Brazilian Journal of Probability and Statistics
2009, Vol. 23, No. 1, 68–81
DOI: 10.1214/08-BJPS006

Bayesian estimation of performance measures of screening tests in the presence of covariates and absence of a gold standard

Edson Zangiacomi Martinez[a], Francisco Louzada-Neto[b], Jorge Alberto Achcar[a], Kari Juhani Syrjänen[c], Sophie Françoise Mauricette Derchain[d], Renata Clementino Gontijo[d] and Luis Otávio Zanatta Sarian[d]

[a]*Universidade de São Paulo, USP*
[b]*Universidade Federal de São Carlos, UFSCar*
[c]*Turku University Hospital*
[d]*Universidade Estadual de Campinas, UNICAMP*

Abstract. The sensitivity (S_e) and the specificity (S_p) are the two most common measures of the performance of a diagnostic test, where S_e is the probability of a diseased individual to be correctly identified by the test while S_p is the probability of a healthy individual to be correctly identified by the same test. A problem appears when all individuals cannot be verified by a gold standard. This occurs when there is not a definitive test for detection of the disease or the verification by a gold standard is an impracticable procedure according to its cost, accessibility or risks. In this paper we develop a Bayesian analysis to estimate the disease prevalence, the sensitivity and specificity of screening tests in the presence of a covariate and in the absence of a gold standard. We use the Metropolis–Hastings algorithm to obtain the posterior summaries of interest. We have as motivation for the investigation the LAMS (Latin American Screening) Study, an extensive project designed for comparing screening tools for cervical cancer in Brazil and Argentina. When applied to the analysis of data from LAMS Study, the proposed Bayesian method shows to be a useful alternative to estimate measures of performance of screening tests in the presence of covariates and when a gold standard is not available. An advantage of the method is the fact that the number of parameters to be estimated is not limited by the number of observations, as it happens with several frequentist approaches. However, it is important to point out that the Bayesian analysis requires informative priors in order for the parameters to be identifiable. The method can be easily extended for the analysis of other medical data sets.

1 Introduction

The performance of a diagnostic or laboratorial test is usually measured by its sensitivity (S_e) and specificity (S_p). Sensitivity is the probability of a diseased individual to be correctly identified by the test while specificity is the probability

Key words and phrases. Bayesian analysis, diagnostic tests, latent variables, biostatistics.
Received September 2007; accepted March 2008.

of a healthy individual to be correctly identified by the same test. When a reference test is available, sensitivity and specificity can be estimated directly through comparing the test results with the reference test. However, it is common to find situations where a proportion of the sampled individuals cannot be verified on their real disease status. The problem can occur especially when the reference test (usually called "gold standard" in health applications) is an invasive and/or risky procedure and the definitive verification for apparently healthy individuals is thus neither practical nor ethical. In order to overcome this problem, many studies on the evaluation of the diagnostic test are carried out by considering only verified individuals. However, this approach can lead to measures that are usually biased (Begg (1987)). This bias is called verification bias or workup bias (Ransohoff and Feinstein (1978)). In this situation, estimators for S_e and S_p are introduced by Begg and Greenes (1983) and Zhou (1983). Studies from the literature that describe how to estimate performance measures in the presence of the verification bias was reviewed by Zhou (1998).

Another problem happens when all individuals cannot be verified by a gold standard. Most likely, this occurs due to a lack of a perfect reference test for the disease or in situations where the verification by a gold standard is an impracticable procedure according to its cost, accessibility or risks. Studies related to the performance estimation of diagnostic and laboratorial tests in the absence of a gold standard are commons in preventive veterinary medicine, given that the correct classification of the true status of herds is an important component of animal disease-control programs (Christensen and Gardner (2000); Greiner and Gardner (2000)). Enøe, Georgiadis and Johnson (2000) review some of the statistical methods usual in veterinary medicine for estimation of the accuracy of diagnostic tests when a reference test does not exist, with an illustration of the evaluation of a nested polymerase chain reaction and microscopic examination of kidney imprints for detection of Nucleospora salmonis in rainbow trout. The authors discuss the estimation of parameters of interest by maximum likelihood method, using Newton–Raphson and EM algorithm, and present a Bayesian approach based in Gibbs sampling method. However, Enøe, Georgiadis and Johnson (2000) do not discuss the presence of covariates in the reviewed models. Toft, Jørgensen and Højsgaard (2005) discuss the usual assumptions underlying the estimation of sensitivity and specificity in veterinary medicine studies in situations that a gold standard is not available.

A literature review of the frequentist methods developed to estimate the sensitivity and specificity of screening tests without a gold standard was conducted by Hui and Zhou (1998). Hui and Walter (1980) derived equations that compute maximum likelihood estimates and standard errors of sensitivity, specificity and prevalence, without considering a reference test. Under the assumption that the tests are conditionally independents, Joseph, Gyorkos and Coupal (1985) introduced a Bayesian model using latent variables, and Dendukuri and Joseph (2001)

extended this method to account for conditional dependence between the diagnostic tests. A software that yields maximum-likelihood estimates of sensitivity, specificity and disease prevalence in the absence of a gold standard was introduced by Pouillot, Gerbier and Gardner (2002). The program uses two estimation methods, a Newton–Raphson procedure and an expectation–maximization (EM) procedure. Other interesting contributions were provided by Faraone and Tsuang (1994), Qu, Tan and Kutner (1996), Hadgu and Qu (1998), Martinez, Achcar and Louzada-Neto (2005, 2006) and Stamey, Boese and Young (2008).

The objective of the present study is to propose a Bayesian approach for the problem of estimating the sensitivities and specificities of multiple diagnostic tests, considering that part of the sampled individuals was not verified by a gold standard. We also consider the presence of covariates in our statistical model. We have as motivation for the investigation, the LAMS (Latin American Screening) study, where PAP smear/liquid-based cytology and screening colposcopy were compared with three optional screening tools (visual inspection with acetic acid or Lugol's iodine and cervicography) and with Hybrid Capture II (HC II) from conventional samples and from self-samples in women at different risk for cervical cancer in three Brazilian arms (São Paulo, Campinas and Porto Alegre) and one Argentine arm (Buenos Aires). Readers interested in more details about the LAMS study can find them in Syrjänen et al. (2005), Sarian et al. (2005) and Gontijo et al. (2007). Here, we considered the data from Campinas, collected in 2002.

In the next section, we present a description of the method of Joseph, Gyorkos and Coupal (1985) for estimating S_e and S_p related to two diagnostic tests in the absence of a gold standard. We also introduce in this section the methodology for estimating S_e and S_p in the presence of a covariate. The application of the proposed methodology on the analysis of the data set is presented in Section 3. Concluding remarks are given in Section 4.

2 The model

We will consider k diagnostic tests, where $T_m = 1$ denotes a positive result for the test m and $T_m = 0$ denotes a negative result for the same test, for $m = 1, \ldots, k$. Let S_{e_m} and S_{p_m} be the sensitivity and the specificity of the test m, respectively, and let g be an observation of a binary latent variable G, introduced in the model aiming to simulate a nonobservable gold standard (Tanner and Wong (1987)). If we assume that the set of the observations and the latent variable for the ith individual is denoted by $\mathbf{x}_i = \{t_{1_i}, t_{2_i}, \ldots, t_{k_i}, g_i\}$, where t_m is an observation of T_m and the joint density function is given by

$$f(x_i) = p^{g_i}(1-p)^{1-g_i} \prod_{m=1}^{k} S_{e_m}^{t_{m_i} g_i} (1 - S_{e_m})^{(1-t_{m_i})g_i} S_{p_m}^{(1-t_{m_i})(1-g_i)} \times (1 - S_{p_m})^{t_{m_i}(1-g_i)}.$$

We are assuming that the test results are independents. Note that we have $2k+1$ parameter to be estimated, or say, k pairs (S_{e_m}, S_{p_m}), and the prevalence p. The likelihood function $L(\boldsymbol{\theta})$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}) = {} & p^{\sum_{i=1}^{n} g_i}(1-p)^{n-\sum_{i=1}^{n} g_i} \\ & \times \prod_{m=1}^{k} S_{e_m}^{\sum_{i=1}^{n} t_{m_i} g_i}(1-S_{e_m})^{\sum_{i=1}^{n}(1-t_{m_i})g_i} \\ & \times S_{p_m}^{\sum_{i=1}^{n}(1-t_{m_i})(1-g_i)}(1-S_{p_m})^{\sum_{i=1}^{n} t_{m_i}(1-g_i)}, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\theta} = (S_{e_m}, S_{p_m}, p : m = 1, \ldots, k)$. The latent variable G has a Bernoulli distribution with success probability given by an application of Bayes' rule, that is,

$$\begin{aligned} G_i|\boldsymbol{\theta}, t_{1_i}, \ldots, t_{k_i} \sim Bernoulli\Bigg(& p\prod_{m=1}^{k} S_{e_m}^{t_{m_i}}(1-S_{e_m})^{1-t_{m_i}} \\ & \Big/\Bigg(p\prod_{m=1}^{k} S_{e_m}^{t_{m_i}}(1-S_{e_m})^{1-t_{m_i}} \\ & + (1-p)\prod_{m=1}^{k} S_{p_m}^{1-t_{m_i}}(1-S_{p_m})^{t_{m_i}}\Bigg)\Bigg). \end{aligned} \tag{2.2}$$

Considering beta prior densities $Beta(\alpha_\theta, \beta_\theta)$ for all parameters in $\boldsymbol{\theta}$, where α_θ and β_θ generically denotes fixed hyperparameters and combining the likelihood function for $\boldsymbol{\theta}$ (2.1) with the prior densities, we use the Gibbs sampling algorithm (Gelfand and Smith (1990); Gelfand (2000)) to simulate samples for the posterior distribution for $\boldsymbol{\theta}$. A prior beta distribution is appropriated in this model because all parameters in $\boldsymbol{\theta}$ can be interpreted as proportions (varies from 0 to 1) and due to its flexibility. The posterior samples are simulated from the full conditional posterior distributions for p, S_{e_m} and S_{p_m}, $m = 1, \ldots, k$. Following equations (2.1) and (2.2) and considering k diagnostic tests, the conditional posterior distributions for the components of $\boldsymbol{\theta}$ needed for the Gibbs sampling algorithm are given by

$$p|\mathbf{X}, \alpha_p, \beta_p \sim Beta\Bigg(\sum_{i=1}^{n} g_i + \alpha_P; n - \sum_{i=1}^{n} g_i + \beta_P\Bigg),$$

$$S_{e_m}|\mathbf{X}, \alpha_{S_{e_m}}, \beta_{S_{e_m}} \sim Beta\Bigg(\sum_{i=1}^{n} t_{m_i} g_i + \alpha_{S_{e_m}}; \sum_{i=1}^{n}(1-t_{m_i})g_i + \beta_{S_{e_m}}\Bigg) \quad \text{and}$$

$$S_{p_m}|\mathbf{X}, \alpha_{S_{p_m}}, \beta_{S_{p_m}} \sim Beta\Bigg(\sum_{i=1}^{n}(1-t_{m_i})(1-g_i) + \alpha_{S_{p_m}}; \sum_{i=1}^{n} t_{m_i}(1-g_i) + \beta_{S_{p_m}}\Bigg),$$

for $m = 1, \ldots, k$. This model is analogous to the model proposed by Joseph, Gyorkos and Coupal (1985). However, the method of Joseph, Gyorkos and Coupal

(1985) does not consider the presence of covariates, which are very common on data from diagnostic test studies.

Let $\mathbf{w}_i$ be a sample observation of $\mathbf{W}_i$, a vector of L covariates. For the sake of simplicity and without lack of generality, we assume that T_i is a random variable (with observation t_i) related to the result of only one diagnostic test, with Bernoulli distribution with success probability $p_i S_{e_i} + (1-p_i)(1-S_{p_i})$, $i = 1, \ldots, n$. In the presence of a vector of covariates, let us assume the logit links for S_{e_i}, S_{p_i} and p_i, given by $\theta_{li} = \exp(\sum_{j=0}^{L} \beta_{lj} w_{ji})[1 + \exp(\sum_{j=0}^{L} \beta_{lj} w_{ji})]^{-1}$, where $l = 1, 2, 3$, $W_{0i} = 1$, $\theta_{1i} = S_{e_i}$, $\theta_{2i} = S_{p_i}$, $\theta_{3i} = p_i$, for $i = 1, \ldots, n$. In this way, we have a vector of parameters given by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, where $\beta_l = (\beta_{l0}, \beta_{l1}, \ldots, \beta_{lL})$, $l = 1, 2, 3$. We are using a logit link function to relate the covariates linearly to the screening performance measures, but it is possible to use other link functions in place of the logit function, according to the nature of the data. Assuming prior independence among the parameters, we consider as prior densities for β_{lj}, normal distributions with fixed hyperparameters a_{lj} (means) and b_{lj}^2 (variances), $l = 1, 2, 3$, $j = 0, 1, \ldots, L$. The likelihood function for $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
L(\boldsymbol{\beta}) = \exp\Bigg[& \sum_{j=0}^{L} \beta_{1j} \sum_{i=1}^{n} w_{ji} t_i g_i \\
& + \sum_{j=0}^{L} \beta_{2j} \sum_{i=1}^{n} w_{ji} (1-t_i)(1-g_i) + \sum_{j=0}^{L} \beta_{3j} \sum_{i=1}^{n} w_{ji} g_i \Bigg] \\
& \Bigg/ \prod_{i=1}^{n} \Bigg\{ \Bigg[1 + \exp\Bigg(\sum_{j=0}^{L} \beta_{1j} w_{ji} \Bigg) \Bigg]^{g_i} \Bigg[1 + \exp\Bigg(\sum_{j=0}^{L} \beta_{2j} w_{ji} \Bigg) \Bigg]^{1-g_i} \\
& \times \Bigg[1 + \exp\Bigg(\sum_{j=0}^{L} \beta_{3j} w_{ji} \Bigg) \Bigg] \Bigg\},
\end{aligned}
$$

where g is an observation of the latent variable G, given by (2.2). Combining the prior distributions with $L(\boldsymbol{\beta})$, we have the conditional posterior distributions for $\boldsymbol{\beta}$ given by

$$
\begin{aligned}
& \pi(\beta_{1j} | \boldsymbol{\beta}_{(\beta_{1j})}, \mathbf{X}, \mathbf{W}) \\
& \quad \propto N(a_{1j}; b_{1j}^2) \times \exp\Bigg\{ \beta_{1j} \sum_{i=1}^{n} w_{ji} t_i g_i - \sum_{i=1}^{n} g_i \ln\Bigg[1 + \exp\Bigg(\sum_{k=0}^{L} \beta_{1k} w_{ki} \Bigg) \Bigg] \Bigg\}, \\
& \pi(\beta_{2j} | \boldsymbol{\beta}_{(\beta_{2j})}, \mathbf{X}, \mathbf{W}) \\
& \quad \propto N(a_{2j}; b_{2j}^2) \times \exp\Bigg\{ \beta_{2j} \sum_{i=1}^{n} w_{ji} (1-t_i)(1-g_i)
\end{aligned}
$$

$$- \sum_{i=1}^{n} (1 - g_i) \ln \left[1 + \exp \left(\sum_{k=0}^{L} \beta_{2k} w_{ki} \right) \right] \Bigg\} \quad \text{and}$$

$$\pi \left(\beta_{3j} | \boldsymbol{\beta}_{(\beta_{3j})}, \mathbf{X}, \mathbf{W} \right)$$

$$\propto N(a_{3j}; b_{3j}^2) \times \exp \left\{ \beta_{3j} \sum_{i=1}^{n} w_{ji} g_i - \sum_{i=1}^{n} \ln \left[1 + \exp \left(\sum_{k=0}^{L} \beta_{3k} w_{ki} \right) \right] \right\},$$

where $j = 0, 1, \ldots, L$, and $\boldsymbol{\beta}_{(\beta_{10})}$ is the vector of all parameters except β_{10} (for example). Observe that we should simulate samples for all parameters considering the Metropolis–Hastings algorithm since their conditional distributions are difficult to sample. In each cycle of the algorithm is generated a new value for the latent variable G as (2.2).

In studies of the performance of two or more independent diagnostic tests applied to a selected group of individuals, where none of these tests can be considered the gold standard, a straightforward extension of this model can be used. Considering the three diagnostic tests, the vector of unknown parameters is now given by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_7)$, where $\boldsymbol{\beta}_l = (\beta_{l0}, \beta_{l1}, \ldots, \beta_{lL})$, $l = 1, \ldots, 7$, are vectors of parameters related to the sensitivity and the specificity of each test and the prevalence of cervical lesions. Let T_{m_i} be a random variable with observation t_{m_i} related to test m, $m = 1, 2, 3$. Using a logit link function to relate the vector $\mathbf{W}_i$ of L covariates to the screening performance measures, $i = 1, \ldots, n$, the likelihood function for $\boldsymbol{\beta}$ is now given by

$$\begin{aligned} L(\boldsymbol{\beta}) = \exp \Bigg[& \sum_{l=1}^{3} \sum_{j=0}^{L} \beta_{lj} \sum_{i=1}^{n} w_{ji} t_{l_i} g_i + \sum_{l=4}^{6} \sum_{j=0}^{L} \beta_{lj} \sum_{i=1}^{n} w_{ji} (1 - t_{(l-3)_i})(1 - g_i) \\ & + \sum_{j=0}^{L} \beta_{7j} \sum_{i=1}^{n} w_{ji} g_i \Bigg] \\ \Big/ & \prod_{i=1}^{n} \left\{ \prod_{l=1}^{3} \left[1 + \exp \left(\sum_{j=0}^{L} \beta_{lj} w_{ji} \right) \right]^{g_i} \right. \\ & \left. \times \prod_{l=4}^{6} \left[1 + \exp \left(\sum_{j=0}^{L} \beta_{lj} w_{ji} \right) \right]^{1-g_i} \left[1 + \exp \left(\sum_{j=0}^{L} \beta_{7j} w_{ji} \right) \right] \right\}. \end{aligned} \tag{2.3}$$

3 Application to data from the LAMS study

LAMS is an ongoing, cross-sectional, multi-centre study sponsored by the European Commission (see details in Sarian et al. (2005)). In this study, consecutive women from the cities of Campinas (Brazil), São Paulo (Brazil), Porto Alegre

Table 1 *Results of cervical cytology, visual inspection with acetic acid (VIA) and Second-Generation Hybrid Capture (HC II)*

	Cervical cytology +		Cervical cytology −		
	HC II +	HC II −	HC II +	HC II −	Total
VIA +	9	2	15	35	61
VIA −	21	19	87	621	748
Total	30	21	102	656	809

(Brazil) and Buenos Aires (Argentina) were recruited to undergo gynecological consultations and examination with conventional Pap smear, VIA and VILI (visual inspection with acetic acid or Lugol's iodine), cervicography, screening colposcopy and Second-Generation Hybrid Capture (HC II). In assessing the accuracy of cervical cancer screening tests, it is not straightforward to define an ideal gold standard. In many studies, the gold standard for evaluating the accuracy of screening tests in detecting true positive lesions is the histopathology. If biopsies are not obtained, colposcopy is accepted as the final diagnosis. However, colposcopy can give many false negative results when used to discriminate between normal and abnormal tissues (see, e.g., Mitchell et al. (1998) and Hopman, Kenemans and Helmerhorst (1998)). In the present study, let us consider the data from the LAMS Study collected in Campinas. Thus, we use a sample of 809 women with test results for cervical cytology, VIA and HC II. Table 1 shows the results of the three tests for the 809 available cases.

In a first step, we estimate the sensitivity and specificity of cervical cytology (T_1), VIA (T_2) and HC II (T_3) to detect cervical preneoplasic or neoplasic lesions by the Bayesian approach proposed by Joseph, Gyorkos and Coupal (1985). Seven parameters were estimated, including the prevalence of preneoplasic or neoplasic lesions and the sensitivity and specificity pairs relative to the three diagnostic methods under evaluation.

The prior information about the sensitivity and the specificity of cervical cytology was based on the systematic review of Nanda et al. (2000), who presented sensitivity for atypical squamous cells of undetermined significance (ASC-US) or worse being ranged from 29 percent to 56 percent and specificity from 97 percent to 100 percent. The studies of Belinson et al. (2001) and of the University of Zimbabwe and JHPIEGO Cervical Cancer Project (1999) were used as references for the choice of the prior information about the sensitivity and the specificity for the VIA. In these studies, the sensitivity of VIA for at least CIN II was estimated in 55 percent and 64 percent, respectively, and the specificity was estimated as 76 percent and 67 percent, respectively. The prior information of the accuracy measures of HC II test was based on the studies of Schiffman et al. (2000) and Wright et al. (2000), who estimated sensitivities by 88.4 percent and 81.3 percent (at 1 RLU cutoff), respectively, and specificities by 89.0 percent and 84.5 percent, respectively.

Table 2 *Bayesian estimates for sensitivities and specificities for each screening test, and for the prevalence of cervical lesions (CI: credible interval)*

Test		%	95% CI	
Cervical cytology	sensitivity (S_{e_1})	53.8	42.6	65.1
	specificity (S_{p_1})	97.0	95.4	98.4
Visual inspection with acetic-acid	sensitivity (S_{e_2})	53.0	43.4	62.3
	specificity (S_{p_2})	93.0	91.0	94.7
Hybrid Capture II	sensitivity (S_{e_3})	90.3	76.5	98.6
	specificity (S_{p_3})	88.7	85.9	91.4
Prevalence (p)		6.3	3.9	9.1

However, the choice of informative prior distributions based only in a summary of previous studies can be a complex task, since each study has elements of subjectivity, error-proneness and possible potential for bias. Thus, a panel of experts on cervical cancer was asked to provide their best estimate for the sensitivities and specificities of the tests, and the prior distributions that summarise the information provided by the literature review corrected by the experts were derived. The assessment of beta distribution priors for each test parameter considered the method presented by Joseph, Gyorkos and Coupal (1985), where the hyperparameters are defined by matching the center of a range of plausible values of sensitivity and sensitivity with the mean of the beta distribution and matching the standard deviation of the beta distribution with one quarter of the total range. We considered a vague prior distribution for the prevalence of precursor cervical lesions (a Beta distribution with hyperparameters 0.5 and 0.5; see Box and Tiao (1992)) motivated by a little background knowledge about this parameter.

We generated 220,000 samples using the MCMC procedure, sampling every 20 to assure that successive samples were independent. We removed the first 20,000 samples in the chain to avoid including any generated values that might have been sampled before convergence of the Markov chain. For each parameter of interest, the arithmetic mean of these Gibbs samples is a natural Bayesian estimator. These arithmetic means are showed in Table 2, with the respective 95 percent credible intervals.

The results suggest a low sensitivity for cervical cytology to detect ASC-US or worse (53.8 percent) as well as for VIA (53 percent), but indicate a high sensitivity for the HC II (90.3 percent). All screening methods presented relatively high specificities: 97.0 percent for cervical cytology, 93.0 percent for VIA and 88.7 percent for the HC II.

In a second instance, we introduced in the model the age of the women (X) as a continuous covariate. Under the notation introduced in Section 2, the covariate W_1 is given by $(X - \overline{x})/10$, where $\overline{x}$ is the sample mean of X. The quotient 10 is only considered for avoiding numerical instability related to large values in exponential functions present in the conditional posterior densities of interest.

We also introduced in the model the variable W_2, a dichotomous variable that denotes whether or not the woman is actually pregnant (1 if pregnant and zero otherwise). Firstly, we considered the interaction between W_1 and W_2 in the model. However, all interaction parameters were estimated to be close to zero (ranged from -0.031 to 0.009) and were excluded from the final model. In the expression (2.3), the vectors of parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are related to the sensitivities of the cervical cytology, VIA and HC II, respectively; $\boldsymbol{\beta}_4$, $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ are related to the specificities of the cervical cytology, VIA and HC II, respectively; and the vector $\boldsymbol{\beta}_7$ is related to the prevalence of cervical lesions. We consider the prior densities for β_{lj} with normal distribution with fixed hyperparameters a_{lj} (means) and b_{lj}^2 (variances), $l = 1, \ldots, 7,\ j = 0, 1, \ldots, L$. Combining the prior distributions with $L(\boldsymbol{\beta})$, we obtain the conditional posterior distributions for $\boldsymbol{\beta}$ and the Metropolis–Hastings algorithm is used to generate samples from the each parameter.

From the conditional densities for the parameters in $\boldsymbol{\beta}$ given in Section 2, we generated 220,000 Gibbs samples. From this chain, we discarded the first 20,000 (regarded as burn-in samples). The convergence of the Gibbs samples was monitored by standard existing methods (Geweke (1992)). The convergence was observed for all parameters. Prior distributions for the intercept parameters β_{10} to β_{70} were assumed with fixed hyperparameters based in estimates obtained in the previous analysis without covariates. For example, we noted that the estimated sensitivity of the cervical cytology was 0.536 (see Table 2), and considering the inverse of the logit function, the hyperparameter a_{10} is thus given by $\log(0.536/(1 - 0.536))$. All the other hyperparameter values were chosen to have noninformative priors. Thus, we used an empirical Bayesian modelling approach (Carlin and Louis (2000)). For each parameter, we considered every 50th draw to assure that successive samples were independent. Considering that a logit link was used, the regression coefficients in $\boldsymbol{\beta}$ are interpreted as being the logarithm of the odds ratios (OR). These odds ratios represent an association measure between the variables W_1 and W_2 and the operating characteristics of the screening tests.

In Table 3, we have the posterior summaries for the exponential function of the parameters of interest in $\boldsymbol{\beta}$, interpreted as odds ratios. We observe that the 95% credible intervals for the parameters $e^{\beta_{11}}$ to $e^{\beta_{71}}$ included the value 1, suggesting that there is no evidence for the effect of pregnancy in S_e and S_p measures for all tests. The parameters $e^{\beta_{22}}$ and $e^{\beta_{52}}$ were estimated in 2.033 and 1.615, respectively, and its credible interval does not include the value 1. This result suggests that the sensitivity and the specificity of VIA increases as the age of the women increases. In fact, in the medical literature several authors have described that methods for detection of precursor lesions of cervical cancer have different performances according to the age of women (see University of Zimbabwe (1999), Koss (2000) and Shlay et al. (2000)). The prevalence of cervical lesions, as expected, tends to

Table 3 *Posterior odds ratios as association measures between pregnancy and age and the performance measures of the cervical cytology, VIA and HC II (SD: standard deviation; CI: credible interval)*

Parameter	Measure	Mean	SD	95% CI	
$e^{\beta_{11}}$	effect of pregnancy on S_{e_1}	1.142	0.566	0.420	2.496
$e^{\beta_{21}}$	effect of pregnancy on S_{e_2}	1.079	0.502	0.400	2.291
$e^{\beta_{31}}$	effect of pregnancy on S_{e_3}	1.214	0.667	0.413	2.864
$e^{\beta_{41}}$	effect of pregnancy on S_{p_1}	1.105	0.528	0.416	2.384
$e^{\beta_{51}}$	effect of pregnancy on S_{p_2}	1.626	0.710	0.670	3.425
$e^{\beta_{61}}$	effect of pregnancy on S_{p_3}	0.840	0.343	0.381	1.692
$e^{\beta_{71}}$	effect of pregnancy on p	1.168	0.509	0.479	2.451
$e^{\beta_{12}}$	effect of age on S_{e_1}	0.973	0.318	0.481	1.694
$e^{\beta_{22}}$	effect of age on S_{e_2}	2.033	0.699	1.013	3.737
$e^{\beta_{32}}$	effect of age on S_{e_3}	0.914	0.364	0.393	1.856
$e^{\beta_{42}}$	effect of age on S_{p_1}	0.804	0.202	0.469	1.239
$e^{\beta_{52}}$	effect of age on S_{p_2}	1.615	0.282	1.131	2.272
$e^{\beta_{62}}$	effect of age on S_{p_3}	1.422	0.238	0.996	1.925
$e^{\beta_{72}}$	effect of age on p	0.483	0.109	0.313	0.738

increase as the age of the women increases (*OR* estimated in 0.483, and the respective 95% credible interval do not include the value 1). This effect of age on the prevalence is well known in the medical literature since the disease is more incident in sexually active women.

4 Concluding remarks

The contribution of the present study is to provide an extension of the Bayesian method proposed by Joseph, Gyorkos and Coupal (1985) for estimating the performance measures of screening tests introducing a vector of covariates. For this purpose, we introduced a Bayesian approach based on a Markov chain Monte Carlo (MCMC) algorithm. The lack of a gold standard is counterbalanced by the introduction of a latent variable G (2.2) that best describe the data simulating a reference test. This latent variable has a Bernoulli distribution with success probability given in function of the performance screening measures and its subjectiveness from the respective prior distributions. An advantage of the proposed methodology is the fact that the number of parameters to be estimated is not limited by the number of observations as it happens when we use the method introduced by Hui and Walter (1980). Even when the sample size is not large, the number of parameters to be estimated is not limited. However, the problem of identifiability is not fully avoided, given that the proposed Bayesian model is very sensitive to choice of the hyperparameters of the prior distributions. This can be demonstrated by a brief sensitivity analysis, and evidences that plausible results would be given

if we incorporate reasonable prior distributions based on prior knowledgement of clinical experts.

The choice of the prior distribution specification of the parameters in the proposed model was based in articles from the literature, followed by a revision by a panel of experts (authors of this article with clinical education). This procedure explicitly incorporates subjective views, and any conclusions drawn from the Bayesian analysis will potentially be sensitive to the choice of prior distribution. Although eliciting and quantifying the prior opinions of clinicians is a difficult task, researchers should be able to choose the prior distributions with thoroughness. A careful verification of the prior information and a subsequent analysis of its changes in the outcomes can result in reasonable estimates for the tests performance measures.

The major shortcoming of the Bayesian estimating method resides in the necessary presumption that the diagnostic tests are statistically and conditionally independent. This presupposition might not be invariably true (Brenner (1996)), and alternative methods were proposed by Espeland and Handelman (1989), Yang and Becker (1997) and Dendukuri and Joseph (2001) to address those situations. However, all of these approaches address situations in which the correlation between two screening tests is considered, and extensions for three or more tests are not found in the literature. Bayesian models that include the conditional dependence between multiple screening tests should be considered in future studies.

Acknowledgments

We are sincerely thankful for the comments made by the anonymous reviewers. This research was funded by Research European Committee of the European Economical Comunity process INCO DEV 4-CT-2001-10013; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant number 99/11264-0 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant number 300354/01-0.

References

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **39** 207–215. MR0712747

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine* **6** 411–423.

Belinson, J. L., Pretorius, R. G., Zhang, W. H., Wu, L. Y., Qiao, Y. L. and Elson, P. (2001). Cervical cancer screening by simple visual inspection after acetic acid. *Obstetrics and Gynecology* **98** 441–444.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*, Reprint ed. Wiley, New York.

Brenner, H. (1996). How independent are multiple "independent" diagnostic classifications? *Statistics in Medicine* **15** 1377–1386.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Chapman & Hall/CRC, London. MR1427749

Christensen, J. and Gardner, I. A. (2000). Herd-level interpretation of test results for epidemiologic studies of animal diseases. *Preventive Veterinary Medicine* **45** 83–106.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. *Biometrics* **57** 208–217. MR1833302

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.

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

Faraone, S. V. and Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry* **151** 650–657.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85** 398–409. MR1141740

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association* **95** 1300–1304. MR1825281

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., eds.) 169–194. Clarendom Press, Oxford. MR1380276

Gontijo, R., Derchain, S., Roteli-Martins, C., Bragança, J., Sarian, L., Morais, S., Maeda, M., Longatto-Filho, A. and Syrjänen, K. (2007). Human papillomavirus (HPV) infections as risk factors for cytological and histological abnormalities in baseline PAP smear-negative women followed-up for 2 years in the LAMS study. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **133** 239–246.

Greiner, M. and Gardner, I. A. (2000). Application of diagnostic tests in veterinary epidemiologic studies. *Preventive Veterinary Medicine* **45** 43–59.

Hadgu, A. and Qu, Y. A. (1998). Biomedical application of latent models with random effects. *Applied Statistics* **47** 603–616.

Hopman, E. H., Kenemans, P. and Helmerhorst, T. J. (1998). Positive predictive rate of colposcopic examination of the cervix uteri: An overview of literature. *Obstetrical & Gynecological Survey* **53** 97–106.

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.

Joseph, L., Gyorkos, T. W. and Coupal, L. (1985). 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.

Koss, L. G. (2000). Human papillomavirus testing as a screening tool for cervical cancer. *Journal of the American Medical Association* **283** 2525–2526.

Martinez, E. Z., Achcar, J. A. and Louzada-Neto, F. (2006). Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach. *Computational Statistics & Data Analysis* **51** 601–611. MR2297474

Martinez, E. Z., Achcar, J. A. and Louzada-Neto, F. (2005). Bayesian estimation of diagnostic tests accuracy for semi-latent data with covariates. *Journal of Biopharmaceutical Statistics* **15** 809–821. MR2190587

Mitchell, M. F., Schottenfeld, D., Tortolero-Luna, G., Cantor, S. B. and Richards-Kortum, R. (1998). Colposcopy for the diagnosis of squamous intraepithelial lesions: A meta-analysis. *Obstetrics and Gynecology* **91** 626–631.

Nanda, K., McCrory, D. C., Myers, E. R., Bastian, L. A., Hasselblad, V., Hickey, J. D. and Matchar, D. B. (2000). Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: A systematic review. *Annals of Internal Medicine* **132** 810–819.

Pouillot, R., Gerbier, G. and Gardner, I. A. (2002). "TAGS", a program for the evaluation of test accuracy in the absence of a gold standard. *Preventive Veterinary Medicine* **53** 67–81.

Qu, Y., Tan, M. and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810. MR1411731

Ransohoff, D. F. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine* **299** 926–930.

Sarian, L. O., Derchain, S. F., Naud, P., Roteli-Martins, C., Longatto-Filho, A., Tatti, S., Branca, M., Erzen, M., Serpa-Hammes, L., Matos, J., Gontijo, R., Bragança, J. F., Lima, T. P., Maeda, M. Y., Lörincz, A., Dores, G. B., Costa, S., Syrjänen, S. and Syrjänen, K. (2005). Evaluation of visual inspection with acetic acid (VIA), Lugol's iodine (VILI), cervical cytology and HPV testing as cervical screening tools in Latin America. This report refers to partial results from the LAMS (Latin AMerican Screening) study. *Journal of Medical Screening* **12** 142–149.

Schiffman, M., Herrero, R., Hildensheim, A., Sherman, M. E., Bratti, M., Wacholder, S., Alfaro, M., Hutchinson, M., Morales, J., Greenberg, M. D. and Lorincz, A. T. (2000). HPV DNA testing in cervical cancer screening: Results from women in a high-risk province of Costa Rica. *Journal of the American Medical Association* **283** 87–93.

Shlay, J. C., Dunn, T., Byers, T., Barón, A. E. and Douglas, J. M. (2000). Prediction of cervical intraepithelial neoplasia grade 2–3 using risk assessment and human papillomavirus testing in women with atypia on Papanicolaou Smears. *Obstetrics and Gynecology* **96** 410–416.

Stamey, J. D., Boese, D. H. and Young, D. M. (2008). Confidence intervals for parameters of two diagnostic tests in the absence of a gold standard. *Computational Statistics & Data Analysis* **52** 1335–1346.

Syrjänen, K., Naud, P., Derchain, S., Roteli-Martins, C., Longatto-Filho, A., Tatti, S., Branca, M., Erzen, M., Hammes, L. S., Matos, J., Gontijo, R., Sarian, L., Bragança, J., Arlindo, F. C., Maeda, M. Y., Lörincz, A., Dores, G. B., Costa, S. and Syrjänen, S. (2005). Comparing PAP smear cytology, aided visual inspection, screening colposcopy, cervicography and HPV testing as optional screening tools in Latin America. Study design and baseline data of the LAMS study. *Anticancer Research* **25** 3469–3480.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82** 528–550. MR0898357

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.

University of Zimbabwe/JHPIEGO Cervical Cancer Project (1999). Visual inspection with acetic acid for cervical-cancer screening: Test qualities in a primary care setting. *The Lancet* **353** 869–873.

Yang, I. and Becker, M. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics* **53** 948–958.

Wright, T. C. Jr., Lynette, D., Kuhn, L., Pollack, A. and Lorincz, A. (2000). HPV DNA testing of self-collected vaginal samples compared with cytologic screening to detect cervical cancer. *Journal of the American Medical Association* **283** 81–86.

Zhou, X. (1983). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics Theory and Methods* **22** 3177–3198. MR1245539

Zhou, X. H. (1998). Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research* **7** 337–353.

E. Z. Martinez
J. A. Achcar
Departamento de Medicina Social
Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo, USP
Ribeirão Preto, SP, 14049-900
Brazil
E-mail: edson@fmrp.usp.br
achcar@fmrp.usp.br

F. Louzada-Neto
Departamento de Estatística
Universidade Federal de São Carlos, UFSCar
São Carlos, SP
Brazil
E-mail: dfln@power.ufscar.br

K. J. Syrjänen
Coordinator of the LAMS Study
Department of Oncology and Radiotherapy
Turku University Hospital
Turku
Finland
E-mail: Kari.Syrjanen@tyks.fi

S. F. M. Derchain
R. C. Gontijo
L. O. Z. Sarian
Brazilian partners of the LAMS Study
Departamento de Tocoginecologia
Faculdade de Ciências Médicas
Universidade Estadual de Campinas, UNICAMP
Campinas, SP
Brazil
E-mail: derchain@fcm.unicamp.br
rgontijo@terra.com.br
sarian@fcm.unicamp.br