected herds, are sampled independently using the method of adaptive rejection (Gilks and Wild, 1992). The sampling is implemented when two or more herds are positive, i.e., $\sum t_i \geq 2$, which gives a log-concave function. The sampled value of α is drawn from

$$\prod_{i=1}^{k}\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}\right]^{t_i} p(\alpha|t_i=1)$$

and similarly for β.

The remaining parameters in the model all have conditional beta distributions. The proportion of infected countries that are infected has the following conditional distribution

$$\gamma|Z \sim Beta\left(a_\gamma+Z,b_\gamma+1-Z\right).$$

The proportion of infected herds has the conditional distribution

$$\tau|Z=1,\{t_i\} \sim Beta\left(a_\tau+\sum_{i=1}^{k} t_i,b_\tau+k-\sum_{i=1}^{k} t_i\right).$$

The sensitivity of the diagnostic test used has the conditional distribution

$$\eta|\{X_{ij}\}, Z=1, \{v_{ij}\} \sim Beta\left(a_\eta + \sum_{ij} X_{ij}v_{ij}, b_\eta + \sum_{ij}(1-X_{ij})v_{ij}\right)$$

assuming at least one animal sampled is infected. Otherwise, η is not sampled. The specificity of the test used has the conditional distribution

$$\theta|\{X_{ij}\}, \{v_{ij}\} \sim Beta\left(a_\theta + \sum_{ij}(1-X_{ij})(1-v_{ij}), b_\theta + \sum_{ij} X_{ij}(1-v_{ij})\right).$$

And finally, for the country-level infection status if any v_{ij} or t_i is greater than zero, then $Z=1$ with probability 1, otherwise,

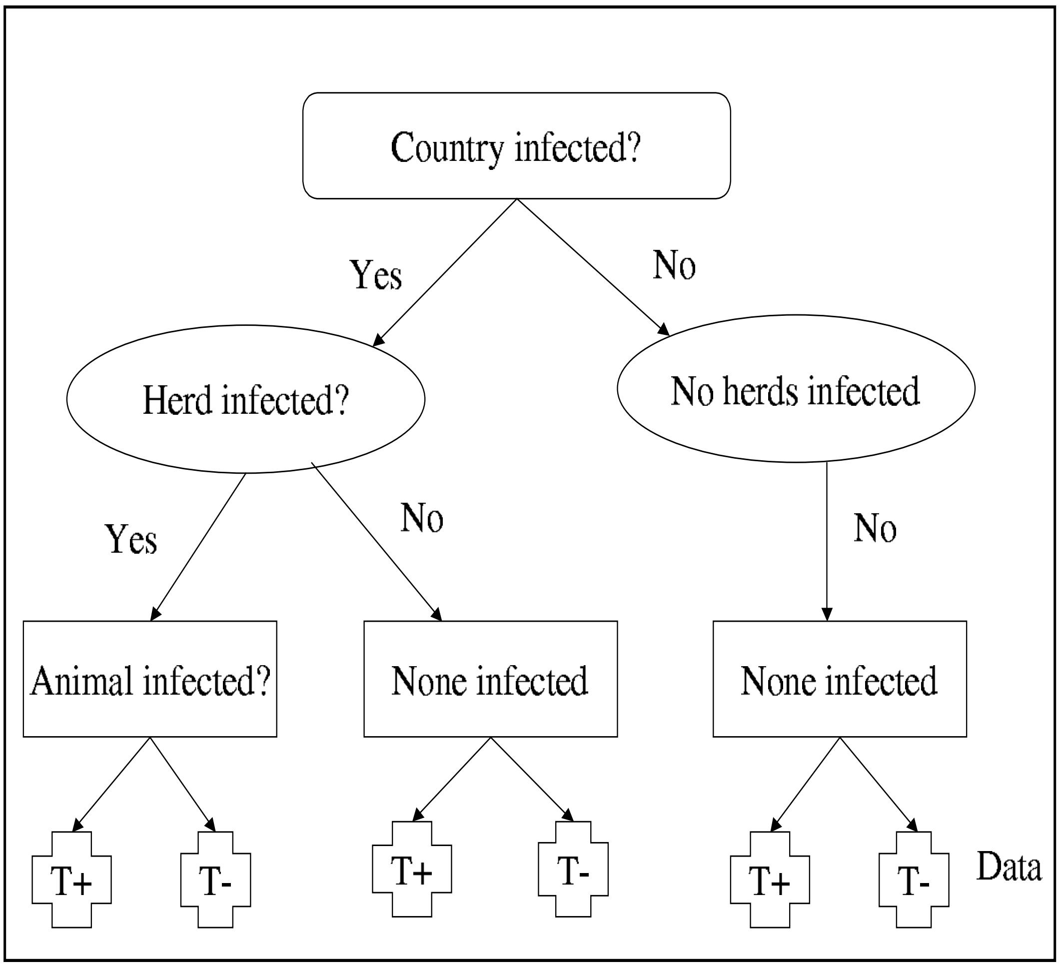
$$Z|\{X_{ij}\}, \{v_{ij}=0\}, \{t_i=0\}, \gamma, \tau \sim Ber\left(\frac{(1-\tau)^k\gamma}{(1-\tau)^k\gamma + 1(1-\gamma)}\right).$$

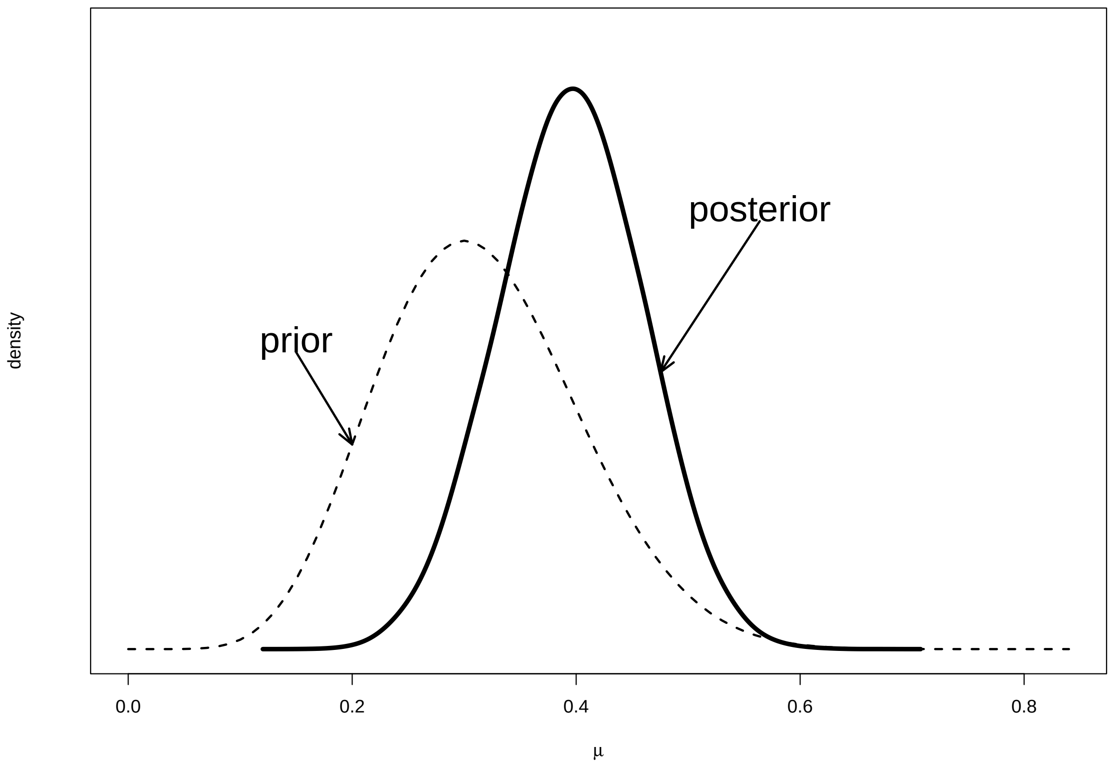
Figures

(1) Flowchart of the levels of questions that are asked to elicit information about the unknown true infection status of a country's animal population. Top is the country level, second is the herd level, third is the animal level, and the test results at the bottom indicate the data collected.

(2) Prior and posterior distribution of the mean, μ, of the prevalence distribution for the analysis of the ND virus survey data from Switzerland.

(3) Estimated prevalence distribution among infected herds, $\pi_i|t_i = 1$, for the analysis of the ND virus survey data from Switzerland.


Country infected?
Yes
No
Herd infected?
No herds infected
Yes
No
No
Animal infected?
None infected
None infected
T+
T-
T+
T-
T+
T-
Data


posterior
prior
density
0.0
0.2
0.4
0.6
0.8
μ

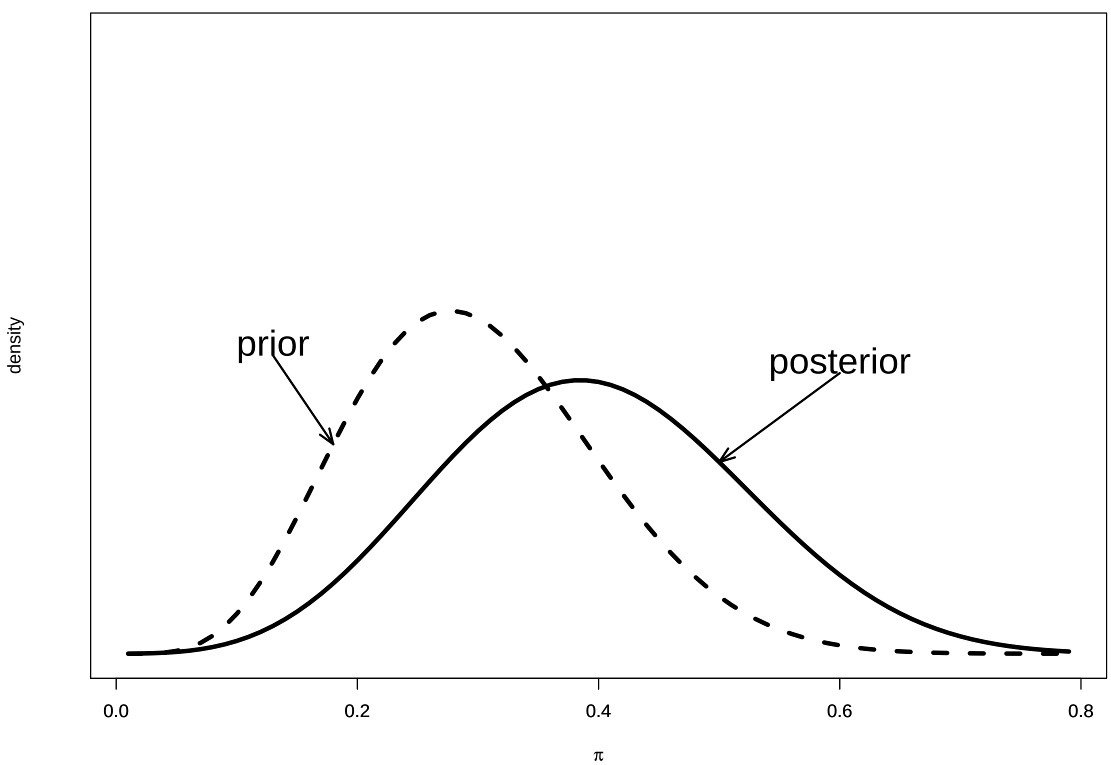

prior
posterior
density
0.0
0.2
0.4
0.6
0.8
π

Table 1

Parameters used in the hierarchical model.

Parameter	Definition
k	number of herds sampled, $i = 1, ..., k$
n_i	number of animals sampled within each herd $j = 1, ..., n_i$
X_{ij}	test result of j^{th} animal within i^{th} herd
Z	true infection-status of the country, infected/not
v_{ij}	true infection-status of j^{th} animal within i^{th} herd, infected/not
t_i	true infection-status of the i^{th} herd, infected/not
π_i	prevalence within the i^{th} herd if infected
(α, β)	unknown parameters for the beta distribution of π_i
μ	average prevalence among infected herds
σ	standard deviation of prevalences among the infected herds
λ_i	prevalence for the i^{th} herd (note: $\lambda_i = \pi_i t_i$)
γ	probability $Z = 1$
τ	proportion of infected herds
η	sensitivity, $P(+\|I)$
θ	specificity, $P(-\|\bar{I})$

Table 2

Description of the prior distributions for prevalence and ELISA accuracy for evaluation of the Newcastle disease virus survey data from Switzerland. Posterior modes, 95% intervals and other model outputs from the analysis. See Table 1 for key to notation.

	Prior				Posterior			
	Mode	95% Interval	Mean	SD	Mode	95% Interval	Mean	SD
τ	0.01	(0.002, 0.059)	0.020	0.149	0.018	(0.007, 0.035)	0.018	0.007
μ	0.30	(0.156, 0.500)	0.315	0.089	0.390	(0.269, 0.512)	0.386	0.062
σ	0.100	-	NA	-	0.106	(0.089, 0.130)	0.107	0.010
α	5.813	(2.914, 11.775)	6.603	2.284	7.217	(4.161, 12.265)	7.666	2.086
β	13.565	(8.531, 21.668)	14.35	3.367	11.62	(7.602, 17.640)	12.089	2.567
η	0.995	(0.965, 0.999)	0.988	0.009	0.995	(0.966, 0.999)	0.989	0.009
θ	0.995	(0.977, 0.999)	0.991	0.006	0.990	(0.988, 0.992)	0.990	0.001
γ	0.20	(0.055, 0.551)	0.259	0.131	0.301	(0.100, 0.611)	0.326	0.134

Table 3

Description of prior distributions for prevalence and ELISA accuracy for evaluation of PRRS virus survey data from Switzerland. Posterior modes, 95% intervals and other model outputs from the analysis. See Table 1 for key to notation.

	Prior				Posterior			
	Mode	95% Interval	Mean	SD	Mode	95% Interval	Mean	SD
θ	0.995	(0.965, 0.999)	0.988	0.009	0.999	(0.993, 1.000)	0.998	0.002
γ	0.01	(0.002, 0.059)	0.021	0.015	0.016	(0.002, 0.057)	0.021	0.015