Nonparametric Estimation of ROC Curves Based on Bayesian Models When the True Disease State Is Unknown

Chong Wang[1], Bruce W. Turnbull[1,4], Yrjö T. Gröhn[2] and Søren S. Nielsen[3]

[1]Department of Statistical Science, Cornell University, Ithaca, NY 14853, U.S.A.
[2]Department of Population Medicine and Diagnostic Sciences,
College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, U.S.A.
[3]Department of Large Animal Science, The Royal Veterinary and
Agricultural University, Frederiksberg, Denmark

[4] Corresponding author.

E-mail: cw245@cornell.edu, bwt2@cornell.edu, ytg1@cornell.edu, ssn@kvl.dk

July 17, 2006

Abstract

We develop a Bayesian methodology for nonparametric estimation of ROC curves used for evaluation of the accuracy of a diagnostic procedure. We consider the situation where there is no perfect reference test, i.e., no "gold standard". The method is based on a multinomial model for the joint distribution of test-positive and test-negative observations. We use a Bayesian approach which assures the natural monotonicity property of the resulting ROC curve estimate. MCMC methods are used to compute the posterior estimates of the sensitivities and specificities that provide the basis for inference concerning the accuracy of the diagnostic procedure. Because there is no gold standard, identifiability requires that the data come from at least two populations with

different prevalences. No assumption is needed concerning the shape of the distributions of test values of the diseased and non-diseased in these populations. We discuss an application to an analysis of ELISA scores in the diagnostic testing of paratuberculosis (Johne's Disease) for several herds of dairy cows and compare the results to those obtained from some previously proposed methods.

Key Words: Bayesian, Johne's Disease, Markov chain Monte Carlo, No Gold Standard, Nonparametric, ROC curves

1 Introduction

In the diagnostic testing of certain diseases, "gold standard" (GS) tests (i.e., those with error-free classification of diseased and disease-free individuals) are often too expensive to use on a large scale, or may not even exist. Therefore, imperfect diagnostic tests are widely used in medical and epidemiological research, especially when investigating disease in large populations (Hui and Walter 1980; Joseph, Gyorkos, and Coupal 1995; Nielsen, Gronbak, Agger, and Houe 2002). Yet when no GS tests are available for comparison, the performances of these imperfect tests are difficult to evaluate.

This paper was motivated by the need for accurate diagnostic tests for Johne's disease in cattle. As part of the New York State Cattle Health Assurance Program, the Animal Health Diagnostic Laboratory at Cornell University provides diagnostic services to dairy herds throughout the northeastern United States. More than 760 herds and 80,000 cows are tested annually. A recent report of the National Research Council of the U.S. National Academies of Science (Rideout et al. 2003) has identified Johne's disease as a significant animal health problem whose study and control deserves high priority from the USDA and other national and state agencies. It recognized that one of the problems with Johne's disease is the lack of specific and sensitive tests for detecting early infections. In particular,

no gold standard test exists. Therefore statistical methods are needed to assess both existing diagnostic tests and ones yet to be proposed.

Consider first diagnostic tests with binary outcomes (positive or negative). Sensitivity (Se, probability of a positive outcome in a diseased individual) and specificity (Sp, probability of a negative outcome in a non-diseased individual) are used to assess the accuracy of such tests – see e.g., Pepe (2003, Chapter 2). In the case when there is no gold standard to determine an individual's disease status, methods have been developed to evaluate the sensitivities and specificities of two imperfect tests by comparing one test to the other. A simple and popular maximum likelihood (ML) approach was introduced by Hui and Walter (1980), assuming the existence of two or more population strata with different prevalences. They also require the assumption of independence of test results conditional on disease status. Joseph et al. (1995) developed Bayesian methods for evaluating conditionally independent diagnostic tests, avoiding the need to stratify the population. However, their method has been criticized by Andersen (1997) and Johnson, Gastwirth, and Pearson (2001) for lack of identifiability; that is, the posterior distributions for some of the quantities of interest need not become concentrated around their true values as the sample size increases. Several models allowing for conditional dependence have been proposed using either ML or Bayesian estimation, e.g., Qu, Tan, and Kutner (1996), Yang and Becker (1997), Dendukuri and Joseph (2001), Black and Craig (2002), Garrett, Eaton, and Zeger (2002), Georgiadis, Johnson, Gardner, and Singh (2003) and Hanson, Johnson, and Gardner (2003). Reviews of methods for estimating the validity of two imperfect binary tests are presented in Walter and Irwig (1988), Hui and Zhou (1998), Enøe, Georgiadis, and Johnson (2000) and Branscum, Gardner, and Johnson (2005). All these methods deal with evaluation of imperfect binary tests and cannot be applied to multichromatic or continuous-scaled tests.

The measurement scale of most serodiagnostic tests is usually ordinal or continuous. A

value on this scale is selected as a decision threshold (cutoff value) to define positive and negative test outcomes. Of course, sensitivity and specificity of a diagnostic test depend on the choice of cutoff value and are inversely related. To assess the accuracy of a diagnostic procedure, it is useful to consider the receiver operating characteristic (ROC) curve, which is a graph of pairs of sensitivity and 1-specificity values that result as the test's cutoff value is varied. ROC analysis is important for full test evaluation and test comparison, as well as for the optimal choice of cutoff value (Greiner, Pfeiffer, and Smith 2000). Principles of ROC curve estimation using parametric and nonparametric methods are well described in Pepe (2003). Traditional ROC analysis assumes the existence of a GS reference test that has perfect sensitivity and specificity. Less work has been done for estimating ROC curves without gold standard tests; this is summarized in the following two paragraphs.

When trying to compute an ROC curve when the true disease state is unknown, one approach in the literature has been to estimate the pair (sensitivity, 1-specificity) for every single cutoff-value separately by the maximum likelihood (ML) method of Hui and Walter (1980), and then link them together to get the curve (Nielsen et al. 2002). We term this the "separate MLE" method. However, an ROC curve with (Se,1-Sp) produced by this method is not necessarily monotone, as in Nielsen et al. (2002, Figures 2, 3, 4, 5), and as we will see in Figure 1 later in Section 4 of this paper. This non-monotonicity phenomenon is due to the random aspect of the counts observed, and it contradicts the fact that ROC curves should be monotone. Henkelman, Kay, and Bronskill (1990) have proposed a maximum likelihood estimation method for the ROC curve of an ordinal-scale test using a multivariate normal mixture latent model, however several limitations were pointed out by Begg and Metz (1990), including the necessity to make a normality assumption . The method of Henkelman et al. (1990) was later developed further by Beiden, Campbell, Meier, and Wagner (2000). Choi, Johnson, Collins, and Gardner (2006) have proposed a parametric Bayesian method for ROC

curve estimation in the absence of a gold standard where the test values (or transformed test values) of both diseased and non-diseased individuals were measured on a continuous-scale and normally distributed. Both methods of Henkelman et al. (1990) and Choi et al. (2006) guarantee monotonicity; however, in the absence of a GS test, it is impossible to determine the validity of the parametric distribution assumptions. Choi et al. (2006) have discussed this in their paper and commented that a nonparametric approach could overcome such difficulties.

Hall and Zhou (2003) proposed a nonparametric estimator for the ROC curves of continuous-scale tests under the conditional independence assumption when the number of tests is three or more. Zhou, Castelluccio, and Zhou (2005) applied the idea of Hall and Zhou (2003) to estimate an ROC curve and its associated AUC (area under curve) of a ordinal-scale test with the EM algorithm. The methods of both articles deal specifically with the situation of one population (i.e., probability of disease must be same for every individual in the study); thus identifiability requires at least three tests available. Also in the work of Zhou et al. (2005), all the different tests should be based on the same ordinal scale. The computations in the methods of both articles are complicated by the existence of many local maxima of the likelihood function, as discussed by the authors.

For our study of Johne's Disease, we need to evaluate an imperfect continuous-scaled test by comparing it to an imperfect reference binary test (see Section 4 for further details). Hence, the method by Nielsen et al. (2002) is the only one in the current literature that could be applied in this situation. Yet, as we have discussed previously, this method can suffer from the disadvantage of producing a non-monotone estimated ROC curve. In this paper, we propose a nonparametric method based on Bayesian models to estimate the (Se, 1-Sp) pairs jointly, without any assumption concerning the shapes of the distributions of test values. MCMC methods are used to compute the posterior estimates of the sensitivities and

specificities that provide the basis for inference concerning the accuracy of the diagnostic procedure. Our approach assures the monotone property of the resulting ROC curves. We discuss an application to the assessment of ELISA scores measured on a continuous scale for the diagnostic testing of Johne's Disease by comparing with fecal culture (FC) test result – an imperfect binary test. Our data come from 2662 cows in 55 herds. We also carry out the method of Nielsen et al. (2002) and compare the results.

In Section 6 we discuss extensions of our methods. In particular, we indicate how our method can be generalized when the scale of the imperfect reference test is ordinal or continuous.

2 Ordered Multinomial Model Structure

The goal is to estimate the ROC curve of a new diagnostic procedure (Test 1) measured on an ordinal or continuous scale by comparing it to an imperfect binary reference test (Test 2). Let the two kinds of tests be applied simultaneously to each individual in samples from G populations (or G strata within a population). Without loss of generality, high values for Test 1 are associated with a high likelihood of presence of disease. For any binary test T, denote the event that the test is positive for a given individual by $T+$, and negative by $T-$. For a given value of K, we select K increasing cutoff values for Test 1. Here K and the cutoff values may be chosen arbitrarily; we discuss desirable choices later. For the i'th cutoff value ($1 \leq i \leq K$) of Test 1, let $T_{1,i}$ denote the corresponding dichotomized test. Let $\alpha_{1,i}$, $\beta_{1,i}$ be the corresponding unknown false positive rate (1−specificity) and false negative rate (1−sensitivity), respectively, of this test $T_{1,i}$. In addition define $T_{1,0}$ as a test that always yields a positive outcome and $T_{1,K+1}$ as a test that is always negative. Correspondingly, define $\alpha_{1,0} = \beta_{1,K+1} = 1$ and $\alpha_{1,K+1} = \beta_{1,0} = 0$. Note that, because the K cutoff values for

Test 1 are increasing, the error rates are ordered:

$$1 = \alpha_{1,0} \geq \alpha_{1,1} \geq \alpha_{1,2} \geq ... \geq \alpha_{1,K} \geq \alpha_{1,K+1} = 0,$$

$$0 = \beta_{1,0} \leq \beta_{1,1} \leq \beta_{1,2} \leq ... \leq \beta_{1,K} \leq \beta_{1,K+1} = 1.$$

Let α_2, β_2 be the unknown false positive and false negative rates of Test 2, respectively. Recall Test 2 is measured on a binary scale. We assume that these error rates for Test 1 and Test 2 do not depend on the population g. This is reasonable as they are typically solely properties of the tests themselves. Note that if Test 2 is based on an ordinal or continuous scale, it could be dichotomized using a single cutoff value in order to use the methodology described here. However, this will lead to some consequent loss of information. We discuss this point further in Section 6.

In the gth population ($1 \leq g \leq G$), let θ_g be the unknown prevalence of disease and N_g be the sample size. Of these N_g suppose m_g are declared positive by Test 2 and $n_g = N_g - m_g$ are negative by Test 2. Using the i'th cutoff value ($1 \leq i \leq K$) for Test 1, let $y_{i,g,1}$, be the frequency of $(T_{1,i}+, T_2+)$ outcomes, that is the number of individuals in population g that test positive both by test $T_{1,i}$ and by test T_2. Similarly define $y_{i,g,2}$ to be the frequency of $(T_{1,i}+, T_2-)$ outcomes ($1 \leq i \leq K, 1 \leq g \leq G$). Further, corresponding to our definitions of $T_{1,0}$ and $T_{1,K+1}$, we define $y_{0,g,1} = m_g, y_{0,g,2} = n_g$ and $y_{K+1,g,1} = y_{K+1,g,2} = 0$. Note that, because the K cutoff values for Test 1 are increasing, we have that the frequencies are decreasing. That is

$$m_g = y_{0,g,1} \geq y_{1,g,1} \geq \ldots \geq y_{K,g,1} \geq y_{K+1,g,1} = 0,$$

$$n_g = y_{0,g,2} \geq y_{1,g,2} \geq \ldots \geq y_{K,g,2} \geq y_{K+1,g,2} = 0.$$

Assuming independence between the two tests conditional on the true disease state, we

can write down expressions for the corresponding probabilities:

$$\Pr\{T_{1,i}+, T_2+\} = \theta_g(1-\beta_{1,i})(1-\beta_2) + (1-\theta_g)\alpha_{1,i}\alpha_2, \quad (1)$$

$$\Pr\{T_{1,i}+, T_2-\} = \theta_g(1-\beta_{1,i})\beta_2 + (1-\theta_g)\alpha_{1,i}(1-\alpha_2). \quad (2)$$

To obtain the entire ROC curve for Test 1, we now need consider all the possible K cutoff values for Test 1 simultaneously. For $i = 1, \ldots, K+1$ and $g = 1, \ldots G$, let

$$p_{i,g} = \Pr\{T_{1,i-1}+, T_{1,i}-, T_2+\} = \Pr\{T_{1,i-1}+, T_2+\} - \Pr\{T_{1,i}+, T_2+\} \quad (3)$$

be the probability that an individual in group g, who was positive by Test 2, was positive by Test 1 using the $(i-1)$th cutoff but negative using the ith cutoff. Similarly define

$$q_{i,g} = \Pr\{T_{1,i-1}+, T_{1,i}-, T_2-\} = \Pr\{T_{1,i-1}+, T_2-\} - \Pr\{T_{1,i}+, T_2-\} \quad (4)$$

for individuals who were negative by Test 2. Finally let the observed frequencies, corresponding to the $\{p_{i,g}\}$ and $\{q_{i,g}\}$, be denoted by:

$$x_{i,g,1} = y_{i-1,g,1} - y_{i,g,1},$$

$$x_{i,g,2} = y_{i-1,g,2} - y_{i,g,2}$$

for $i = 1, \ldots, K+1$ and $g = 1, \ldots, G$, respectively. These frequencies and probabilities for population g can be displayed in a $2 \times (K+1)$ table — see Table 1. In this table, the columns labelled $\{T_{1,i}-, T_{1,i-1}+\}$ are for individuals classified positive by $T_{1,i-1}$ but negative by $T_{1,i}$. These are divided between those classified positive by Test 2 (first row) and those classified negative by Test 2 (second row).

[Table 1 about here.]

Using (1–4), we have that the cell probabilities are given by:

$$p_{i,g} = \theta_g b_i(1-\beta_2) + (1-\theta_g)a_i\alpha_2, \quad (5)$$

$$q_{i,g} = \theta_g b_i\beta_2 + (1-\theta_g)a_i(1-\alpha_2) \quad (6)$$

where, for $i = 1, 2, .., K+1$, we have defined:

$$a_i = \alpha_{1,i-1} - \alpha_{1,i}, \tag{7}$$

$$b_i = \beta_{1,i} - \beta_{1,i-1} \tag{8}$$

and where $\mathbf{a} = (a_1, \ldots, a_{K+1})$, etc. Note relationships (7) and (8) define one-to-one mappings between the $\{\alpha_i\}$ and the $\{a_i\}$ and between the $\{\beta_i\}$ and the $\{b_i\}$. Also note that:

$$\begin{aligned} \sum_{i=1}^{K+1} a_i &= \sum_{i=1}^{K+1} \alpha_{1,i-1} - \alpha_{1,i} = \alpha_{1,0} - \alpha_{1,K+1} = 1, \\ \sum_{i=1}^{K+1} b_i &= \sum_{i=1}^{K+1} \beta_{1,i} - \beta_{1,i-1} = \beta_{1,K+1} - \beta_{1,0} = 1. \end{aligned}$$

The observed frequencies $\{x_{i,g,1}\}$ and $\{x_{i,g,2}\}$ for $1 \leq i \leq K+1,\ 1 \leq g \leq G$ have a likelihood of a product of G multinomials each with $K+1$ categories:

$$\begin{aligned} &L(\mathbf{x}|\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}) \\ &\propto \prod_{g=1}^{G} \prod_{i=1}^{K+1} [\theta_g b_i (1-\beta_2) + (1-\theta_g) a_i \alpha_2]^{x_{i,g,1}} \cdot [\theta_g b_i \beta_2 + (1-\theta_g) a_i (1-\alpha_2)]^{x_{i,g,2}} \end{aligned} \tag{9}$$

where $\mathbf{a} = (a_1, \ldots, a_{K+1})$ etc.

In the model, there are a total of $2K + G + 2$ unknown parameters, namely $\alpha_{1,1}, ..., \alpha_{1,K}$, $\beta_{1,1}, ..., \beta_{1,K}$, $\theta_1, ..., \theta_G$, α_2, β_2, and $G(2K+1)$ degrees of freedom. Thus we need $G \geq 2$ with at least two of the θ_g-values distinct in order to make the model identifiable. This can also be seen by examining the likelihood (9) directly. In practice, no two distinct populations will have identical prevalences. However, the true values could be "close"; we discuss this situation further in Section 5. In addition, we see that we also need $\alpha_2 + \beta_2 \neq 1$ for the model to be identifiable for otherwise (9) depends on $\theta_g, \{a_i\}, \{b_i\}$ only through the $\{\theta_g b_i + (1-\theta_g) a_i\}$. Intuitively, $\alpha_2 + \beta_2 = 1$ implies that Test 2 is no better than random guessing and so contributes no information. Presumably, such a test would not be under

consideration – see further comments in Section 5. It can also be seen that (9) remains unchanged under the transformation $b_i' = a_i,\ a_i' = b_i, \theta_g' = 1 - \theta_g, \alpha_2' = 1 - \beta_2,\ \beta_2' = 1 - \alpha_2$. Hence for identifiability, we may assume $\alpha_2 + \beta_2 < 1$. In fact, this is without loss of generality: If $\alpha_2 + \beta_2 > 1$, which is worse than random guessing, we could always use the test criterion in the opposite direction to obtain a test which is better than random guessing and now has $\alpha_2 + \beta_2 < 1$.

In order to guarantee the monotonicity of the ROC curve, $\mathbf{a}$ and $\mathbf{b}$ need to be constrained as non-negative. We implement this by taking a Bayesian approach and placing Dirichlet priors on the parameters $\mathbf{a}$ and $\mathbf{b}$.

3 The Bayesian Approach

Suppose we specify a prior distribution with density $\pi(\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta})$ for all the parameters in our model. Here we use appropriate conjugate priors, as will be described later in this section. In the Bayesian approach, inference concerning the parameters is based on the posterior density:

$$\begin{aligned} &\pi(\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}|\mathbf{x}) \qquad (10)\\ &\quad \propto \quad L(\mathbf{x}|\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}) \cdot \pi(\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}) \\ &\quad \propto \quad \prod_{g=1}^{G} \prod_{i=1}^{K+1} [\theta_g b_i (1 - \beta_2) + (1 - \theta_g) a_i \alpha_2]^{x_{i,g,1}} \cdot [\theta_g b_i \beta_2 + (1 - \theta_g) a_i (1 - \alpha_2)]^{x_{i,g,2}} \\ &\qquad\qquad \times \pi(\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}). \end{aligned}$$

The Bayesian estimates of the parameters are given by their means from this posterior distribution. These posterior means are computed using Markov chain Monte Carlo (MCMC) methods — e.g., see Robert and Casella (2004). In particular, we propose to use the Gibbs sampling algorithm. The Gibbs sampler is an iterative algorithm for generation of samples of variables (parameter values) from the posterior multivariate distribution. It proceeds

by successively updating each variable by sampling from its conditional distribution given current values of all other variables. After a sufficiently large number of iterations, under mild conditions it can be proven that the values of the updated variables so obtained form a sample from the joint posterior distribution — see for example, Robert and Casella (2004).

First however, we note from inspection of (10) that it is not easy to construct the conditional distributions needed for the Gibbs sampler directly from the density. In order to take advantage of Gibbs sampling, we solve this problem by using data augmentation as described by Tanner and Wong (1987). This is as follows.

In the g'th population, the vector $\mathbf{x}_g = (x_{1,g,1}, ..., x_{K+1,g,1}, x_{1,g,2}, ..., x_{K+1,g,2})$ has a 2(K+1)-dimensional multinomial likelihood, namely

$$\mathbf{x}_g \sim \text{Mul}_{2(K+1)}(N_g; p_{1,g}, ..., p_{K+1,g}, q_{1,g}, ..., q_{K+1,g})$$

where N_g is the sample size of the g'th population.

As calculated before,

$$\begin{aligned} p_{i,g} &= \theta_g b_i(1-\beta_2) + (1-\theta_g)a_i\alpha_2, \\ q_{i,g} &= \theta_g b_i\beta_2 + (1-\theta_g)a_i(1-\alpha_2). \end{aligned}$$

This may be viewed as a grouped multinomial problem, if we introduce new random vectors $\mathbf{u}_g = (u_{1,g,1}, ..., u_{K+1,g,1}, u_{1,g,2}, ..., u_{K+1,g,2})$ for all $g = 1, .., G$. Let

$$\begin{aligned} &(u_{1,g,1}, x_{1,g,1} - u_{1,g,1}, ..., u_{K+1,g,1}, x_{K+1,g,1} - u_{K+1,g,1}, \\ &\qquad u_{1,g,2}, x_{1,g,2} - u_{1,g,2}, ..., u_{K+1,g,2}, x_{K+1,g,2} - u_{K+1,g,2}) \\ &\sim \text{Mul}_{4(K+1)}(N_g; \theta_g b_1(1-\beta_2), (1-\theta_g)a_1\alpha_2, ..., \theta_g b_{K+1}(1-\beta_2), (1-\theta_g)a_{K+1}\alpha_2, \\ &\qquad \theta_g b_1\beta_2, (1-\theta_g)a_1(1-\alpha_2), ..., \theta_g b_{K+1}\beta_2, (1-\theta_g)a_{K+1}(1-\alpha_2)). \end{aligned}$$

Then the marginal distribution of $\mathbf{x}_g$ will remain the same as in the original model. However the augmented data likelihood in the new multinomial distribution is of a more tractable

form, permitting use of conjugate priors:

$$
\begin{aligned}
&L(\mathbf{x},\mathbf{u}|\mathbf{a},\mathbf{b},\alpha_2,\beta_2,\boldsymbol{\theta}) \\
&\quad\propto \prod_{g=1}^{G}\prod_{i=1}^{K+1}\{[\theta_g b_i(1-\beta_2)]^{u_{i,g,1}}\cdot[(1-\theta_g)a_i\alpha_2]^{x_{i,g,1}-u_{i,g,1}} \\
&\qquad\qquad\qquad\qquad \cdot[\theta_g b_i\beta_2]^{u_{i,g,2}}\cdot[(1-\theta_g)a_i(1-\alpha_2)]^{x_{i,g,2}-u_{i,g,2}}\} \\
&\quad= [\prod_{i=1}^{K+1} a_i^{\sum_{g=1}^{G}(x_{i,g,1}-u_{i,g,1}+x_{i,g,2}-u_{i,g,2})}]\cdot[\prod_{i=1}^{K+1} b_i^{\sum_{g=1}^{G}(u_{i,g,1}+u_{i,g,2})}]\cdot\alpha_2^{\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,1}-u_{i,g,1})} \\
&\qquad\qquad\cdot(1-\alpha_2)^{\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,2}-u_{i,g,2})}\cdot\beta_2^{\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,2}}\cdot(1-\beta_2)^{\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,1}} \\
&\qquad\qquad\cdot[\prod_{g=1}^{G}\theta_g^{\sum_{i=1}^{K+1}(u_{i,g,1}+u_{i,g,2})}(1-\theta_g)^{\sum_{i=1}^{K+1}(x_{i,g,1}-u_{i,g,1}+x_{i,g,2}-u_{i,g,2})}].
\end{aligned}
$$

Now $\sum_{i=1}^{K+1} a_i = \sum_{i=1}^{K+1} b_i = 1$, so the natural conjugate priors for $\mathbf{a}$, $\mathbf{b}$ are Dirichlet distributions, while for $\alpha_2, \beta_2, \theta_1, \ldots, \theta_G$ they are all Beta distributions. Specifically, as a prior for $\mathbf{a}$, we take $Dirich(h_1, ..., h_{K+1})$:

$$\pi(\mathbf{a}) \propto \prod_{i=1}^{K+1} a_i^{h_i-1};$$

for $\mathbf{b}$, we use prior $Dirich(h_{K+2}, ..., h_{2K+2})$:

$$\pi(\mathbf{b}) \propto \prod_{i=1}^{K+1} b_i^{h_{K+1+i}-1};$$

for α_2, we use prior $Beta(h_{2K+3}, h_{2K+4})$:

$$\pi(\alpha_2) \propto \alpha_2^{h_{2K+3}-1}(1-\alpha_2)^{h_{2K+4}-1};$$

for β_2, we use prior $Beta(h_{2K+5}, h_{2K+6})$:

$$\pi(\beta_2) \propto \beta_2^{h_{2K+5}-1}(1-\beta_2)^{h_{2K+6}-1};$$

and for θ_g, $g = 1, ..., G$, we use prior $Beta(h_{2K+5+2g}, h_{2K+6+2g})$:

$$\pi(\theta_g) \propto \theta_g^{h_{2K+5+2g}-1}(1-\theta_g)^{h_{2K+6+2g}-1}.$$

Informative priors will be preferred whenever previous knowledge is available for any of the parameters (error rates or prevalences). Otherwise $h_1 = ... = h_{K+1} = 1/2$, $h_{K+2} = ... = h_{2K+2} = 1/2$, $h_{2K+3} = h_{2K+4} = 1/2$, $h_{2K+5} = h_{2K+6} = 1/2$, $h_{2K+5+2g} = h_{2K+6+2g} = 1/2$ can be taken as noninformative priors for $\mathbf{a}$, $\mathbf{b}$, α_2, β_2, θ_g respectively. Sensitivity analyses for different priors have been performed by Gustafson (2005) in the context of binary tests evaluation without a GS. Our priors are natural to use for the models in this paper. As discussed in the previous section, we add the identifiability constraint $\alpha_2 + \beta_2 < 1$ into the prior.

Finally, the augmented data posterior is obtained as:

$$
\begin{aligned}
&\pi(\mathbf{a}, \mathbf{b}, \alpha_2, \beta_2, \boldsymbol{\theta}|\mathbf{x}, \mathbf{u}) \qquad (11)\\
&\propto \left[\prod_{i=1}^{K+1} a_i^{h_i - 1 + \sum_{g=1}^{G}(x_{i,g,1} - u_{i,g,1} + x_{i,g,2} - u_{i,g,2})}\right] \times \left[\prod_{i=1}^{K+1} b_i^{h_{K+1+i} - 1 + \sum_{g=1}^{G}(u_{i,g,1} + u_{i,g,2})}\right]\\
&\quad \times \left[\alpha_2^{h_{2K+3} - 1 + \sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,1} - u_{i,g,1})} \cdot (1 - \alpha_2)^{h_{2K+4} - 1 + \sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,2} - u_{i,g,2})}\right]\\
&\quad \times \left[\beta_2^{h_{2K+5} - 1 + \sum_{g=1}^{G}\sum_{i=1}^{K+1} u_{i,g,2}} \cdot (1 - \beta_2)^{h_{2K+6} - 1 + \sum_{g=1}^{G}\sum_{i=1}^{K+1} u_{i,g,1}}\right]\\
&\quad \times \left[\prod_{g=1}^{G} \theta_g^{h_{2K+5+2g} - 1 + \sum_{i=1}^{K+1}(u_{i,g,1} + u_{i,g,2})} (1 - \theta_g)^{h_{2K+6+2g} - 1 + \sum_{i=1}^{K+1}(x_{i,g,1} - u_{i,g,1} + x_{i,g,2} - u_{i,g,2})}\right]\\
&\quad \times I(\alpha_2 + \beta_2 < 1).
\end{aligned}
$$

We can update the values in the chain by Gibbs sampling from the full conditional probabilities. These are given in the Appendix. The Gibbs sampling is programmed in MATLAB version 7.0 (The MathWorks, Inc. 2004). We have found this approach to be preferred to WinBUGS (Spiegelhalter, Thomas, Best, and Gilks 1995) on grounds of computing speed. Source code is available upon request from the first author.

4 Application to Johne's Disease

As stated in Section 1, our original motivation for the methodology developed in this paper came from the diagnostic testing for paratuberculosis (Johne's Disease) in cattle at the Animal Health Diagnostic Laboratory at Cornell University . For this disease there is no agreed GS test. Two different (imperfect) tests are employed: fecal culture (FC), a test with categorical outcomes (here, just two – positive or negative); and enzyme-linked immunosorbent assay (ELISA), the readings of which can be regarded as continuous. Because the ELISA test takes a relatively short time and is inexpensive, the Laboratory is interested in evaluating its performance.

The conditional independence of the FC and ELISA tests can be justified on biological grounds. FC measures bacterial shedding, which can occur from the time of infection to the death of the animal, though the probability increases with progression of the disease. ELISA measures occurrence of antibodies, which require that a shift from cell-mediated immune responses to humoral immune responses has occurred (Stabel 2000).

4.1 Johne's Disease Data Analysis

In this section, the ELISA test (Test 1) is to be evaluated against the FC test (Test 2) for the diagnosis of Johne's Disease. The cross-sectional data set to be used is from part of a multipurpose study on infectious disease in a study carried out by the Danish dairy board in Southern Jutland, Denmark, in 1998 (Andersen et al. 2000). It is similar to the data sets that have been used in Nielsen et al. (2002). The 2662 cows from 55 herds in the data set are divided into two groups (G=2) by veterinary practitioner of the farm. The split was based on the median of an ordered list of the veterinary official federal practice codes. This criterion was used by Nielsen et al. to select the groups and, according to their Table 3, it

would seem reasonable to assume the two groups have different prevalences. They considered two other ways to select the groups which we will discuss later in Section 4.2. The ELISA test readings were analyzed based on corrected optical density (OD_C) measurements. There are a total of 1462 distinct corrected OD_C-values, all of which could be used as effective cutoff values in the analysis. After ordering the corrected OD_C-values, we chose the set of $\{1/(K+1), 2/(K+1), ..., K/(K+1)\}$-quantiles of the values to be used as cutoff values in the model. Such a choice insures balanced frequencies in the columns of Table 1. Here we considered several alternative values for K, namely $K = 10, 20, 50, 100$.

Noninformative priors where $h_1 = ... = h_{2K+6+2G} = 1/2$ have been used in this analysis to compare the result with the separate ML approach. The method of Choi et al. (2006) cannot be applied in this case because the outcome of the reference FC test is not measured on a continuous scale. Four models with different numbers of cutoff values have been implemented to analyze the data, where K = 10, 20, 50, 100, respectively. The results are shown in Figure 1. All four models have been run for 100,000 iterations and the samples from 50,001 to 100,000 are used to compute summary statistics.

[Figure 1 about here.]

As can be seen in Figure 1, the four Bayesian ROC estimates are very close in both shape and position to the one produced by the method of separate ML estimates. However, the Bayesian method produces monotone estimated ROC curves, which the other method does not. The values of the area under the ROC curve (AUC) are 0.784, 0.785, 0.758, 0.748, for K = 10, 20, 50, 100 respectively. As the number of cutoff values K increases from 10 to 100, there are more points on the curve being estimated so that the ROC estimate is capable of capturing more details. But, on the other hand, adding each new cutoff value will introduce two more prior parameters into the model. So without increasing the sample size of the

data, increasing K alone will make the model more sensitive to the prior information. In practice, a larger K also leads to a larger number of samplings in each Gibbs iteration, thus increasing the running time of the program. For the four cases where $K = 10$, 20, 50, 100, the running times of the MATLAB program on a 2.80 GHz Pentium 4 processor per 1000 iterations are 3.67, 4.43, 5.56, 17.71 seconds, respectively.

[Figure 2 about here.]

Multiple chains have been simulated and pass the convergence criterion after the first 5,000 iterations. Convergence has been checked using CODA (Convergence Diagnostic and Output Analysis) software (Best, Cowles, and Vines 1995) and passes all six different convergence diagnostic criteria. For the particular model with $K = 20$, the first 10,000 iterations of the error rates $\alpha_{1,1}$, $\alpha_{1,10}$, $\alpha_{1,19}$, $\beta_{1,1}$, $\beta_{1,10}$, $\beta_{1,19}$ are plotted in Figure 2. Figure 3 (upper panel) shows estimated ROC curve based on the posterior means. Also shown are the $\pm$ standard deviation band and the 95% credible band based on 2.5% and 97.5% quantiles of the posterior distribution. In the lower panel of Figure 3, we reproduce our Bayesian estimate of the ROC curve (a), along with the estimate based on using the separate MLE method (b) of Nielsen et al. (2002) and (c) the usual estimate that would be obtained if the fecal culture test was incorrectly assumed to be a GS test. Note that the estimate (b) is non-monotone and that the estimate (c) lies below (a) with a lower AUC.

The estimates (95% credible interval) for the specificity (1-α_2) and sensitivity (1-β_2) for the FC test are estimated as 0.988 (0.975,0.999) and 0.830 (0.507,0.999), respectively. The estimate of sensitivity of the FC test may seem higher than generally accepted; however, its credible interval is wide and contains most of the different estimates reported in Nielsen et al. (2002, Table 5), which were based on the same data set. The estimate for FC sensitivity could be interpreted as the sensitivity of an animal with some level of antibodies. The

prevalences (θ_1, θ_2) in the two populations are estimated as 0.030 and 0.080, with a 95% credible interval for the difference being (0.027,0.083). Note that this interval excludes zero. These estimates are of the same order of magnitude that have been reported in other studies, – e.g., Nielsen et al. (2002, Table 3 and 5).

[Figure 3 about here.]

4.2 Robustness with respect to division of population

Both the ML estimation and the Bayesian methods require division of the full population into two or more distinct subpopulations with different prevalences. In their study of the same Johne's disease data set, Nielsen et al. (2002) examined three division criteria, all of which are based on non-biological risk factors. Besides veterinary practitioner of the farm (Vet) which we used in Section 4.1, their other two criteria were: herd size (Size) and postal zip code (Zip). Size was divided at 100 cows, Zip was divided based on the median of postal codes in an ordered list.

We repeated our analyses of Section 4.1 for the other two methods of division. The ROC curve estimates obtained by the separate ML method (upper panel of Figure 4) appear to be different in shape and position for the three division criteria. However, as can be seen in the lower panel of Figure 4, the Bayesian ROC estimates obtained from the three division criteria are almost the same. Intuitively, in the Bayesian model, the information from different cutoff values are considered jointly, and thus the aberrant observations can not affect the ROC curve estimates as much as they would if the separate ML method were used. Thus the nonparametric method we propose here should be more robust to the impact of different division criteria on the ROC estimation. The estimates of the two prevalence quantities corresponding to the three ways to split the population were Vet: 0.030, 0.080 (as

in Section 4.1); Size: 0.044, 0.107; and Zip: 0.013, 0.068. Also, the 95% credible intervals for the difference in prevalences excluded 0 for all three division methods.

[Figure 4 about here.]

5 Discussion

It is important to investigate how well the proposed ROC estimation procedure performs. To answer this question, a moderate size simulation study was undertaken. First, a hundred data sets mimicking the Johne's disease data set analyzed in Section 4.1 were simulated. Each data set contained two populations ($G = 2$), with 1250 individuals in each ($n_1 = n_2 = 1250$). For each individual in the g'th population ($g = 1, 2$), we first simulated the true disease status (0 for healthy and 1 for diseased) according to a binomial distribution $\text{Bin}(1, \theta_g)$, using prevalences $\theta_1 = 0.03,\ \theta_2 = 0.08$. Then, for each diseased individual in either population, a Test 1 score is generated from a normal distribution $\text{N}(\mu_1 = +0.33, \sigma_1^2 = 0.63\ )$ and a Test 2 score of 0 (negative) or 1 (positive) is generated with probabilities $\beta_2 = 0.17$, and $1 - \beta_2 = 0.83$, respectively. Similarly, for each healthy individual, a Test 1 score is generated from a $\text{N}(\mu_0 = -0.33, \sigma_0^2 0.12)$ distribution and a Test 2 score of 0 or 1 is generated with probabilities $1 - \alpha_2 = 0.98$, and $\alpha_2 = 0.02$, respectively. These α and β values were taken from the estimates reported in Section 4.1; the μ and σ values for the binormal model of Test 1 scores lead to an ROC curve similar to our Bayesian estimate in Figure 3 with a similar AUC of 0.78.

[Table 2 about here.]

It can be seen that the procedure does very well in estimating the true accuracy of Test 1 as measured by the AUC value. The estimates of Se2 are not as accurate; however, it should

be realized that that the disease prevalences here are low and so sensitivities are based on a small proportion of the population. Hence it is not surprising that the standard deviation of the Se2 estimate is higher.

The results reported in Table 2 represent just part of a larger simulation study that was conducted in order to evaluate the performance of the estimation procedure for a variety of situations. That larger study was designed as a $2 \times 3 \times 3$ factorial experiment in which 18 different combinations of parameter values were used — leading to the creation of 1800 data sets. Because of space limitations, the results are reported separately in a Technical Report available from the authors.

For these different 18 scenarios, the procedure performed similarly as in Table 2. However, for true Test 1 AUC values close to 1, the estimation procedure underestimated the AUC. This is natural since 1 is an upper bound. In this case, improved estimates can be obtained by increasing the number K of cutoff points. Alternatively, instead of taking evenly spaced quantiles, an increased proportion of the cutoff values may be taken over the range of the diseased population measurements (assuming the prevalence is less than 0.5). This will lead to more cutoff points in the region where the sensitivity is increasing to 1, which happens rapidly on the ROC curve in this situation when the AUC is close to 1.

It is also of interest to ask what happens with the estimation procedure when the true (unknown) model happens to be non-identifiable. In fact, of the 18 scenarios in our simulation study, 12 had non-identifiable configurations of parameters in which either $\theta_1 = \theta_2$ or $\alpha_2 + \beta_2 = 1$ or both. In this case we would not expect that the procedure to perform well at all, although in fact, for most of the situations considered, the Test 1 AUC values were still estimated quite accurately. Of course, these situations cannot really occur in practice. First, no two populations can have exactly identical prevalences. Second, $\alpha_2 + \beta_2 = 1$ implies that Test 2 is no better than random guessing; it is unlikely that such a test would be employed

in an expensive experiment. However, it is important to know when the true parameters are "close" to a non-identifiable configuration. We propose two rules for flagging potential problem configurations that can be used as warnings.

1. **Flag 1.** The 95% credible interval (CI) for $\theta_1 - \theta_2$ includes 0.

2. **Flag 2.** The sum of upper bounds of the 95% CIs for α_2 and for β_2 is larger than 1.

None of the 100 simulations in Table 2 resulted in any flag, which gives added confidence to the results reported in Section 4.1. In the larger simulation study, 0.5% of the analyses of the 600 datasets for six scenarios with identifiable configurations resulted in a flag. On the contrary, the number of analyses flagged ranged from 92% to 98% for twelve scenarios with non-identifiable configurations. The details are given in the Technical report cited previously.

6 Conclusion

In conclusion, the monotonicity property of the resulting ROC curve estimate is guaranteed by the model structure for the multinomial joint distribution of test-positive and test-negative observations. No assumption is needed concerning the shape of the distributions of test values of the diseased and non-diseased in these populations. The number of populations G may take any value greater than or equal to two.

Our main focus has been to estimate the ROC curve of Test 1; the imperfect reference Test 2 is assumed to have binary outcomes in our model. However, an imperfect reference test with multichromatic or continuous-scaled outcomes may also be analyzed using a generalization of our approach, in which we consider $K'(\geq 2)$ cutoffs for classification rather than simply dichotomizing the Test 2 results. In this case, our model would be based on a $(K'+1)(K+1)$-dimensional multinomial distribution instead of the $2(K+1)$-dimensional multinomial in Section 2.

Our model assumes independence between Test 1 and Test 2 conditional on the true disease status. This assumption is reasonable if the tests have unrelated bases such as can be argued in our application of Section 4. It has been widely used in the literature (Hui and Walter 1980; Joseph et al. 1995; Zhou et al. 2005). This issue has been discussed in detail by Albert and Dodd (2004) and Toft, Jørgensen, and Højsgaard (2005) in the context of binary tests evaluation. It would be an interesting extension to consider ROC curve estimation that allows correlation between tests, still without a gold standard.

Acknowledgements

This study was funded in part by grant USDA-CSREES/2004-35605-14243, JDIP: Johne's Disease Integrated Program in Research, Education, and Extension and in part by grant R01 CA66218 from the National Institutes of Health.

APPENDIX: Full Conditional Distributions for Implementation of the Gibbs Sampler

Based on the augmented data posterior (11), the full conditionals are sampled as follows.

For every $i = 1, .., K; g = 1, .., G$,

$$\begin{aligned} u_{i,g,1}|\bullet &\sim \text{Bin}\left(x_{i,g,1}; \frac{\theta_g b_i(1-\beta_2)}{\theta_g b_i(1-\beta_2) + (1-\theta_g)a_i\alpha_2}\right), \\ u_{i,g,2}|\bullet &\sim \text{Bin}\left(x_{i,g,2}; \frac{\theta_g b_i\beta_2}{\theta_g b_i\beta_2 + (1-\theta_g)a_i(1-\alpha_2)}\right). \end{aligned}$$

All these are Binomial distributions.

$$\pi(\mathbf{a}|\bullet) \propto \prod_{i=1}^{K+1} a_i^{h_i - 1 + \sum_{g=1}^{G}(x_{i,g,1} - u_{i,g,1} + x_{i,g,2} - u_{i,g,2})}.$$

Thus $\mathbf{a}|\bullet$ is simulated from a Dirichlet distribution:

$$Dirich\left(h_1+\sum_{g=1}^{G}(x_{1,g,1}-u_{1,g,1}+x_{1,g,2}-u_{1,g,2})\,,...,\right.$$
$$\left.h_{K+1}+\sum_{g=1}^{G}(x_{K+1,g,1}-u_{K+1,g,1}+x_{K+1,g,2}-u_{K+1,g,2})\right).$$

Similarly

$$\pi(\mathbf{b}|\bullet) \;\propto\; \prod_{i=1}^{K+1} b_i^{h_{K+1+i}-1+\sum_{g=1}^{G}(u_{i,g,1}+u_{i,g,2})}$$

and $\mathbf{b}|\bullet$ is simulated from a Dirichlet distribution:

$$Dirich\left(h_{K+2}+\sum_{g=1}^{G}(u_{1,g,1}+u_{1,g,2}),...,h_{2K+2}+\sum_{g=1}^{G}(u_{K+1,g,1}+u_{K+1,g,2})\right).$$

Also

$$\pi(\alpha_2|\bullet) \;\propto\; \alpha_2^{h_{2K+3}-1+\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,1}-u_{i,g,1})}\cdot(1-\alpha_2)^{h_{2K+4}-1+\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,2}-u_{i,g,2})}$$
$$\times\, I(\alpha_2+\beta_2<1).$$

Thus $\alpha_2|\bullet$ is simulated from a Beta distribution:

$$\alpha_2|\bullet \;\sim Beta\;\left(h_{2K+3}+\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,1}-u_{i,g,1}),h_{2K+4}+\sum_{g=1}^{G}\sum_{i=1}^{K+1}(x_{i,g,2}-u_{i,g,2})\right)$$

truncated by $I(\alpha_2+\beta_2<1)$. Similarly

$$\pi(\beta_2|\bullet) \;\propto\; \beta_2^{h_{2K+5}-1+\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,2}}\cdot(1-\beta_2)^{h_{2K+6}-1+\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,1}}$$
$$\times\, I(\alpha_2+\beta_2<1).$$

Thus $\beta_2|\bullet$ is simulated from a Beta distribution:

$$\beta_2|\bullet \;\sim Beta\;\left(h_{2K+5}+\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,2},h_{2K+6}+\sum_{g=1}^{G}\sum_{i=1}^{K+1}u_{i,g,1}\right)$$

truncated by $I(\alpha_2+\beta_2<1)$.

For every $g=1,..,G$,

$$\pi(\theta_g|\bullet) \;\propto\; \prod_{g=1}^{G}\theta_g^{h_{2K+5+2g}-1+\sum_{i=1}^{K+1}(u_{i,g,1}+u_{i,g,2})}(1-\theta_g)^{h_{2K+6+2g}-1+\sum_{i=1}^{K+1}(x_{i,g,1}-u_{i,g,1}+x_{i,g,2}-u_{i,g,2})}.$$

Thus

$$\theta_g|\bullet \sim Beta \left(h_{2K+5+2g} + \sum_{i=1}^{K+1}(u_{i,g,1} + u_{i,g,2}), \right.$$
$$\left. h_{2K+6+2g} + \sum_{i=1}^{K+1}(x_{i,g,1} - u_{i,g,1} + x_{i,g,2} - u_{i,g,2}) \right).$$

References

Albert, P. S. and Dodd, L. E. (2004), "A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error Without a Gold Standard," *Biometrics*, 60, 427–435.

Andersen, H. J., Aagaard, K., Skjoth, F., Rattenborg, E., and Enevoldsen, C. (2000), "Integration of Research, Development, Health Promotion, and Milk Quality Assurance in the Danish Dairy Industry," in *Proceedings of the Ninth Symposium of the International Society of Veterinary Epidemiology and Economics, Breckenridge, CO, August 6-11*, eds. Salman, M. D., Morley, P., and Ruch-Gallie, R., pp. 258–260.

Andersen, S. (1997), "Re: Bayesian Estimation of Disease Prevalence and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology*, 145, 290–291.

Begg, C. B. and Metz, C. E. (1990), "Consensus Diagnosis and "Gold Standards"," *Medical Decision Making*, 10, 29–30.

Beiden, S. V., Campbell, G., Meier, K. L., and Wagner, R. F. (2000), "On the Problem of ROC Analysis without Truth: The EM Algorithm and the Information Matrix," in *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE): The Inter-*

national Society for Optical Engineering, Bellingham WA,, eds. Salman, M. D., Morley, P., and Ruch-Gallie, R., vol. 3981, pp. 126–134.

Best, N., Cowles, M., and Vines, S. (1995), *CODA Manual version 0.30*, Cambridge, UK: MRC Biostatistics Unit.

Black, M. A. and Craig, B. A. (2002), "Estimating Disease Prevalence in the Absence of a Gold Standard," *Statistics in Medicine*, 21, 2653–2669.

Branscum, A. J., Gardner, I. A., and Johnson, W. O. (2005), "Estimation of Diagnostic-test Sensitivity and Specificity Through Bayesian Modeling," *Preventive Veterinary Medicine*, 68, 145–163.

Choi, Y., Johnson, W. O., Collins, M. T., and Gardner, I. A. (2006), "Bayesian Inferences for Receiver Operating Characteristic Curves in the Absence of a Gold Standard," *Journal of Agricultural, Biological and Enviromental Statistics*, 11, 210 – 229.

Dendukuri, N. and Joseph, L. (2001), "Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests," *Biometrics*, 57, 158 – 167.

Enøe, C., Georgiadis, M. P., and Johnson, W. O. (2000), "Estimation of Sensitivity and Specificity of Diagnostic Tests and Disease Prevalence When the True Disease State is Unknown." *Preventive Veterinary Medicine*, 45, 61–81.

Garrett, E. S., Eaton, E. E., and Zeger, S. (2002), "Methods for Evaluating the Performance of Diagnostic Tests in the Absence of a Gold Standard: a Latent Class Model Approach," *Statistics in Medicine*, 21, 1289–1307.

Georgiadis, 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.

Greiner, M., Pfeiffer, D., and Smith, R. D. (2000), "Principles and Practical Application of Receiver Operating Characteristic Analysis for Diagnostic Tests," *Preventive Veterinary Medicine*, 45, 23–41.

Gustafson, P. (2005), "The Utility of Prior Information and Stratification for Parameter Estimation With Two Screening Tests but No Gold Standard," *Statistics in Medicine*, 24, 1203–1217.

Hall, P. and Zhou, X.-H. (2003), "Nonparametric Estimation of Component Distributions in a Multivariate Mixture." *Annals of Statistics*, 31, 201–224.

Hanson, T. E., Johnson, W. O., and Gardner, I. A. (2003), "Hierarchical Models for the Estimation of Disease Prevalence and the Sensitivity and Specificity of Dependent Tests in the Absence of a Gold-standard," *Journal of Agricultural, Biological and Enviromental Statistics*, 8, 223–239.

Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990), "Receiver Operator Characteristic (ROC) Analysis Without Truth," *Medical Decision Making*, 10, 24–29.

Hui, S. L. and Walter, S. D. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167–171.

Hui, S. L. and Zhou, X. H. (1998), "Evaluation of Diagnostic Tests Without Gold Standards," *Statistical Methods in Medical Research*, 7, 354–370.

Johnson, W. O., Gastwirth, J. L., and Pearson, L. M. (2001), "Screening Without a "Gold Standard": the Hui-Walter Paradigm Revisited," *American Journal of Epidemiology*, 153, 921–924.

Joseph, L., Gyorkos, T., and Coupal, L. (1995), "Bayesian Estimation of Disease Prevalence

and the Parameters of Diagnostic Tests in the Absence of a Gold Standard," *American Journal of Epidemiology*, 141, 263–272.

Nielsen, S. S., Gronbak, C., Agger, J. F., and Houe, H. (2002), "Maximum-likelihood Estimation of Sensitivity and Specificity of ELISAs and Faecal Culture for Diagnosis of Paratuberculosis," *Preventive Veterinary Medicine*, 53, 191–204.

Pepe, M. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, New York: Oxford University Press.

Qu, Y., Tan, M., and Kutner, M. K. (1996), "Random Effects Models for Evaluating Accuracy of Diagnostic Tests," *Biometrics*, 52, 797–810.

Rideout, B. A., Brown, S., Davis, W. C., Gay, J. M., Giannella, R. A., Hines, M. E., Hueston, W. D., Hutchinson, L. J., and Rouse, T. (2003), *The Diagnosis and Control of Johne's Disease*, Washington DC: National Academy Press.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods, 2nd ed.*, New York: Springer.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995), *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*, Cambridge: MRC Biostatistics Unit.

Stabel, J. (2000), "Transitions in Immune Responses to Mycobacterium Paratuberculosis," *Veterinary Microbiology*, 77, 465–473.

Tanner, M. and Wong, W. (1987), "The Calculation Of Posterior Distributions By Data Augmentation," *Journal of the American Statistical Association*, 82, 528–550.

The MathWorks, Inc. (2004), *Getting Started with MATLAB, Version 7.*

Toft, N., Jørgensen, E., and Højsgaard, S. (2005), "Diagnosing Diagnostic Tests: Evaluating the Assumptions Underlying the Estimation of Sensitivity and Specificity in the Absence of a Gold Standard," *Preventive Veterinary Medicine*, 68, 19–33.

Walter, S. D. and Irwig, L. M. (1988), "Estimation of Test Error Rates, Disease Prevalence and Relative Risk From Misclassified Data - a Review," *Journal of Clinical Epidemiology*, 41, 923–937.

Yang, I. and Becker, M. P. (1997), "Latent Variable Modeling of Diagnostic Accuracy," *Biometrics*, 53, 948–958.

Zhou, X.-H., Castelluccio, P., and Zhou, C. (2005), "Nonparametric Estimation of ROC Curves in the Absence of a Gold Standard," *Biometrics*, 61, 600–609.

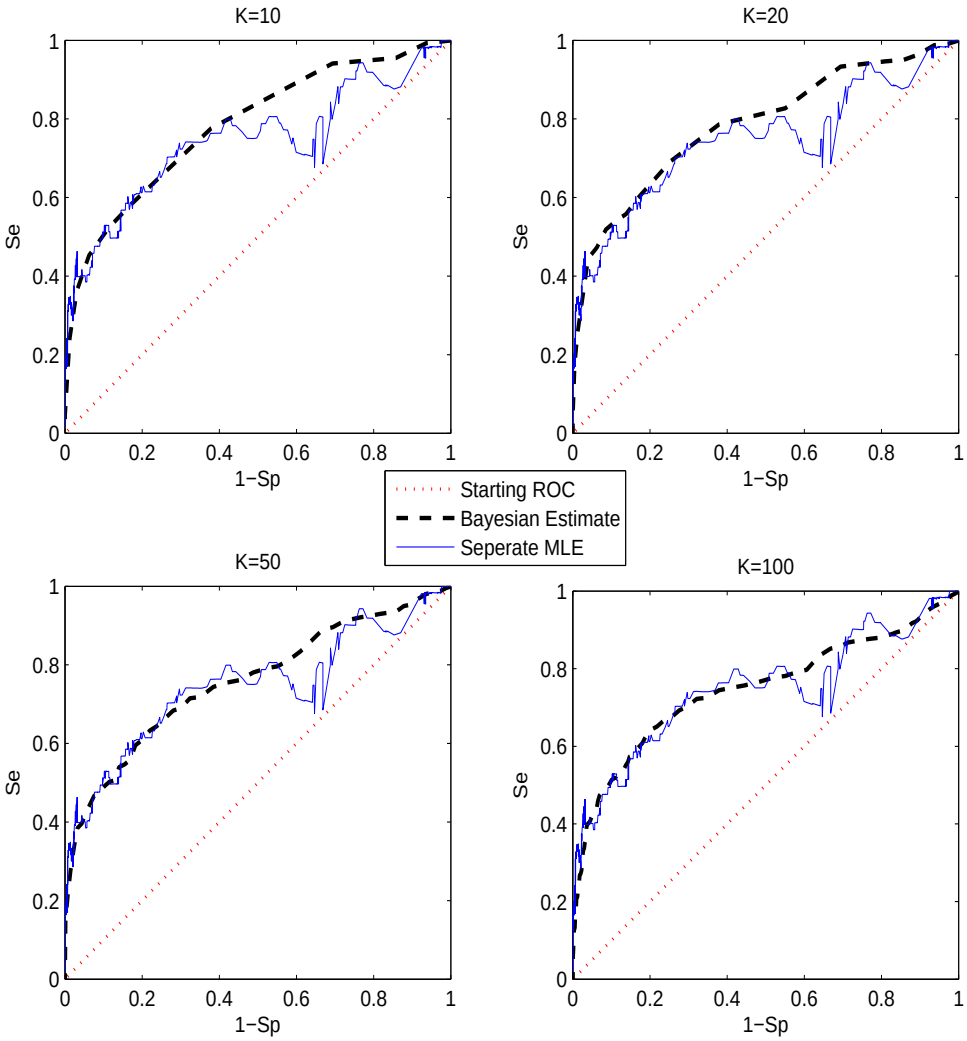

Figure 1: Nonparametric ROC estimation based on Bayesian models compared with the separate MLE method for Johne's disease data. The dotted diagonal lines are the initial starting lines of the MCMC chains. The continuous lines represent the ROC curves estimated by the separate MLE method. The dashed lines represent the ROC curves computed by the nonparametric method proposed here.

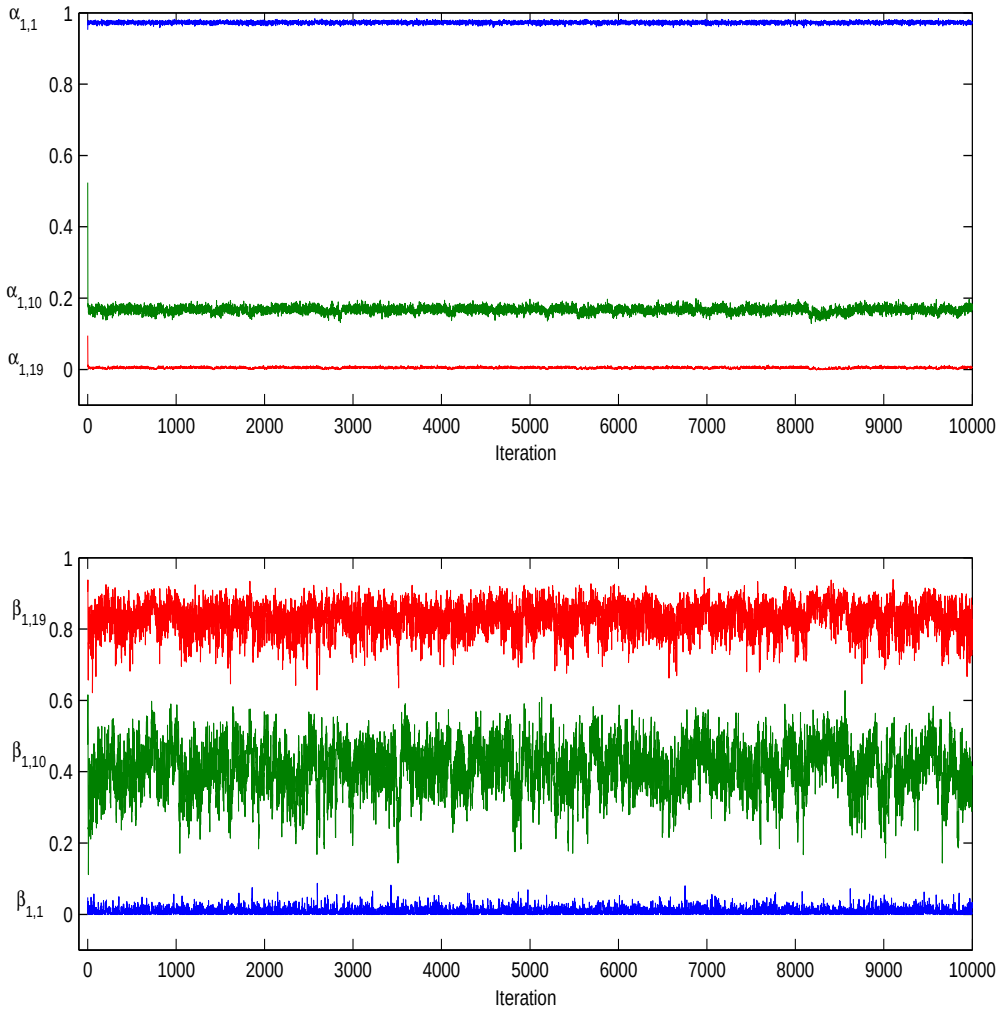

Figure 2: Analysis of Johne's disease data: first 10,000 iterations for error rates $\alpha_{1,1}$, $\alpha_{1,10}$, $\alpha_{1,19}$, $\beta_{1,1}$, $\beta_{1,10}$, $\beta_{1,19}$ in the MCMC simulation. $K = 20$ in the model.

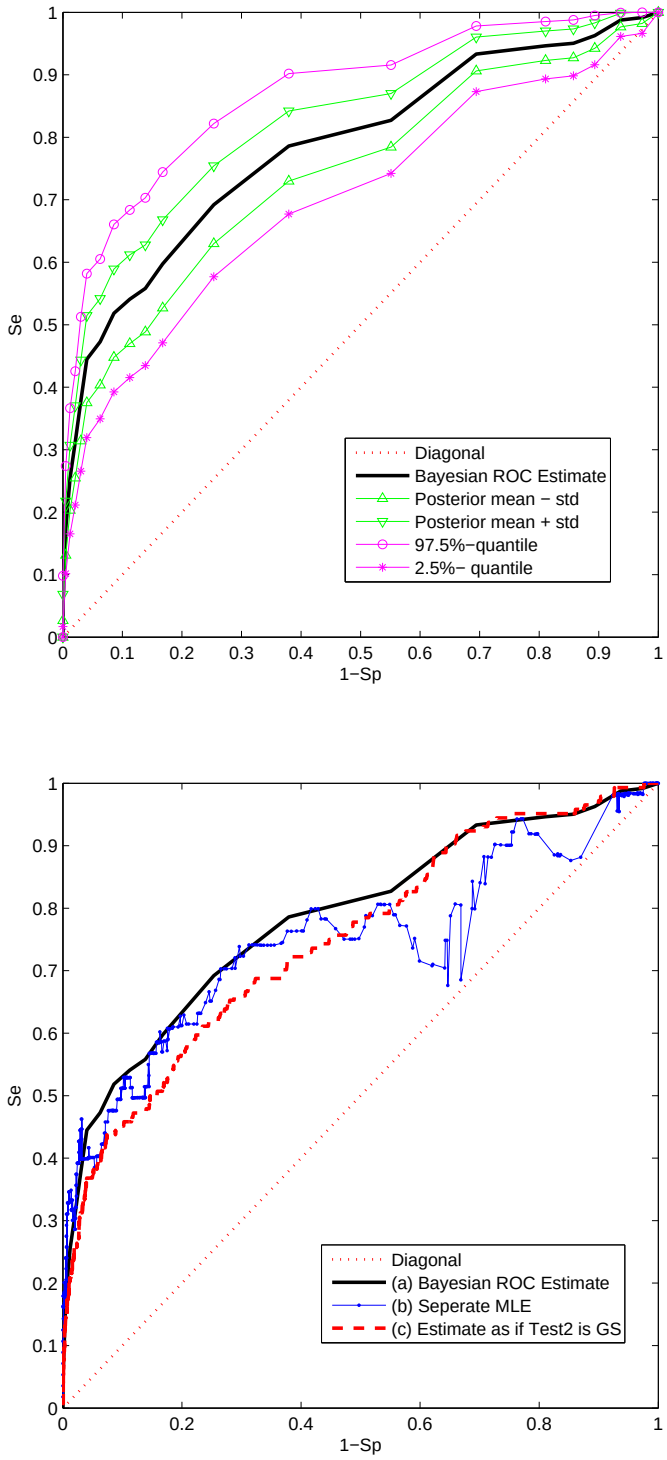

Figure 3: Analysis of Johne's disease data. In upper panel, "$-\triangle-$" and "$-\nabla-$" give the band within one standard deviation of the posterior mean, "$-o-$" and "$-*-$" give the 95% credible interval band. Lower panel compares (a) our Bayesian estimate with (b) separate MLE estimate and (c) estimate as if Test2 is GS. $K = 20$ in the model.

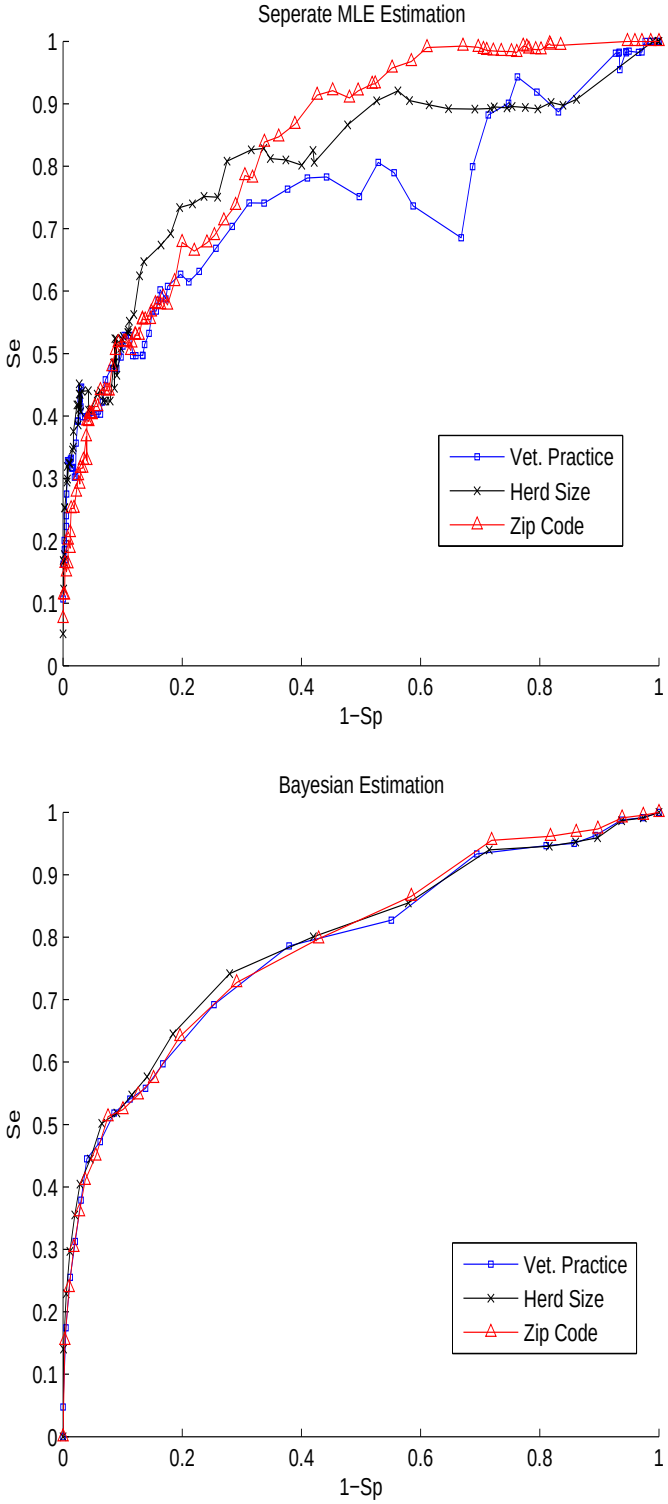

Figure 4: Analysis of Johne's disease data. Robustness of ROC curve estimates to choice of subpopulation division: nonparametric ROC estimates based on Bayesian models (K=20, lower panel), ROC estimates by separate MLE method (upper panel). Subpopulations are created based on divisions by veterinary practitioner, herd size and postal zip code.

Table 1: $2 \times (K+1)$ table of probabilities and frequencies comparing Test 1 and Test 2 in population g

		Test 1						
		$T_{1,1}-$	$T_{1,2}-$, $T_{1,1}+$	... ...	$T_{1,i}-$, $T_{1,i-1}+$	... ...	$T_{1,K}-$, $T_{1,K-1}+$	$T_{1,K}+$
Test 2	T_2+	$p_{1,g}$ $x_{1,g,1}$	$p_{2,g}$ $x_{2,g,1}$	... ...	$p_{i,g}$ $x_{i,g,1}$	... ...	$p_{K,g}$ $x_{K,g,1}$	$p_{K+1,g}$ $x_{K+1,g,1}$
	T_2-	$q_{1,g}$ $x_{1,g,2}$	$q_{2,g}$ $x_{2,g,2}$	... ...	$q_{i,g}$ $x_{i,g,2}$	... ...	$q_{K,g}$ $x_{K,g,2}$	$q_{K+1,g}$ $x_{K+1,g,2}$

See text for explanation of the notation.

Table 2: Results of simulation study for a situation mimicking Johne's disease data set.

	True value	mean(std)	2.5%(std)	97.5%(std)	percentage of 95% CIs that include true value
AUC	0.78	0.764 (0.031)	0.669 (0.035)	0.855 (0.027)	100%
Sp2	0.98	0.977 (0.004)	0.966 (0.004)	0.988 (0.005)	95%
Se2	0.83	0.770 (0.104)	0.507 (0.109)	0.985 (0.038)	99%
Prev1	0.03	0.035 (0.012)	0.016 (0.006)	0.066 (0.023)	98%
Prev2	0.08	0.086 (0.016)	0.054 (0.009)	0.130 (0.026)	100%

AUC is AUC for Test 1. Sp2 and Se2 are the specificity and sensitivity for Test 2. Prev1 and Prev2 are the disease prevalences in populations 1 and 2, respectively.

JABES MS 05075R:

Supplementary Material Not for Publication

Simulation Study: Nonparametric Estimation of ROC Curves Based on Bayesian Models When the True Disease State Is Unknown

Chong Wang[1], Bruce W. Turnbull[1], Yrjö T. Gröhn[2] and Søren S. Nielsen[3]
[1]Department of Statistical Science, Cornell University, Ithaca, NY 14853, U.S.A.
[2]Department of Population Medicine and Diagnostic Sciences,
College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, U.S.A.
[3]Department of Large Animal Science, The Royal Veterinary and
Agricultural University, Frederiksberg, Denmark
E-mail: cw245@cornell.edu, bwt2@cornell.edu, ytg1@cornell.edu, ssn@kvl.dk

April 15, 2006

1. Introduction.

In order to evaluate the performance of the ROC estimation procedure described in the paper [1] "Nonparametric Estimation of ROC Curves Based on Bayesian Models When the True Disease State Is Unknown" by Wang, Turnbull, Gröhn and Nielsen, a moderate size simulation study was undertaken. Here we report the results. The paper [1] describes a procedure for estimating the ROC curve for a diagnostic test (Test 1) with responses measured on a continuous scale. Available also is another diagnostic test (Test 2) which yields results on a binary scale (positive/negative), subject to error. Both tests are applied to individuals from G populations, with samples of size n_g from population g ($1 \leq g \leq G$). The notation used here is as defined in that paper.

2. Creation of a simulated data set.

In this study, each simulated data set contained two populations ($G = 2$), with 1250 individuals in each ($n_1 = n_2 = 1250.$). The sample sizes here were chosen to be similar to those for our Johne's disease application. For the each individual in the g'th population ($g = 1, 2$), we first simulate the true disease status (0 for healthy and 1 for diseased) according to a a binomial distribution $\mathrm{Bin}(1, \theta_g)$, where θ_g is the prevalence of the g'th population. Then, for each diseased individual in either population, a Test 1 score is generated from a normal distribution $\mathrm{N}(\mu_1, \sigma_1^2)$ and a Test 2 score of 0 (negative) or 1 (positive) is generated with probabilities β_2, and $1 - \beta_2$, respectively. Similarly, for each healthy individual, a Test 1 score is generated from a $\mathrm{N}(\mu_0, \sigma_0^2)$ distribution and a Test 2 score of 0 or 1 is generated with probabilities $1 - \alpha_2$, and α_2, respectively. Recall that α_2 and β_2 are the (unknown) false positive rate (1-specificity) and false negative rate (1-sensitivity) for Test 2.

3. Factorial structure of the simulation experiment.

The performance of the estimation procedure was evaluated under 18 different scenarios. These included 12 scenarios where the true parameter configurations implied non-identifiability. The purpose of including these cases was to see the effect on the estimates produced and to see how well these situations might be detected in practice. The effect of varying three factors was studied:

Factor A: Prevalences $\{\theta_1, \theta_2\}$.

Two choices of pairs were used:

A.1. $\{\theta_1 = 0.03, \theta_2 = 0.08\}$. These values are comparable with those found in our Johne's disease application and are of the same order of magnitude reported by other authors.

A.2. $\{\theta_1 = \theta_2 = 0.055\}$. This is a non-identifiable situation because the prevalences are equal. The common value of 0.055 was chosen as the mean of the two prevalences in A1.

Factor B: ROC curves for Test 1.

As noted in Section 2 above, the ROC curve for Test 1 was based on a binormal model. Three such models were considered:

B.I. Distributions from the healthy and diseased individuals are normal with $\mu_0 = -0.33, \sigma_0^2 = 0.12$ and $\mu_1 = 0.33, \sigma_1^2 = 0.63$. These design values lead to an ROC curve similar to the Bayesian estimate in our application (see Figure 3 of [1]). The AUC (area under the curve) is 0.78, approximately the same as that found in the application.

B.II. The distributions are generated using $\mu_0 = \mu_1 = 0; \sigma_0^2 = 0.12, \sigma_1^2 = 0.63$. Here the variances are the same as in B2, but the means are equal. The latter fact implies that the AUC = 0.5 and Test 1 has no diagnostic value at all.

B.III. The distributions are generated using $\mu_0 = -0.56, \mu_1 = 0.56; \sigma_0^2 = 0.12, \sigma_1^2 = 0.63$. This is a scenario where Test 1 is quite accurate with AUC = 0.9.

Factor C: Error rates for Test 2.
Three pairs $\alpha_2,\ \beta_2$ were considered for the false positive and false negative rates for Test 2:

C.*i*. $\alpha_2 = 0.02,\ \beta_2 = 0.17$. These were the values from our Johne's disease data set application, estimated using the method proposed in [1].

C.*ii*. $\alpha_2 = 0,\ \beta_2 = 0$. This is the situation where Test 2 is a "gold standard" (GS) test, i.e. it predicts perfectly.

C.*iii*. $\alpha_2 = 0.5,\ \beta_2 = 0.5$. In this case Test 2 has no diagnostic value at all, i.e. it is equivalent to random guessing. This is a non-identifiable situation because Test 2 provides no information at all.

Thus, considering all combinations of factor levels, we have $2 \times 3 \times 3 = 18$ scenarios. For each scenario, one hundred data sets were simulated, producing a total of 1800 data sets, on each of which the method in [1] was applied. In each case, the number of cutoff values used in the estimation procedure was taken to be $K = 20$.

4. Choosing K cutoff values.

In the Johne's disease data set application in [1], we chose the set of $\{1/(K+1), 2/(K+1), ..., K/(K+1)\}$-quantiles of the observed Test 1 scores to be used as cutoff values in the model. We call this Method CO 1. Such a choice insures balanced frequencies in the columns of Table 1 of [1]. However, for cases when there is a large difference between the number of healthy and diseased individuals and the test scores of the two groups (healthy and diseased) are well separated, Method CO 1 will lead to insufficient cutoff values in the range of test scores of the smaller group. When the disease prevalences are very low and the AUC is close to one, lack of cutoff values in this region, where the sensitivity is increasing rapidly to 1 on the ROC curve, will lead to underestimates of the AUC. And this was borne out by the simulations results presented in the next section.

We can use an alternative method to construct cutoff values — Method CO 2, say, — in order to help to resolve the problem. Denote the number of Test 2 positive and Test 2 negative individuals are $m = \sum_g m_g$ and $n = \sum_g n_g$, respectively. For each individual, we generate duplicate observations all with the same Test 1 score as this individual — n duplicates if it is Test 2 positive, and m duplicates if it is Test 2 negative. By doing this, we will have a sample of

$2 \cdot m \cdot n$ scores, now with equal numbers of scores in the two groups. Then we chose the set of $\{1/(K+1), 2/(K+1), ..., K/(K+1)\}$-quantiles of these Test 1 scores in the new augmented sample to be used as the cutoff values in the model. Compared to Method CO 1, an increased proportion of the cutoff values may be taken over the range of the smaller group. If the disease prevalences are very low, Method CO 2 will lead to more cutoff points in the region where the sensitivity is increasing to 1, which happens rapidly on the ROC curve in this situation when the AUC is close to 1.

5. Results.

Table 1 and 2 list the simulation results for the 18 situations, with 100 simulated samples for each situation, for the two methods to chose cutoff values respectively. In particular, the listed results include, as well as the true value, the means and standard deviations of the estimates and 95% credible intervals, for the quantities of interest, namely the AUC of Test 1, sensitivity $(1-\beta_2)$ and specificity $(1-\alpha_2)$ of Test 2 and the prevalences $\{\theta_1, \theta_2\}$ of the two populations. Also provided are the number of samples (out of 100) that "flag" a potential non-identifiable situation ($\{\theta_1 = \theta_2\}$ or $\alpha_2 + \beta_2 = 1$) according to the two criteria proposed in Section 5 of [1].

6. Discussion.

In Table 1, it can be seen that in Scenario 1 (1,I,i), which mimics the situation for the Johne's disease data set application, the estimation procedure provides quite satisfactory results with point and interval estimates close to the true values. The same is true for the other five identifiable scenarios — {(1,II,i),(1,III,i),(1,I,ii),(1,II,ii),(1,III,ii)}. At most 2 of the 100 simulated data sets flagged a non-identifiable situation in each of these cases. However a high percentage of the remaining 12 non-identifiable situations led to warnings that a non-identifiable situation could exist. Of course, except in such artificially created data sets, it is almost impossible to have $\theta_1 = \theta_2$. It is also unlikely in a real application, that any proposed diagnostic procedure, with sufficient promise that it is being tested in substantial experiment, will be not at all better than random guessing ($\alpha_2 + \beta_2 = 1$). It is interesting to note that in those non-identifiable cases 10–15, where Test 2 is better than random guessing, estimates of the value of Test 1, as measured by AUC, remain quite accurate, despite equality of the true prevalences ($\theta_1 = \theta_2 = 0.055$).

For true Test 1 AUC values close to 1, the estimation procedure using Method CO 1 to choose cutoff values underestimated the AUC (Scenario 3 in Table 1). This is improved by using Method CO 2, as shown in Table 2. For other situations, Table 2 suggests Method CO 2 performs similarly to Method CO 1. Since the latter method is simpler, we stay with this as a preferred method in the submitted revised manuscript.

Of course, it is realized that like any simulation study, this one is necessarily limited in scope. It is impossible to cover all possible choices of design parameters. However this study does suggest that the proposed estimation procedure in [1] is quite feasible and practical to carry out in a variety of situations and that it does produce answers that are reasonable and reliable in the cases considered. In any particular application, a similar simulation study could be undertaken with the appropriate approximate design parameters (sample sizes, prevalences, error rates, etc.) if it is important to assess the performance of the estimation procedure in that precise situation.

Table 1: Results of simulation study for 18 Scenarios. Cutoff values are chosen using Method CO 1 as in the submitted manuscript.

S1 (1,I,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.764 (0.031)	0.669 (0.035)	0.855 (0.027)			
Sp2	0.980	0.977 (0.004)	0.966 (0.004)	0.988 (0.005)	0%		
Se2	0.830	0.770 (0.104)	0.507 (0.109)	0.985 (0.038)			0%
Prev1	0.030	0.035 (0.012)	0.016 (0.006)	0.066 (0.023)		0%	
Prev2	0.080	0.086 (0.016)	0.054 (0.009)	0.130 (0.026)			
S2 (1,II,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.498 (0.043)	0.388 (0.044)	0.607 (0.043)			
Sp2	0.980	0.970 (0.005)	0.957 (0.005)	0.982 (0.006)	0%		
Se2	0.830	0.669 (0.157)	0.360 (0.132)	0.957 (0.118)			2%
Prev1	0.030	0.040 (0.043)	0.011 (0.016)	0.095 (0.077)		2%	
Prev2	0.080	0.099 (0.044)	0.050 (0.019)	0.175 (0.075)			
S3 (1,III,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.840 (0.022)	0.762 (0.027)	0.909 (0.017)			
Sp2	0.980	0.983 (0.003)	0.974 (0.004)	0.992 (0.003)	0%		
Se2	0.830	0.777 (0.094)	0.592 (0.082)	0.956 (0.072)			0%
Prev1	0.030	0.039 (0.009)	0.022 (0.006)	0.061 (0.012)		0%	
Prev2	0.080	0.092 (0.011)	0.065 (0.010)	0.123 (0.013)			
S4 (1,I,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.792 (0.026)	0.715 (0.030)	0.864 (0.022)			
Sp2	1.000	0.992 (0.003)	0.985 (0.004)	0.999 (0.002)	0%		
Se2	1.000	0.864 (0.083)	0.653 (0.113)	0.989 (0.036)			0%
Prev1	0.030	0.030 (0.007)	0.017 (0.005)	0.049 (0.012)		0%	
Prev2	0.080	0.085 (0.013)	0.062 (0.009)	0.115 (0.018)			
S5 (1,II,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.504 (0.036)	0.411 (0.036)	0.598 (0.038)			
Sp2	1.000	0.987 (0.004)	0.976 (0.005)	0.996 (0.004)	0%		
Se2	1.000	0.754 (0.147)	0.474 (0.139)	0.967 (0.105)			1%
Prev1	0.030	0.032 (0.022)	0.013 (0.011)	0.062 (0.037)		1%	
Prev2	0.080	0.096 (0.028)	0.059 (0.016)	0.152 (0.044)			
S6 (1,III,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.885 (0.014)	0.829 (0.018)	0.932 (0.011)			
Sp2	1.000	0.998 (0.001)	0.993 (0.002)	1.000 (0.000)	0%		
Se2	1.000	0.921 (0.052)	0.778 (0.075)	0.995 (0.022)			0%
Prev1	0.030	0.031 (0.006)	0.021 (0.005)	0.045 (0.007)		0%	
Prev2	0.080	0.084 (0.008)	0.066 (0.007)	0.104 (0.009)			

Table 1 - continued from previous page

S7 (1,I,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.515 (0.067)	0.381 (0.074)	0.649 (0.080)			
Sp2	0.500	0.517 (0.012)	0.490 (0.010)	0.548 (0.019)	81%		
Se2	0.500	0.597 (0.094)	0.502 (0.019)	0.730 (0.180)			94%
Prev1	0.030	0.250 (0.157)	0.125 (0.119)	0.394 (0.190)		60%	
Prev2	0.080	0.262 (0.164)	0.137 (0.120)	0.403 (0.199)			
S8 (1,II,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.497 (0.052)	0.354 (0.063)	0.637 (0.071)			
Sp2	0.500	0.517 (0.013)	0.490 (0.011)	0.549 (0.020)	76%		
Se2	0.500	0.616 (0.106)	0.504 (0.025)	0.767 (0.190)			97%
Prev1	0.030	0.238 (0.168)	0.117 (0.134)	0.374 (0.201)		79%	
Prev2	0.080	0.244 (0.164)	0.121 (0.128)	0.383 (0.198)			
S9 (1,III,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.545 (0.086)	0.424 (0.090)	0.667 (0.098)			
Sp2	0.500	0.516 (0.014)	0.490 (0.012)	0.547 (0.019)	87%		
Se2	0.500	0.568 (0.066)	0.497 (0.014)	0.682 (0.152)			92%
Prev1	0.030	0.259 (0.178)	0.135 (0.146)	0.402 (0.210)		34%	
Prev2	0.080	0.301 (0.152)	0.170 (0.133)	0.447 (0.174)			
S10 (2,I,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.772 (0.030)	0.674 (0.034)	0.864 (0.026)			
Sp2	0.980	0.976 (0.005)	0.965 (0.005)	0.987 (0.005)	0%		
Se2	0.830	0.615 (0.128)	0.395 (0.070)	0.883 (0.144)			96%
Prev1	0.055	0.078 (0.016)	0.043 (0.012)	0.122 (0.020)		96%	
Prev2	0.055	0.074 (0.018)	0.040 (0.013)	0.118 (0.023)			
S11 (2,II,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.499 (0.044)	0.392 (0.047)	0.607 (0.045)			
Sp2	0.980	0.971 (0.005)	0.958 (0.006)	0.984 (0.006)	0%		
Se2	0.830	0.294 (0.126)	0.171 (0.048)	0.540 (0.268)			96%
Prev1	0.055	0.175 (0.066)	0.086 (0.052)	0.276 (0.087)		96%	
Prev2	0.055	0.175 (0.064)	0.087 (0.054)	0.276 (0.085)			
S12 (2,III,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.839 (0.021)	0.759 (0.025)	0.910 (0.016)			
Sp2	0.980	0.984 (0.004)	0.974 (0.004)	0.993 (0.004)	0%		
Se2	0.830	0.765 (0.080)	0.574 (0.068)	0.973 (0.051)			98%
Prev1	0.055	0.067 (0.009)	0.043 (0.007)	0.094 (0.012)		98%	
Prev2	0.055	0.066 (0.009)	0.042 (0.006)	0.093 (0.011)			

Table 1 - continued from previous page

S13 (2,I,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.802 (0.029)	0.720 (0.032)	0.878 (0.025)			
Sp2	1.000	0.990 (0.003)	0.982 (0.004)	0.998 (0.002)	0%		
Se2	1.000	0.698 (0.083)	0.476 (0.071)	0.977 (0.053)			93%
Prev1	0.055	0.070 (0.011)	0.040 (0.008)	0.105 (0.014)		93%	
Prev2	0.055	0.071 (0.012)	0.041 (0.008)	0.107 (0.015)			
S14 (2,II,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.508 (0.041)	0.406 (0.041)	0.608 (0.044)			
Sp2	1.000	0.982 (0.004)	0.971 (0.004)	0.992 (0.004)	0%		
Se2	1.000	0.367 (0.158)	0.213 (0.063)	0.594 (0.261)			94%
Prev1	0.055	0.131 (0.038)	0.072 (0.035)	0.202 (0.045)		94%	
Prev2	0.055	0.131 (0.038)	0.072 (0.033)	0.203 (0.046)			
S15 (2,III,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.887 (0.014)	0.830 (0.018)	0.936 (0.010)			
Sp2	1.000	0.997 (0.001)	0.992 (0.002)	1.000 (0.000)	0%		
Se2	1.000	0.856 (0.051)	0.703 (0.062)	0.994 (0.014)			95%
Prev1	0.055	0.063 (0.007)	0.045 (0.006)	0.083 (0.008)		95%	
Prev2	0.055	0.063 (0.007)	0.046 (0.006)	0.083 (0.008)			
S16 (2,I,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.730	0.503 (0.054)	0.364 (0.075)	0.642 (0.057)			
Sp2	0.500	0.520 (0.015)	0.493 (0.010)	0.553 (0.023)	66%		
Se2	0.500	0.619 (0.109)	0.508 (0.030)	0.771 (0.187)			93%
Prev1	0.055	0.255 (0.177)	0.131 (0.148)	0.394 (0.209)		82%	
Prev2	0.055	0.251 (0.176)	0.124 (0.140)	0.394 (0.210)			
S17 (2,II,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.499 (0.049)	0.356 (0.066)	0.641 (0.064)			
Sp2	0.500	0.516 (0.014)	0.489 (0.011)	0.546 (0.019)	76%		
Se2	0.500	0.618 (0.101)	0.504 (0.018)	0.780 (0.190)			95%
Prev1	0.055	0.245 (0.171)	0.116 (0.131)	0.390 (0.198)		88%	
Prev2	0.055	0.247 (0.178)	0.121 (0.143)	0.390 (0.201)			
S18 (2,III,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.508 (0.046)	0.373 (0.061)	0.643 (0.064)			
Sp2	0.500	0.519 (0.016)	0.491 (0.013)	0.553 (0.024)	78%		
Se2	0.500	0.608 (0.099)	0.502 (0.017)	0.757 (0.188)			97%
Prev1	0.055	0.264 (0.172)	0.134 (0.138)	0.407 (0.201)		90%	
Prev2	0.055	0.267 (0.181)	0.142 (0.146)	0.413 (0.214)			

Table 2: Results of simulation study for 18 Scenarios. Cutoff values are chosen by using alternate Method CO 2.

S1 (1,I,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.788 (0.031)	0.690 (0.034)	0.879 (0.025)			
Sp2	0.980	0.974 (0.004)	0.964 (0.004)	0.985 (0.005)	0%		
Se2	0.830	0.878 (0.047)	0.714 (0.068)	0.994 (0.012)			0%
Prev1	0.030	0.024 (0.006)	0.013 (0.004)	0.039 (0.008)		0%	
Prev2	0.080	0.070 (0.010)	0.049 (0.008)	0.095 (0.013)			
S2 (1,II,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.500 (0.046)	0.384 (0.047)	0.616 (0.046)			
Sp2	0.980	0.968 (0.005)	0.956 (0.005)	0.979 (0.006)	0%		
Se2	0.830	0.789 (0.099)	0.516 (0.115)	0.989 (0.031)			2%
Prev1	0.030	0.022 (0.008)	0.008 (0.004)	0.046 (0.015)		2%	
Prev2	0.080	0.070 (0.013)	0.042 (0.009)	0.111 (0.023)			
S3 (1,III,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.868 (0.022)	0.789 (0.027)	0.934 (0.015)			
Sp2	0.980	0.980 (0.003)	0.971 (0.004)	0.989 (0.003)	0%		
Se2	0.830	0.887 (0.052)	0.762 (0.066)	0.988 (0.023)			0%
Prev1	0.030	0.029 (0.005)	0.018 (0.004)	0.043 (0.006)		0%	
Prev2	0.080	0.076 (0.009)	0.057 (0.007)	0.098 (0.010)			
S4 (1,I,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.815 (0.028)	0.734 (0.032)	0.890 (0.023)			
Sp2	1.000	0.991 (0.003)	0.982 (0.004)	0.998 (0.002)	0%		
Se2	1.000	0.978 (0.011)	0.904 (0.035)	1.000 (0.000)			0%
Prev1	0.030	0.023 (0.005)	0.014 (0.004)	0.034 (0.006)		0%	
Prev2	0.080	0.072 (0.008)	0.055 (0.007)	0.090 (0.009)			
S5 (1,II,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.507 (0.038)	0.406 (0.038)	0.608 (0.039)			
Sp2	1.000	0.984 (0.004)	0.974 (0.004)	0.994 (0.004)	0%		
Se2	1.000	0.909 (0.051)	0.719 (0.097)	0.999 (0.005)			0%
Prev1	0.030	0.020 (0.005)	0.010 (0.003)	0.034 (0.007)		0%	
Prev2	0.080	0.070 (0.010)	0.049 (0.008)	0.098 (0.015)			
S6 (1,III,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.907 (0.014)	0.852 (0.018)	0.953 (0.011)			
Sp2	1.000	0.996 (0.002)	0.991 (0.002)	1.000 (0.001)	0%		
Se2	1.000	0.990 (0.003)	0.950 (0.011)	1.000 (0.000)			0%
Prev1	0.030	0.028 (0.005)	0.019 (0.004)	0.038 (0.006)		0%	
Prev2	0.080	0.077 (0.007)	0.062 (0.006)	0.093 (0.008)			

Table 2 - continued from previous page

S7 (1,I,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.509 (0.066)	0.375 (0.070)	0.643 (0.086)			
Sp2	0.500	0.517 (0.013)	0.490 (0.010)	0.548 (0.019)	79%		
Se2	0.500	0.604 (0.104)	0.501 (0.018)	0.741 (0.186)			92%
Prev1	0.030	0.260 (0.185)	0.133 (0.149)	0.398 (0.215)		56%	
Prev2	0.080	0.275 (0.176)	0.145 (0.146)	0.416 (0.202)			
S8 (1,II,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.498 (0.054)	0.361 (0.068)	0.637 (0.064)			
Sp2	0.500	0.518 (0.014)	0.490 (0.011)	0.550 (0.019)	75%		
Se2	0.500	0.605 (0.098)	0.501 (0.014)	0.749 (0.184)			98%
Prev1	0.030	0.257 (0.159)	0.124 (0.122)	0.409 (0.187)		81%	
Prev2	0.080	0.269 (0.169)	0.129 (0.130)	0.420 (0.198)			
S9 (1,III,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.547 (0.078)	0.418 (0.088)	0.678 (0.083)			
Sp2	0.500	0.516 (0.013)	0.490 (0.011)	0.546 (0.019)	81%		
Se2	0.500	0.585 (0.080)	0.501 (0.021)	0.715 (0.165)			92%
Prev1	0.030	0.222 (0.157)	0.107 (0.130)	0.360 (0.174)		44%	
Prev2	0.080	0.271 (0.153)	0.151 (0.130)	0.408 (0.166)			
S10 (2,I,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.795 (0.029)	0.693 (0.033)	0.888 (0.026)			
Sp2	0.980	0.973 (0.004)	0.962 (0.004)	0.984 (0.004)	0%		
Se2	0.830	0.867 (0.059)	0.682 (0.085)	0.995 (0.012)			96%
Prev1	0.055	0.047 (0.007)	0.030 (0.006)	0.068 (0.009)		96%	
Prev2	0.055	0.044 (0.008)	0.028 (0.006)	0.065 (0.011)			
S11 (2,II,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.506 (0.050)	0.385 (0.050)	0.627 (0.050)			
Sp2	0.980	0.966 (0.005)	0.954 (0.006)	0.978 (0.005)	0%		
Se2	0.830	0.673 (0.163)	0.427 (0.151)	0.945 (0.128)			96%
Prev1	0.055	0.058 (0.030)	0.028 (0.010)	0.102 (0.057)		96%	
Prev2	0.055	0.057 (0.025)	0.027 (0.009)	0.101 (0.052)			
S12 (2,III,i)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.863 (0.019)	0.780 (0.026)	0.934 (0.012)			
Sp2	0.980	0.981 (0.004)	0.971 (0.004)	0.990 (0.004)	0%		
Se2	0.830	0.913 (0.036)	0.781 (0.057)	0.998 (0.007)			97%
Prev1	0.055	0.051 (0.007)	0.036 (0.006)	0.070 (0.009)		97%	
Prev2	0.055	0.051 (0.007)	0.036 (0.006)	0.069 (0.008)			

Table 2 - continued from previous page

S13 (2,I,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.780	0.826 (0.028)	0.740 (0.032)	0.903 (0.024)			
Sp2	1.000	0.987 (0.003)	0.979 (0.004)	0.996 (0.003)	0%		
Se2	1.000	0.977 (0.009)	0.897 (0.032)	1.000 (0.000)			96%
Prev1	0.055	0.044 (0.007)	0.032 (0.006)	0.059 (0.008)		96%	
Prev2	0.055	0.045 (0.006)	0.032 (0.005)	0.060 (0.007)			
S14 (2,II,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.511 (0.047)	0.398 (0.045)	0.624 (0.049)			
Sp2	1.000	0.978 (0.004)	0.968 (0.004)	0.988 (0.004)	0%		
Se2	1.000	0.875 (0.079)	0.666 (0.125)	0.997 (0.009)			98%
Prev1	0.055	0.040 (0.008)	0.024 (0.006)	0.061 (0.012)		98%	
Prev2	0.055	0.041 (0.008)	0.025 (0.006)	0.062 (0.013)			
S15 (2,III,ii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.911 (0.014)	0.854 (0.018)	0.958 (0.009)			
Sp2	1.000	0.995 (0.002)	0.990 (0.003)	1.000 (0.001)	0%		
Se2	1.000	0.990 (0.003)	0.951 (0.013)	1.000 (0.000)			94%
Prev1	0.055	0.051 (0.006)	0.039 (0.006)	0.065 (0.007)		94%	
Prev2	0.055	0.053 (0.007)	0.040 (0.006)	0.067 (0.007)			
S16 (2,I,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.730	0.504 (0.049)	0.366 (0.070)	0.648 (0.065)			
Sp2	0.500	0.521 (0.016)	0.493 (0.012)	0.554 (0.023)	63%		
Se2	0.500	0.627 (0.112)	0.509 (0.033)	0.778 (0.188)			95%
Prev1	0.055	0.245 (0.182)	0.133 (0.143)	0.378 (0.218)		86%	
Prev2	0.055	0.245 (0.182)	0.132 (0.145)	0.378 (0.216)			
S17 (2,II,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.500	0.498 (0.043)	0.351 (0.060)	0.646 (0.057)			
Sp2	0.500	0.515 (0.013)	0.489 (0.011)	0.545 (0.017)	79%		
Se2	0.500	0.625 (0.108)	0.506 (0.023)	0.783 (0.183)			96%
Prev1	0.055	0.223 (0.156)	0.103 (0.120)	0.357 (0.196)		81%	
Prev2	0.055	0.221 (0.157)	0.103 (0.123)	0.363 (0.198)			
S18 (2,III,iii)	TRUE	mean(std)	2.5%(std)	97.5%(std)	Flag1	Flag2	at least one flag
AUC	0.900	0.503 (0.054)	0.368 (0.066)	0.636 (0.069)			
Sp2	0.500	0.519 (0.014)	0.491 (0.011)	0.551 (0.020)	80%		
Se2	0.500	0.602 (0.102)	0.501 (0.020)	0.738 (0.190)			99%
Prev1	0.055	0.283 (0.173)	0.154 (0.143)	0.424 (0.198)		84%	
Prev2	0.055	0.269 (0.169)	0.141 (0.135)	0.409 (0.201)			

Reference.

1 Wang, C., Turnbull, B.W., Gröhn, Y.T. and Nielsen, S.S. (2006). Nonparametric estimation of ROC curves based on Bayesian models when the true disease state is unknown.