Biometrics 59, 670–675
September 2003

Binomial Regression with Misclassification

Carlos Daniel Paulino,[1,*] Paulo Soares,[1] and John Neuhaus[2]

[1]Instituto Superior Técnico, e Centro de Matemática e Aplicações,
Universidade Técnica de Lisboa, Portugal
[2]Department of Epidemiology and Biostatistics, University of California,
San Francisco, California, U.S.A.
[*]*email:* dpaulino@math.ist.utl.pt

Summary. Motivated by a study of human papillomavirus infection in women, we present a Bayesian binomial regression analysis in which the response is subject to an unconstrained misclassification process. Our iterative approach provides inferences for the parameters that describe the relationships of the covariates with the response and for the misclassification probabilities. Furthermore, our approach applies to any meaningful generalized linear model, making model selection possible. Finally, it is straightforward to extend it to multinomial settings.

Key words: Bayesian analysis; Binomial regression; Generalized linear model; Misclassification.

1. Introduction

Currently the analysis of binary response data is still dominated by the classical approach under which inferences are based on asymptotic theory. However, the Bayesian literature on this topic and related ones is steadily growing. See, for example, Dey, Ghosh, and Mallick (2000). Most importantly, the greatest contribution to this increasing ability to deal with more-complex problems and models is the continuous development of computational tools, in particular those based on Monte Carlo methods. Under the Bayesian paradigm it is then natural that the treatment of more problematic cases, such as misclassification, incomplete data, or measurement errors, would receive increasing attention. See, for instance, Geng and Asano (1989), Evans et al. (1996), Mendoza-Blanco, Tu, and Iyengar (1996), and Rekaya, Weigel, and Gianola (2001) for different approaches to misclassified categorical data under several sampling schemes. Soares and Paulino (2001) present an analysis of incomplete categorical data under informative censoring and Thürigen et al. (2000) offer a review of methods for measurement errors in the covariates.

Motivation for this work comes from data gathered in an ongoing study of human papillomavirus (HPV) infection at the University of California-San Francisco (UCSF) (see Moscicki et al. (2001)). The purpose of the investigation is to examine the association of several potential risk factors with HPV cervical infection among females who tested negative at entry into the study. The study screened 104 women aged 13 to 21 years who attended family planning clinics in the San Francisco Bay Area; it recorded for each woman her infection status at the end of the study (HPVS) by testing for HPV DNA in cervical samples, whether she had a history of vulvar warts (VW), whether she had any new sexual partner in the last two months at baseline (NSP), and whether she had an history of herpes simplex (HS). The median follow-up time was 26 months for those women who remained HPV-negative.

HPV is really a family of viruses responsible for various epithelial lesions of which over 90 subtypes have been described. From those, around 30 subtypes have a clear preference for the genital tissues and certain ones are commonly associated with cervical cancer. Since any test for HPV infection is limited to one subtype or a group of subtypes, it will miss a certain number of infections and, therefore, the response variable HPVS is bound to be affected by misclassification, producing some false negative results. Although less probable, false positive results are also possible due to sample contamination and other reasons associated with laboratory work. The definition of initial negative results was made using a conservative criterion: only those from the cohort who at baseline and first follow-up had negative results were included.

In this work, we present a fully Bayesian analysis of binomial regression data with a misclassified response and error-free covariates. In Section 2, the problem is described, along with an informative misclassification model. Section 3 focuses on the use of generalized linear models to analyze the association of the response with the covariates, embodying the ideas found in Bedrick, Christensen, and Johnson (1996) on prior specification. Section 4 discusses the use of data augmentation and other computational issues. In Section 5, the HPV data above mentioned serve as an illustration of the analysis developed and, finally, Section 6 contains some concluding remarks.

2. Model for the Misclassified Binomial Regression Data

Consider regression data $(n_k,\ N_k,\ \mathbf{x}_k)$, $k = 1, \ldots, N$, where the n_k's represent the number of successes from independent binomial distributions, Binomial$(N_k,\ \phi_k)$, the $\mathbf{x}_k$'s are known $p \times 1$ vectors of covariates, and the index k denotes covariate patterns. Due to the action of some corrupting mechanism, the response variable is often classified incorrectly. To accommodate this misclassified response, we conceptually split the data-collecting process into two stages: an unobserved *sampling stage* related to the true response R^T followed by a *reporting stage* where an observed and possibly wrong response R^O is reported.

If we associate a success with $R=1$ and write $\theta_{ki} = P(R^T = i \,|\, \mathbf{x}_k)$, $k=1,\ldots,N$, $i=0,1$, with $\sum_i \theta_{ki} = 1$ and $\lambda_{kij} = P(R^O = j \,|\, R^T = i,\ \mathbf{x}_k)$, $k=1,\ldots,N$, $i,\ j = 0,\ 1$, with $\sum_j \lambda_{kij} = 1$, then the probability of success for the observed response of an individual with covariates $\mathbf{x}_k$ is $\phi_k = \sum_i \lambda_{ki1}\theta_{ki}$. The probability model for the observed regression data, $\mathbf{n} = (n_k,\ N_k)$, is described by the nonidentifiable product-binomial likelihood

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathbf{n}) = \prod_{k=1}^{N} \binom{N_k}{n_k} \left(\sum_i \lambda_{ki1}\theta_{ki} \right)^{n_k} \left(\sum_i \lambda_{ki0}\theta_{ki} \right)^{N_k - n_k}, \tag{1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ stand for the sets of parameters θ_{ki} and λ_{kij}, respectively. Note that both kinds of parameters may depend on the covariates. When the covariate dependence of $\boldsymbol{\lambda}$ is allowed, this means that a differential misclassification mechanism is being adopted. A standard method of analyzing the association between a response variable and several covariates is using generalized linear models. In this case, we can express the expected value of the success proportions as a function of a linear predictor

$$E\left(\frac{n_k}{N_k} \middle| \boldsymbol{\theta} \right) = \theta_{k1} = \theta_1(\mathbf{x}_k'\boldsymbol{\beta}) = g(\mathbf{x}_k'\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector of regression coefficients and $g(\cdot)$ can be an arbitrary c.d.f. Common choices for g include the logistic, normal, and Gumbel (for minima) distribution functions, defined by

$$g(\mathbf{x}'\boldsymbol{\beta}) = \begin{cases} e^{\mathbf{x}'\boldsymbol{\beta}}/(1+e^{\mathbf{x}'\boldsymbol{\beta}}) \\ \Phi(\mathbf{x}'\boldsymbol{\beta}) \\ 1 - \exp(-e^{\mathbf{x}'\boldsymbol{\beta}}). \end{cases} \tag{2}$$

Henceforth, we will always assume that, *a priori,* the distinct quantities $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are independent, motivated by the occurrence of misclassification being inherent to the clinical procedure regardless of $\boldsymbol{\theta}$. Moreover, the inclusion of any form of dependence between these parameters would make the analysis more difficult, without any obvious practical benefit.

3. Subjective Prior Distribution

In a subjective Bayesian analysis, the introduction of further regression parameters leads to a potentially serious problem. Because these parameters do not relate directly to the data, they can look rather esoteric to the practitioner's eyes. This is particularly true when competing choices of the link function, $g^{-1}(\cdot)$, must be considered, each one with a different interpretation of $\boldsymbol{\beta}$. In practice, this can be a subtle, but nevertheless powerful inducement for using convenience priors and, sometimes, to skip altogether the expert prior elicitation step and stick to noninformative or diffuse prior distributions. A recent example of this approach can be seen in Rekaya et al. (2001); they use hierarchical priors of convenience in problems of this type, but with a simplified misclassification structure.

Bedrick et al. (1996) considered this problem of generalized linear models and proposed a method of overcoming it using a so-called conditional means prior (CMP). In brief, a CMP is obtained as follows:

1. Choose p covariate vectors $\tilde{\mathbf{x}}_l$, $l = 1, \ldots, p$;
2. Assign a prior on $\{\theta_1(\tilde{\mathbf{x}}_1'\boldsymbol{\beta}), \ldots, \theta_1(\tilde{\mathbf{x}}_p'\boldsymbol{\beta})\}$;
3. Obtain the induced prior on $\boldsymbol{\beta}$ by using the change-of-variables method.

The choice of the p covariate vectors $\tilde{\mathbf{x}}_l$ should also be subjected to expert opinion and be made so that those vectors are widely spaced in the predictor space in order to make it reasonable to assume prior independence of the quantities $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$. On this subject, we refer the reader to Bedrick et al. (1996) for general guidelines and diagnostic measures for the choice of the $\tilde{\mathbf{x}}_l$'s, and to Bedrick, Christensen, and Johnson (1997) for an illustration.

In the following, we will abandon some of the previous generality and focus on a particular case that will be important in the applications. This is when one specifies that, independently, $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta}) \sim \text{Beta}(c_l, d_l)$, $l = 1, \ldots, p$, and that $g(\cdot)$ is a continuous c.d.f. $F(\cdot)$ with density $f(\cdot)$. Denoting $\theta_0(\tilde{\mathbf{x}}_l'\boldsymbol{\beta}) = 1 - \theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$, the induced prior on $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}) \propto \prod_{l=1}^{p} \{\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}^{c_l - 1} \{\theta_0(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}^{d_l - 1} f(\tilde{\mathbf{x}}_l'\boldsymbol{\beta}). \tag{3}$$

To illustrate a simple logistic regression model, the Jacobian of the transformation $\{\theta_1(\tilde{\mathbf{x}}_1'\boldsymbol{\beta}), \theta_1(\tilde{\mathbf{x}}_2'\boldsymbol{\beta})\} \to \boldsymbol{\beta} \equiv (\beta_0, \beta_1)$, where $\tilde{\mathbf{x}}_l' = (1 \ \ x_l)$, is $(x_2 - x_1)\prod_{l=1}^{2} \theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\theta_0(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$ and hence (3) is $\pi(\boldsymbol{\beta}) = (x_2 - x_1)\prod_{l=1}^{2}[\{\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}^{c_l}\{\theta_0(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}^{d_l}/B(c_l, d_l)]$.

The hyperparameters c_l and d_l are determined (indirectly in general) from expert prior judgments on features of the prior distribution for $\{\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}$. An analogous process is used to obtain the (possibly beta) prior hyperparameters for $\boldsymbol{\lambda}$.

Comparing this prior with the likelihood $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda} \,|\, \mathbf{n}) = \mathcal{L}\{\boldsymbol{\theta}(\boldsymbol{\beta}), \boldsymbol{\lambda} \,|\, \mathbf{n}\}$, we see that the presence of a misclassified response, along with the probabilistic model described in Section 2 to explain that misclassification, leads to a complicated posterior kernel that makes impossible any straightforward inferences using analytical methods. We will see in the next section how the use of data augmentation can alleviate this problem by separating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ in the likelihood.

4. Data Augmentation

Let m_{kij} be the number of observations with $R^T = i$ and $R^O = j$ among those observations with a covariate pattern $\mathbf{x}_k$. For these unobserved quantities, we have $m_{k+1} = \sum_i m_{ki1} = n_k$ and $m_{k+0} = N_k - n_k$, while any other partial sum is also unobserved. The augmented data $\mathbf{m} = (m_{kij})$ is an hypothetical sample from a product of multinomial distributions

$M\{N_k,\ (\lambda_{kij}\theta_{ki})\}$ with a corresponding likelihood under the generalized linear model parameterization,

$$\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{m}\right) \propto \prod_{k,i} \left\{\theta_i(\mathbf{x}_k'\boldsymbol{\beta})\right\}^{m_{ki+}} \prod_{k,i,j} \lambda_{kij}^{m_{kij}}.$$

This likelihood shows that a data augmentation approach can serve our inferential purposes, in the sense that it leads to a factorization $\mathcal{L}\left(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{m}\right) = \mathcal{L}\left(\boldsymbol{\beta} \mid \mathbf{m}\right) \times \mathcal{L}\left(\boldsymbol{\lambda} \mid \mathbf{m}\right)$ that, in a sense, conjugates nicely with the induced prior for $\boldsymbol{\beta}$ in expression (3). In fact, the augmented data posterior distribution is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{m}) \propto \pi(\boldsymbol{\beta} \mid \mathbf{m})\pi(\boldsymbol{\lambda}) \prod_{k,i,j} \lambda_{kij}^{m_{kij}}, \tag{4}$$

where $\pi(\boldsymbol{\lambda})$ is a prior distribution for $\boldsymbol{\lambda}$ and

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \mathbf{m}) \propto & \prod_{l=1}^{p} \left\{\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\right\}^{c_l - 1} \left\{\theta_0(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\right\}^{d_l - 1} f(\tilde{\mathbf{x}}_l'\boldsymbol{\beta}) \\ & \times \prod_{k,i} \left\{\theta_i(\mathbf{x}_k'\boldsymbol{\beta})\right\}^{m_{ki+}}. \end{aligned} \tag{5}$$

For the reporting stage parameters $\boldsymbol{\lambda}$, it may be generally reasonable to assume prior independence among the sets $\{\lambda_{kij}, j = 0, 1\}, \forall k, i$, and to use a beta distribution for each. In this case, their posterior distribution given $\mathbf{m}$ is a product of beta distributions with their hyperparameters updated by $\mathbf{m}$.

Conditionally on the observed data, the augmented data is distributed according to independent binomial distributions for each k

$$\begin{aligned} m_{k01} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{n} &\sim \text{Binomial}\left\{ n_k, \frac{\lambda_{k01}\theta_0(\mathbf{x}_k'\boldsymbol{\beta})}{\sum_i \lambda_{ki1}\theta_i(\mathbf{x}_k'\boldsymbol{\beta})} \right\} \\ m_{k10} \mid \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{n} &\sim \text{Binomial}\left\{ N_k - n_k, \frac{\lambda_{k10}\theta_1\left(\mathbf{x}_k'\boldsymbol{\beta}\right)}{\sum_i \lambda_{ki0}\theta_i\left(\mathbf{x}_k'\boldsymbol{\beta}\right)} \right\}. \end{aligned} \tag{6}$$

From this setup, it seems now possible to draw inferences based on a data augmentation algorithm (see Tanner, 1996), such as the chained data augmentation algorithm (CDA). This algorithm can be viewed as a Gibbs sampler and is formed by the following steps:

(1) Choose adequate initial values $\boldsymbol{\beta}^0$ and $\boldsymbol{\lambda}^0$;
(2) For $i = 1, \ldots, t$:

(a) Imputation step

(i) sample $\mathbf{m}^i$ from the independent binomial distributions in (6) given $\boldsymbol{\beta}^{i-1}$, $\boldsymbol{\lambda}^{i-1}$, and $\mathbf{n}$;

(b) Posterior step

(i) sample $\boldsymbol{\lambda}^i$ from the independent beta distributions given $\mathbf{m}^i$;
(ii) sample $\boldsymbol{\beta}^i$ from $\pi(\boldsymbol{\beta} \mid \mathbf{m})$ in (5) given $\mathbf{m}^i$.

Then, under general conditions, $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{m}^i)$ will converge to $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{n})$ as i goes to infinity (Tanner and Wong, 1987).

The remaining problematic issue is to sample from $\pi(\boldsymbol{\beta} \mid \mathbf{m})$ in step (2)(b)(ii). A possible solution is to resort to the sampling-importance resampling (SIR) method as described by Gelman et al. (1995). For the importance distribution, we found it adequate to use a multivariate student t-density with ν degrees of freedom, with a mode equal to the mode of $\pi(\boldsymbol{\beta} \mid \mathbf{m}^i)$ and dispersion proportional to the asymptotic covariance matrix evaluated at the mode. The value of ν should ensure that the tails of the importance distribution decay slower than those of the target distribution $\pi(\boldsymbol{\beta} \mid \mathbf{m}^i)$.

Table 1
The HPV infection data

$\mathbf{x}_k^*$	n_k	N_k
(0,0,0)	12	44
(0,0,1)	1	2
(0,1,0)	29	40
(0,1,1)	3	3
(1,0,0)	6	9
(1,1,0)	1	4
(1,1,1)	2	2

5. The HPV Data

We will now analyze the problem described in Section 1. Writing $\mathbf{x}^* = (\text{VW, NSP, HS})$, the collected data is presented in Table 1 and a brief descriptive analysis is given in Figure 1, where we can see that, for example, 54 women were diagnosed as infected by the HPV during the study, while 50 others maintained a negative result in the clinical tests.

Following Bedrick et al. (1997), we carried out the prior elicitation with the kind collaboration of Dr Anna-Barbara Moscicki from the Department of Pediatrics at UCSF. Since we have three binary covariates, we needed to choose four covariate configurations to induce the prior on $\boldsymbol{\beta}$. So, Dr Moscicki was asked to choose those four configurations and to provide values for the 1%, 50%, and 99% quantiles of her prior density for the probability of HPV infection of someone with those covariate characteristics. This is a purposeful overspecification of the beta distributions, so we used the two quantiles on which the expert was more confident, 50% and 99%, to calculate the prior hyperparameters, and the third one to assess the consistency of the choice. This process was iterated until a consistent set of values was found and then the hyperparameters were obtained numerically.

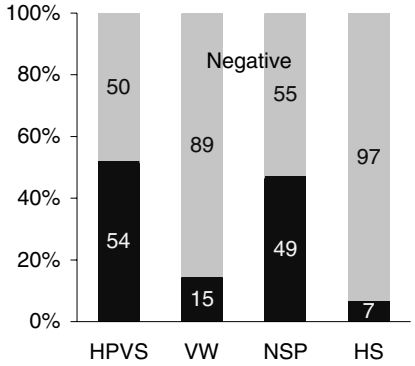

Figure 1. Brief descriptive analysis of the HPV data.

Table 2
Expert elicitation of prior quantiles for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$

	$q_{1\%}$	$q_{50\%}$	$q_{99\%}$	Hyperparameters	
$\tilde{\mathbf{x}}_1 = (1,1,1)$	0.30	0.70	0.95	5.944	2.731
$\tilde{\mathbf{x}}_2 = (1,0,0)$	0.09	0.30	0.60	4.692	10.512
$\tilde{\mathbf{x}}_3 = (0,1,0)$	0.15	0.40	0.70	5.897	8.682
$\tilde{\mathbf{x}}_4 = (0,0,1)$	0.14	0.30	0.50	10.142	23.224
P(false positive)	0.02	0.05	0.10	8.386	153.386
P(false negative)	0.03	0.07	0.20	3.135	37.650

Regarding the misclassification probabilities, it is quite reasonable to argue that the presence of misclassification is not related to the covariates and, consequently, a nondifferential misclassification mechanism can be used. This means that we only have to deal with two parameters: P(false positive) = 1-specificity = λ_{01} and P(false negative) = 1-sensitivity = λ_{10}. The elicitation of the prior hyperparameters for these two misclassification probabilities followed the lines described above. Assessments of the sensitivity and specificity of the HPV diagnostic using known samples (Moscicki et al., 2001) formed the basis of Dr Moscicki's prior opinion about misclassification rates. The complete set of elicited quantiles is given in Table 2.

In this article, we will restrict the analysis to three common generalized linear models, the logistic (M_1), probit (M_2), and complementary log-log models (M_3), for which the c.d.f. is defined by the function g in (2) where $\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1$VW + β_2NSP + β_3HS.

A preliminary simulation study of the importance-sampling weights showed that the degrees of freedom of the importance distribution, ν, need not be very small and that a value of 30 would be fine for all three models in terms of the usual requirement of a small variability for the importance weights. We believe that this is due to the light tails of the posterior of $\boldsymbol{\beta}$, since we found that the posterior results were only mildly affected by the variation of the importance-sampling degrees of freedom.

To select the most appropriate model for the data, we computed Bayes factors to compare M_i and M_j, $\mathrm{BF}_{ij} = P(\mathbf{n} \mid M_i)/P(\mathbf{n} \mid M_j)$, where $P(\mathbf{n} \mid M_i)$ is the marginal probability of observing the data $\mathbf{n}$ from model M_i, that is,

$$P(\mathbf{n} \mid M_i) = \int \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{n})\, \pi(\boldsymbol{\beta}, \boldsymbol{\lambda})\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda},$$

where $\mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\lambda} \mid \mathbf{n})$ is the likelihood for the model M_i. That marginal probability can be estimated by sampling from the prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda})$ of Sections 3 and 4, and then evaluating

$$P(\mathbf{n} \mid M_i) \approx \frac{1}{t} \sum_{l=1}^{t} \mathcal{L}_i(\boldsymbol{\beta}^l, \boldsymbol{\lambda}^l \mid \mathbf{n}).$$

To sample from the marginal prior of $\boldsymbol{\beta}$, it is sufficient to sample from the prior distributions of $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$, $l = 1, \ldots, 4$, and then solve the set of equations $\{\theta_{l1} = g(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})\}$ to obtain samples from $\boldsymbol{\beta}$.

To obtain sufficiently stable estimates, it was necessary to use $t = 10^7$, from which the values obtained were $\mathrm{BF}_{12} = 1.0058$, $\mathrm{BF}_{23} = 1.0015$, and $\mathrm{BF}_{13} = 1.0073$. These values are all close to 1 so if the prior odds for any pair of models are equal to 1, then they are practically unaltered by the data. Therefore, this criterion for model selection does not provide enough evidence for choosing any particular one. This being so, the following results were computed from a sample of 100,000 points obtained with the logistic model, after a proper convergence analysis using the methods available in the CODA (Best, Cowles, and Vines, 1997) or BOA (Smith, 2001) software.

Figures 2 and 3 display the plots of the prior (dashed lines) and posterior densities (solid lines) for λ_{01}, λ_{10}, and $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$, $l = 1, \ldots, 4$, where the posteriors are drawn as smoothed histograms. These figures show that there is no conflict between the prior information and the data, and that the expert's opinion about the misclassification probabilities was very precise; this is particularly true in the false positive case, in the sense that there are no significant differences between the respective prior and posterior densities. In this regard, we should add that a further simulation study (based on uniform priors for λ_{01} and λ_{10}) showed that the data can strongly affect the estimation of the probabilities of a false positive and a false negative (and of $\boldsymbol{\theta}$ as well). Furthermore, point and interval summaries of the posterior distribution of the misclassification probabilities were found to be rather insensitive to changes in $\boldsymbol{\theta}$'s prior distribution strength.

Other posterior estimates are shown in Table 3. The highest priority density (HPD) credible intervals were obtained with Chen and Shao's method (described in Chen, Shao, and Ibrahim (2000)). Those estimates show that NSP is the covariate with the strongest association with HPV infection.

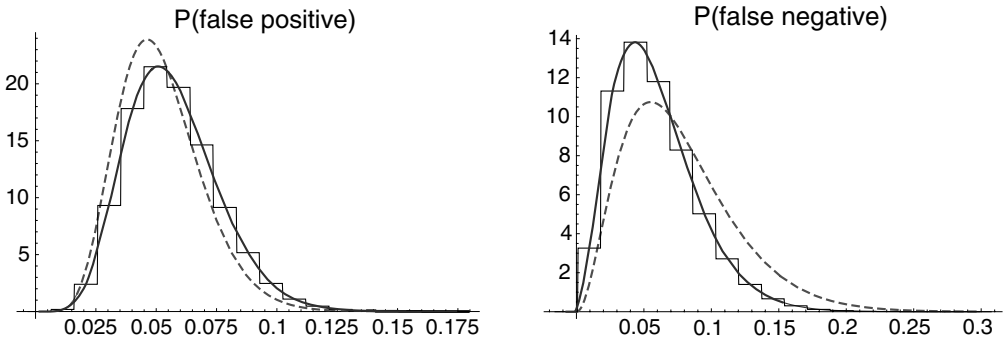

Figure 2. Prior and posterior distributions for $\boldsymbol{\lambda}$ (the dashed lines are for the priors and the solid ones for the posteriors).

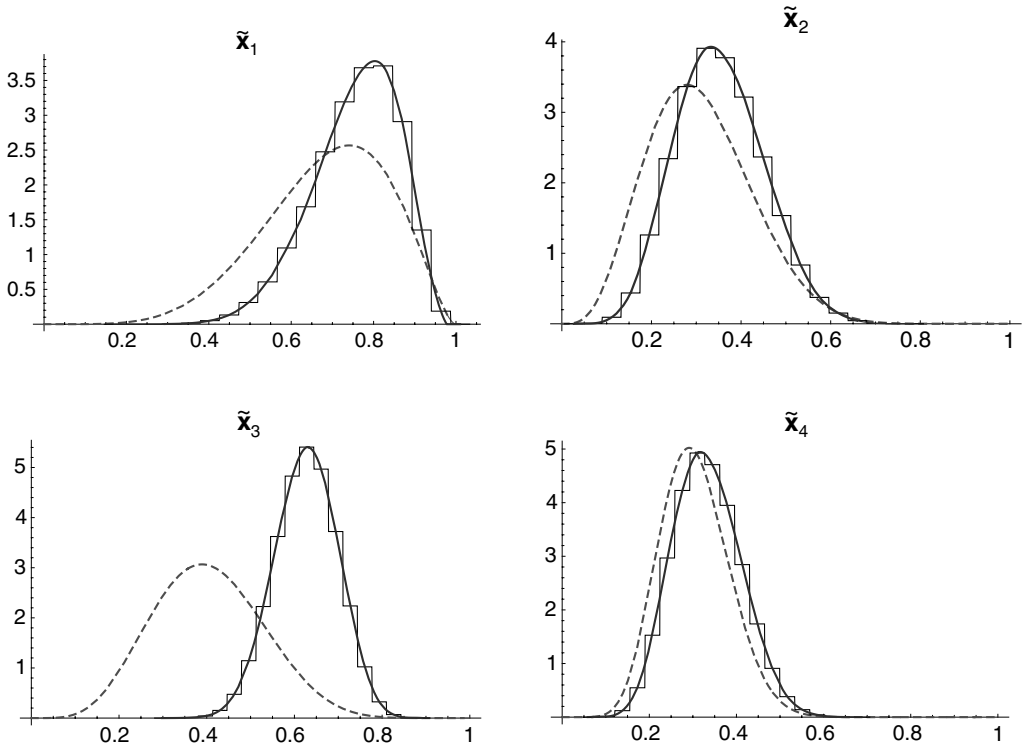

Figure 3. Prior and posterior distributions for $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$ (the dashed lines are for the priors and the solid ones for the posteriors).

This result supports the hypothesis that HPV is a sexually transmitted virus and that most new infections are due to exposure rather than to the development of latent infections.

All the computations were made with specially written OX or Mathematica code that is available from the authors upon request.

6. Concluding Remarks

The presented approach extends the work of Bedrick et al. (1996) to accommodate the observation of binomial regression data with misclassification on the response. In Section 5, the described illustration misses an important point—there are no continuous covariates. However, by construction, this approach does not depend on the covariate nature and so it is immediately applicable to the general case. Naturally, the larger the number of covariates, the more extensive the eliciting work will be and the more complex it is to assure the prior independence between the $\theta_1(\tilde{\mathbf{x}}_l'\boldsymbol{\beta})$'s. Despite the undeniable advantages of the Bayesian operation on drawing inferences in settings like this, model nonidentifiability requires special care in the prior elicitation process because, as is well known, the posterior inferences of nonidentifiable parameters are strongly influenced by the prior, even for increasingly large sample sizes. Another simple extension would be the consideration of polychotomous response data.

On the negative side, the use of the SIR method within the chained data augmentation algorithm is currently a weakness of this approach. Since there is no way to completely ensure the "quality" of the SIR results, alternatives, such as advanced importance sampling, slice sampling, or even perfect sampling (see Liu (2001)), are badly needed. Ongoing preliminary work has already allowed us to validate the SIR results using a much faster slice sampling algorithm.

Table 3
Some posterior estimates for the HPV data

	Mean	MC error	S.d.	Median	HPD CI (95%)	
P(false positive)	0.0568	0.0001	0.0188	0.0549	0.0228	0.0941
P(false negative)	0.0586	0.0001	0.0311	0.0535	0.0082	0.1202
Intercept	−1.0215	0.0013	0.3321	−1.0108	−1.6715	−0.3786
VW	0.3861	0.0016	0.4579	0.3824	−0.5071	1.2803
NSP	1.5506	0.0014	0.3933	1.5452	0.7732	2.3059
HS	0.3028	0.0014	0.4230	0.3017	−0.5249	1.1233

Acknowledgements

The authors thank Dr Anna-Barbara Moscicki from the University of California-San Francisco for the use of the data and for her kind collaboration in the prior elicitation. They also gratefully acknowledge the helpful comments and questions raised by the associate editor and the referees. The first author's research was partially supported by the Luso-American Foundation for Development (FLAD) and the Centre for Mathematics and Its Applications (CEMAT). Grants from the U.S. National Institutes of Health and the Portuguese Studies Program at the University of California-Berkeley also supported this research.

Résumé

Motivés par l'étude de l'infection à virus papilloma humain chez la femme, nous présentons une analyse de régression binomiale bayésienne dans laquelle la réponse est sujette à un processus d'erreur de classement sans contrainte. Notre approche itérative fournit des inférences pour les paramètres qui décrivent la relation des covariables avec la réponse, et pour les probabilités d'erreur de classement. De plus, notre approche s'applique à n'importe quel modèle linéaire généralisé pertinent, rendant possible la sélection de modèle, et il est immédiat de l'étendre à un contexte multinomial.

References

Bedrick, E. J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91,** 1450–1460.

Bedrick, E. J., Christensen, R., and Johnson, W. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *American Statistician* **51,** 211–218.

Best, N., Cowles, M., and Vines, K. (1997). *CODA—Convergence Diagnosis and Output Analysis software for Gibbs Sampling Output,* Version 0.4. Cambridge: MRC Biostatistics Unit.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation.* New York: Springer.

Dey, D. K., Ghosh, S. K., and Mallick, B. K. (2000). *Generalized Linear Models: A Bayesian Perspective.* New York: Marcel Dekker.

Evans, M., Guttman, I., Haitovsky, Y., and Swartz, T. (1996). Bayesian analysis of binary data subject to misclassification. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner,* D. Berry, K. Chaloner, and J. Geweke (eds), 67–77. New York: North Holland.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall.

Geng, Z. and Asano, C. (1989). Bayesian estimation methods for categorical data with misclassification. *Communications in Statistics* **8,** 2935–2954.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* New York: Springer.

Mendoza-Blanco, J. R., Tu, X. M., and Iyengar, S. (1996). Bayesian inference on prevalence using a missing-data approach with simulation-based techniques: Applications to HIV screening. *Statistics in Medicine* **15,** 2161–2176.

Moscicki, A.-B., Hills, N., Shiboski, S., et al. (2001). Risks for incident human papillomavirus infection and low-grade squamous intraepithelial lesion development in young females. *Journal of the American Medical Association* **285,** 2995–3002.

Rekaya, R., Weigel, K. A., and Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics* **57,** 1123–1129.

Soares, P., and Paulino, C. D. (2001). Incomplete categorical data analysis: A Bayesian perspective. *Journal of Statistical Computation and Simulation* **69,** 157–170.

Smith, B. J. (2001). *Bayesian Output Analysis Program (BOA), User's Manual,* Version 1.0.0. Brian J. Smith, Dept. of Biostatistics, University of Iowa, College of Public Health.

Tanner, M. A. (1996). *Tools for Statistical Inference,* 3rd edition. New York: Springer.

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

Thüringen, D., Spiegelman, D., Blettner, M., Heuer, C., and Brenner, H. (2000). Measurement error correction using validation data: A review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research* **9,** 447–474.

Received March 2002. *Revised February* 2003.
Accepted February 2003.