A Bayesian approach for misclassified ordinal response data

Article *in* Journal of Applied Statistics · February 2019

DOI: 10.1080/02664763.2019.1582613

A Bayesian approach for misclassified ordinal response data

Lizbeth Naranjo[a*], Carlos J. Pérez[b], Jacinto Martín[b], Timothy Mutsvari[c], and Emmanuel Lesaffre[d]

[a]*Departamento de Matemáticas, Facultad de Ciencias, Universidad Nacional Autónoma de M éxico (UNAM), 04510 México D.F., Mexico*; [b]*Departamento de Matemáticas, Facultad de Ciencias, Universidad de Extremadura, 06006 Badajoz, Spain*; [c]*Arlenda S.A., Chaussée Verte 93, 4470 Saint-Georges-Sur-Meuse, Belgium*; [d]*L-BioStat, School of Public Health, KU Leuven, 3000 Leuven, Belgium*

(*January 31, 2019*)

Motivated by a longitudinal oral health study, the Signal-Tandmobiel® study, a Bayesian approach has been developed to model misclassified ordinal response data. Two regression models have been considered to incorporate misclassification in the categorical response. Specifically, probit and logit models have been developed. The computational difficulties have been avoided by using data augmentation. This idea is exploited to derive efficient Markov chain Monte Carlo methods. Although the method is proposed for ordered categories, it can also be implemented for unordered ones in a simple way. The model performance is shown through a simulation-based example, and the analysis of the motivating study.

Keywords: Bayesian analysis; Data augmentation; Markov chain Monte Carlo methods; Misclassification; Ordinal regression model

Classification codes: 62F15; 62J99; 62P10

1. Introduction

Dental caries is one of the most prevalent chronic diseases worldwide, affecting individuals of all ages. Many epidemiological surveys and clinical studies are carried out to obtain a further understanding of this disease. However, the process of detecting caries experience (CE) is not obvious. CE scoring may not perfectly reflect the tooth's true condition, and therefore, the presence of CE can be misdiagnosed, leading to misclassified outcomes. In order to standardize data collection techniques in epidemiological surveys and clinical trials, CE assessment guidelines have been developed by the International Caries Detection and Assessment System ([14]). These guidelines highlight the need for training the examiners and measuring the reliability of the obtained scores. However, despite these criteria, the process of CE detection is still subject to misclassification.

In situations where misclassification may happen, additional parameters are necessary to correct the bias yielded by using error-prone data. If misclassification in a data-generating process is not properly modeled, the information may be perceived as being more accurate than it actually is, leading, in many cases, to a non-optimal decision-making. Therefore, correction for misclassification is needed to obtain unbiased estimates

*Corresponding author. Email: lizbethna@ciencias.unam.mx

for the regression coefficients. Statistical models addressing misclassification should be available in these contexts.

Several models to address misclassification on binary regression have been proposed in the statistical literature, see, e.g., [26], [23], [19], [22], [18], and [21]. [4, 5, 13] presented reviews on the effects of misclassification on model estimates. However, there are few models that consider measurement error for polychotomous responses. Computations in multidimensional settings are more difficult, and this case is not an exception. [2] proposed a class of models to analyze repeated monotonic ordinal responses with diagnostic misclassification. They separately modeled the underlying monotonic response and the misclassification process, by developing an EM algorithm for maximum likelihood estimation that incorporates covariates and randomly missing data. [20] presented a Bayesian approach for correcting interobserver measurement error in an ordinal logistic regression model taking into account the variability of the estimated correction terms. [28] considered a multivariate probit model for correlated binary responses. Some of the responses were subject to classification errors and hence they were not directly observable. Besides, measurements on some of the predictors were not available, instead measurements on their surrogate were available. However, the conditional distribution of the unobservable predictors given the surrogate was completely specified. The authors proposed models based on the likelihood approach that takes into account one or both of these sources of errors. [25] developed a class of parametric models that generalizes the multivariate model and the errors-in-variables model to analyze ordinal data. They assumed a general model structure to accommodate the information that is obtained via surrogate variables. The authors developed a hybrid Gibbs sampler to estimate the model parameters and applied the parameter expansion technique to the correlation structure of the multivariate probit models to obtain a rapidly converging algorithm. Recently, [27] developed an ordered probit model that corrects for the classification errors in ordinal responses and/or measurement error in covariates by using maximum likelihood.

Motivated by a longitudinal oral health study, in this paper, the Signal-Tandmobiel® study (ST), a Bayesian approach is proposed to address misclassified ordinal response data. A regression model is developed to incorporate misclassification in the categorical response. A data augmentation framework is proposed to derive Markov chain Monte Carlo (MCMC) algorithms to polychotomous response data that are subject to misclassification. Although only the ordered case is explored, the approach can be adapted for unordered categories. This approach generalizes the binary probit regression model addressing misclassification proposed by [21] and the data augmentation scheme for ordinal regression models proposed by [1] and [9]. The model performance is illustrated with a simulated-based example, and the analysis of the motivating ST data is presented.

The outline of the paper is as follows. The ST study is introduced in Section 2, illustrating the need of addressing misclassification. The way misclassification is addressed in polychotomous response data models is presented in Section 3. In Section 4, the prior distributions are described and the posterior distributions are explored. Section 5 shows the model performance for a simulation-based example, whereas the analysis of the ST data is presented in Section 6. Finally, Section 7 presents the conclusion and some future research lines. The paper is completed with Appendix A showing details of the algorithms of the ordinal response data model considering misclassification by using probit and logit link functions.

2. The Signal-Tandmobiel® study

The Signal-Tandmobiel® study is a longitudinal prospective oral health intervention project conducted in Flanders (North of Belgium), between 1996 and 2001. For this project, 4468 children (2315 boys and 2153 girls) were examined on a yearly basis during their primary school time (between 7 and 12 years of age) by one of sixteen trained dentists (examiners) based on visual and tactile observations. The clinical examinations took place in a mobile dental clinic, with a standard chair and artificial dental light. No radiographs were taken. Data on oral hygiene and dietary habits were obtained through structured questionnaires completed by the parents. For a more detailed description of the study design and research methods, see [32].

The status of surfaces of the teeth were recorded using various modalities. In this work, two ordinal outcomes at both mouth and tooth levels have been used, which have been generated from these original records. Specifically, the ordinal outcome at mouth level is based on the *dmft* index, which is a count that measures CE in deciduous teeth and it is a sum of the number of decayed, missing due to caries and filled teeth, ranging from 0 (no caries experience) to 20 (all teeth affected). Then, a percentage is obtained from the *dmft* index divided by the total teeth in the mouth, and the ordinal outcome at mouth is defined as the following: 1="$0\% - 1\%$", 2="$> 1\% - 10\%$", 3="$> 10\% - 50\%$", and 4="$> 50\% - 100\%$". The ordinal outcome at tooth level denotes the number of CE surfaces for a specific tooth. The number of surfaces is 4 for incisors and canines, while it is 5 for pre-molars and molars. Consequently, on surface level the response is maximally 4 or 5 depending on the tooth. Therefore, this ordinal outcome at tooth level has four ordinal categories.

The statistical findings were applied to the scoring of the four permanent first molars, i.e., teeth 16 and 26 on the maxilla (upper quadrants), and teeth 36 and 46 on the mandible (lower quadrants). The numbering of teeth follows the notation of the Federation Dentaire Internationale, which indicates the position of the tooth in the mouth. Diagnosing CE is difficult for a variety of reasons. For instance, composite materials can imitate the natural enamel so well that it is difficult to spot a restored lesion; or the location of the cavity, far back in the mouth, hampers the view of the dental examiner. But the dental examiner could also classify discolorations as CE. Hence, it is likely to happen in practice that CE is underrated or overrated.

In the ST study, 16 dental examiners were calibrated for scoring CE. The calibration exercises were performed according to the guidelines of training and calibration published by the British Association for the Study of Community Dentistry (BASCD, [24]). The calibration of the dental examiners was performed by comparing their scores on the tooth surfaces of a group of children to those of a benchmark examiner. Note that there exists no infallible scorer for CE. The best one can do is to take a very experienced dental examiner, called benchmark (see [33]), which is assumed to be error-free or is nearly so. In order to maintain a high level of intra- and inter-examiner reliability, calibration exercises were carried out twice a year for all examiners involved. During the study period (1996-2001), three calibration exercises were devoted to the scoring of CE (1996, 1998, 2000), involving 92, 32 and 224 children, respectively. A contingency table of dental examiners and the benchmark examiner was determined, yielding a table with misclassified scores. Data of the three calibration exercises were combined into one validation dataset, and also examiners' data were combined into one. All examiners were lumped together, but the approach can be generalized to take into account multiple examiners. The results suggested that examiners overscore or underscore the true CE status.

In the main dataset the dental examiners scored the children, but their scores are likely to be prone to error. Ignoring in the statistical analysis that the levels of CE lesion severity are prone to misclassification might lead to wrong estimates, and so, to wrong conclusions. Bayesian ordinal regression models considering misclassification can help to provide better predictions than standard models.

3. Addressing misclassification in polychotomous response data models

Suppose that n independent random variables $Y_1, \dots, Y_n$ are observed, where Y_i takes one of J categories, $i = 1, \dots, n$. Suppose that $Y_1, \dots, Y_n$ are prone to error. Let $\theta_{is} = \mathrm{p}(Y_i = s|\boldsymbol{x}_i)$ denote the probability that the i-th observation with covariate pattern $\boldsymbol{x}_i$ is classified in the s-th category (it is possibly misclassified), $s = 1, \dots, J$. The parameters θ_{is} are related to a set of covariates $\boldsymbol{x}_i$ through a regression model that considers misclassification. They are defined as

$$\theta_{is} = \mathrm{p}(Y_i = s|\boldsymbol{x}_i) = \sum_{r=1}^{J} \mathrm{p}(Y_i = s|v_i = r)\mathrm{p}(v_i = r|\boldsymbol{x}_i) = \sum_{r=1}^{J} \lambda_{sr} p_{ir},$$

where $\lambda_{sr} = \mathrm{p}(Y_i = s|v_i = r)$ is the probability that an observation y_i is classified in the s-th category when the true category v_i is the r-th one, and $p_{ir} = \mathrm{p}(v_i = r|\boldsymbol{x}_i)$ denotes the probability that the true category for an observation with covariate pattern $\boldsymbol{x}_i$ is the r-th.

Note that $\boldsymbol{v} = (v_1, \dots, v_n)$, is an unknown random vector of the true classifications, and v_i has a categorical distribution $v_i \sim \mathrm{Cat}(p_{i1}, \dots, p_{iJ})$, with the vector of probabilities $(p_{i1}, \dots, p_{iJ})$, and $\sum_{r=1}^{J} p_{ir} = 1$, where p_{ir} depends on $g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown regression parameters, and g is the link function that usually depends on a cumulative distribution function (cdf). The most common link functions are the logit and probit links, coming from the cdf of a logistic and a normal distribution, respectively (see, for example, [15] and [8]). These two link functions will be considered to develop the proposed regression models.

The likelihood function for a model considering misclassification can be expressed as

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda}|\boldsymbol{y}, \boldsymbol{x}) \propto \prod_{i=1}^{n} \left(\sum_{s=1}^{J} \sum_{r=1}^{J} \lambda_{sr} p_{ir} I[y_i = s] \right),$$

where $I[\cdot]$ denotes the indicator function, i.e., $I[A] = 1$ if A is true, and $I[A] = 0$ otherwise, and $\boldsymbol{\lambda}$ is a matrix

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_J \end{pmatrix}' = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1J} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{J1} & \lambda_{J2} & \cdots & \lambda_{JJ} \end{pmatrix},$$

where $\sum_{s=1}^{J} \lambda_{sr} = 1$, and the elements of the diagonal, λ_{rr} for $r = 1, \dots, J$, denote the probabilities of correct classification.

Latent variables related with the misclassification are introduced to simplify the generation process. Binary latent variables c_{sr}^i are defined, $r, s = 1, \dots, J$, where r represents the index for the true value and s represents the index for the observed value. When the latent variable takes value one, it denotes the group where the ith observation has been assigned: the true category is r and the observed category is s, i.e. $c_{sr}^i = 1$ if $v_i = r$ and

$y_i = s$. Note that $c^i_{s+} = \sum_{r=1}^J c^i_{sr} = 1$ when the observed category is s, i.e. $y_i = s$, and $c^i_{+r} = \sum_{s=1}^J c^i_{sr} = 1$ means that the true category is r, i.e. $v_i = r$, other sums are zero. For each $i = 1, \ldots, n$, a latent matrix $\boldsymbol{c}^i$ is defined

$$\boldsymbol{c}^i = \begin{pmatrix} c^i_{11} & c^i_{12} & \cdots & c^i_{1J} \\ c^i_{21} & c^i_{22} & \cdots & c^i_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ c^i_{J1} & c^i_{J2} & \cdots & c^i_{JJ} \end{pmatrix}.$$

Then, an augmented likelihood function is considered

$$\mathcal{L}(\boldsymbol{p}, \boldsymbol{\lambda} | \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{i=1}^n \left(\left[\prod_{r=1}^J \prod_{s=1}^J \lambda_{sr}^{c^i_{sr}} \right] \left[\prod_{r=1}^J p_{ir}^{c^i_{+r}} \right] \left[\sum_{s=1}^J I[y_i = s] I[c^i_{s+} = 1] \right] \right).$$

This data augmentation scheme allows us to derive easy-to-implement Gibbs sampling algorithms in the context of polychotomous regression models considering misclassification.

4. Exploring the posterior distributions in ordered categories

In this section the prior distributions are presented, which together with the specifications of the previous section allow us to derive MCMC sampling algorithms (see [12]) to sample from the posterior distributions.

For ordered response categories the ordinal regression model is defined by cutpoints $\kappa_0, \kappa_1, \ldots, \kappa_{J-1}, \kappa_J$ considering that $p_{ir} = \Psi(\kappa_r - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Psi(\kappa_{r-1} - \boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{J-1})'$ is the vector of unknown cutpoints, $\kappa_0 = -\infty$, $\kappa_J = \infty$, and Ψ is a cdf (see [1]). Note that if a constant term is included in $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ includes an intercept, then there are only $J-2$ unknown cutpoints $\kappa_2, \ldots, \kappa_{J-1}$, with $\kappa_1 = 0$ (see [15]), $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{J-1})$. This option is often less complex for numerical stability in sampling (see [7]).

4.1 *Prior distributions*

There is some literature addressing informative prior elicitation for generalized linear models. [3] proposed the conditional means prior approach to introduce a prior distribution on the regression parameters, and [6] proposed power prior distributions for the regression parameters based on the notion of availability of historical data. All of them assume models without errors. The literature about informative prior elicitation to binomial regression models with misclassification is mainly focused on the regression parameter vector, see some applications in [19], [22] and [21].

The prior distributions for the regression parameter vector and the cutpoints are flat. For the misclassification parameters, since $\lambda_{sr} \in (0,1)$ and $\sum_{s=1}^J \lambda_{sr} = 1$, for $r, s = 1, \ldots, J$, the natural prior distributions for $\boldsymbol{\lambda}_r$ are Dirichlet, i.e. $\boldsymbol{\lambda}_r \sim \text{Dirichlet}(a_{1r}, \ldots, a_{Jr})$, where $a_{sr} > 0$, whose probability density function (pdf) is

$$\pi(\boldsymbol{\lambda}_r) = \frac{\Gamma\left(\sum_{s=1}^J a_{sr}\right)}{\prod_{s=1}^J \Gamma(a_{sr})} \prod_{s=1}^J \lambda_{sr}^{a_{sr}-1} \propto \prod_{s=1}^J \lambda_{sr}^{a_{sr}-1}.$$

When the response is presented in an ordinal scale, adjacent categories have bigger risk to be misclassified, so that the natural constraints are $\lambda_{1r} < \cdots < \lambda_{r-1,r} < \lambda_{rr}$ and $\lambda_{rr} > \lambda_{r+1,r} > \cdots > \lambda_{Jr}$. In case of nominal response data, the categories are not related, however, a natural constraint is to assume that the correct classification probability is greater than the misclassification probabilities, i.e., $\lambda_{rr} > \lambda_{sr}$. Alternative sets of constraints on the parameters are presented by [30]. Besides, some restrictions on the parameters values such like $a_{sr} \leq \lambda_{sr} \leq d_{sr}$, with a_{sr} and d_{sr} fixed values, can be specified by using truncated Dirichlet distributions (see [11]).

4.2 *Posterior distributions*

The joint posterior distribution of the unobservables $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, $\boldsymbol{c}$, and $\boldsymbol{\lambda}$ is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{c}, \boldsymbol{\lambda} | \boldsymbol{y}, \boldsymbol{x}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\kappa})\pi(\boldsymbol{\lambda})\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\lambda} | \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{x}).$$

In order to derive the Gibbs sampling algorithm, the full conditional distributions must be obtained. The full conditional distributions for $\boldsymbol{\lambda}$ and $\boldsymbol{c}$ are easy to obtain. Specifically, the full conditional distributions for $\boldsymbol{c}$ given $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, $\boldsymbol{\lambda}$, the data $\boldsymbol{y}$, and the covariates $\boldsymbol{x}$ is

$$\pi(\boldsymbol{c} | \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{i=1}^{n} \left(\left[\prod_{r=1}^{J} \prod_{s=1}^{J} \lambda_{sr}^{c_{sr}^{i}} \right] \left[\prod_{r=1}^{J} p_{ir}^{c_{+r}^{i}} \right] \left[\sum_{s=1}^{J} I[y_i = s] I[c_{s+}^{i} = 1] \right] \right),$$

where $\sum_{r=1}^{J} \sum_{s=1}^{J} c_{sr}^{i} = 1$, so that for $i = 1, \ldots, n$,

$$\left[(c_{s1}^{i}, \ldots, c_{sJ}^{i}) | \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x} \right] \sim \text{Multinomial}\left(1, (\pi_{c_{s1}^{i}}, \ldots, \pi_{c_{sJ}^{i}})\right) I[y_i = s], \tag{1}$$

and $(c_{j1}^{i}, \ldots, c_{jJ}^{i}) = (0, \ldots, 0)$ for $j \neq s$, where

$$\pi_{c_{sr}^{i}} = \frac{\lambda_{sr} p_{ir}}{\sum_{j=1}^{J} \lambda_{sj} p_{ij}} I[y_i = s].$$

The full conditional distribution for $\boldsymbol{\lambda}$ given $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, $\boldsymbol{c}$, the data $\boldsymbol{y}$, and the covariates $\boldsymbol{x}$ is

$$\pi(\boldsymbol{\lambda} | \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{r=1}^{J} \prod_{s=1}^{J} \lambda_{sr}^{\sum_{i=1}^{n} c_{sr}^{i} + a_{sr} - 1},$$

where $\sum_{s=1}^{J} \lambda_{sr} = 1$, so that for $r = 1, \ldots, J$,

$$[\boldsymbol{\lambda}_r | \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{x}] \sim \text{Dirichlet}\left(\sum_{i=1}^{n} c_{1r}^{i} + a_{1r}, \ldots, \sum_{i=1}^{n} c_{Jr}^{i} + a_{Jr} \right). \tag{2}$$

However, the full conditional distributions $\pi(\boldsymbol{\beta} | \boldsymbol{\kappa}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x})$ and $\pi(\boldsymbol{\kappa} | \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x})$ do not have closed expressions to easily generate from, because these are given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{\kappa}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \prod_{r=1}^{J} p_{ir}^{c_{+r}^{i}},$$
$$\pi(\boldsymbol{\kappa}|\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) \propto \pi(\boldsymbol{\kappa}) \prod_{i=1}^{n} \prod_{r=1}^{J} p_{ir}^{c_{+r}^{i}}.$$

Our proposal considers the introduction of latent variables in order to allow other easy-to-sample steps within this Gibbs sampling. These latent variables are based on the data augmentation framework of the ordinal regression model proposed by [1] and [9]. Independent latent continuous random variables $z_1, \ldots, z_n$ are assumed, $\boldsymbol{z} = (z_1, \ldots, z_n)$, whose cdf is given by Ψ, $c_{+r}^{i} = 1$ if $\kappa_{r-1} < z_i < \kappa_r$, and $c_{+r}^{i} = 0$ otherwise. The new joint full conditional posterior distribution of $\boldsymbol{z}$ and $\boldsymbol{\kappa}$ is the analogous to the one defined in [9], and it is given by

$$\pi(\boldsymbol{z}, \boldsymbol{\kappa}|\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) = \pi(\boldsymbol{\kappa}|\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x})\pi(\boldsymbol{z}|\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}),$$

where

$$\pi(\boldsymbol{\kappa}|\boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{r=1}^{J} \prod_{\{i:c_{+r}^{i}=1\}} \left[\Psi\left(\kappa_r - \boldsymbol{x}_i'\boldsymbol{\beta}\right) - \Psi\left(\kappa_{r-1} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right] I\left[0 \leq \kappa_r\right],$$
$$\pi(\boldsymbol{z}|\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{x}) \propto \prod_{r=1}^{J} \prod_{\{i:c_{+r}^{i}=1\}} \frac{(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})}{\left[\Psi\left(\kappa_r - \boldsymbol{x}_i'\boldsymbol{\beta}\right) - \Psi\left(\kappa_{r-1} - \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right]},$$

where ψ is the pdf of z_i and $0 = \kappa_1 < \kappa_2 < \cdots < \kappa_{J-1}$. See [9] for more details.

In order to generate from these distributions, Metropolis-Hastings steps are considered. Then, the conditional posterior distributions of $\boldsymbol{z}$, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ are obtained by using the multivariate Metropolis-Hastings-within-Gibbs update step algorithm defined by [9] for ordinal models. Appendix A shows the Metropolis-Hastings steps for probit and logit models.

5. Simulation-based example

A simulation-based study has been carried out to analyze the model performance of the proposed approach. In this section, an example is presented to illustrate the advantages of using this approach.

Different criteria have been considered for model performance. The deviance information criterion (DIC) proposed by [29] is evaluated as:

$$DIC = 2\overline{D(\eta)} - D(\overline{\eta}),$$

where $D(\eta) = -2\log L(\eta)$ is the deviance of the model, $L(\eta)$ is the likelihood, $\overline{D(\eta)} = \mathrm{E}(D(\eta)|\text{data})$ is the posterior mean of the deviance, and $D(\overline{\eta})$ is the deviance at the posterior means of the parameters of interest $\overline{\eta} = \mathrm{E}(\eta|\text{data})$. Another criterion is the total variation distance (TVD) between the true and estimated probabilities. It is defined as:

$$TVD = \sum_{i=1}^{n} \sum_{r=1}^{J} \left| \theta_{ir}^{true} - \widehat{\theta}_{ir} \right| ,$$

and it has been proposed to measure the discrepancy between the true probabilities (which are known for simulated data) and the estimated ones in simulation-based scenarios (see [21]). Finally, the third criterion that will be considered in this section is the pseudo-predictors method (see [10]). The variables θ_{ir}^{obs}, where $\theta_{ir}^{obs} = 1$ if $y_i = r$ and $\theta_{ir}^{obs} = 0$ otherwise, correspond to the observed probabilities for category r at the ith observation in contrast to the predicted probabilities $\widehat{\theta}_{ir}$. When category r is observed at the observation i, it is clear that a good model fit leads to a high probability $\widehat{\theta}_{ir}$, and to small probabilities $\widehat{\theta}_{ij}$ for other categories $j \neq r$. Large differences should be penalized more than small differences. Then, the verification score is defined as:

$$S = \frac{1}{n} \sum_{r=1}^{J} \sum_{i=1}^{n} \left(\theta_{ir}^{obs} - \widehat{\theta}_{ir} \right)^2 ,$$

providing an idea of the goodness-of-fit. For all the criteria, models with smaller criteria values are preferred over models with large values.

Multiple misclassified ordinal response data are generated. The main objective is to compare the performance of the proposed ordinal regression models addressing misclassification with the standard ordinal regression ones. These simulation-based scenarios allow to compare the predictive outcomes with the true ones instead of comparing them with the observed ones (which are subject to misclassification). Also, these scenarios allow to know which model performs better.

The generating process is as follows. A covariate set is generated by $x_{il} \sim \mathrm{U}(0,1)$, for $i = 1, \ldots, n$ and $l = 2, \ldots, k$ and $x_{i1} = 1$, by using several sample sizes $n = \{300, 500\}$ and different number of categories and $J = \{3, 4\}$. The vectors of regression parameters are $\boldsymbol{\beta} = (-2, 3, 3)'$ ($k = 3$) and $\boldsymbol{\beta} = (-2, 3, 3, 3)'$ ($k = 4$), and the vectors of cutpoints are $\boldsymbol{\kappa} = (0, 2)'$ and $\boldsymbol{\kappa} = (0, 2, 4)'$ ($\kappa_0 = -\infty$ and $\kappa_J = \infty$). Two link functions are considered and $\mathcal{D}_\Psi$ denotes the distribution related with the cdf Ψ of the standard normal distribution, $\mathrm{N}(0,1)$, or with the cdf Ψ of the standard logistic distribution, $\mathrm{L}(0,1)$. The true ordinal dependent variable v is randomly generated by using the following process: (i) compute $p_{ir} = \Psi(\kappa_r - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Psi(\kappa_{r-1} - \boldsymbol{x}_i'\boldsymbol{\beta})$ for $r = 1, \ldots, J$, (ii) generate v_i from the distribution Categorical$(p_{i1}, \ldots, p_{iJ})$. Now, the outcomes are randomly misclassified according to the following process: (iii) generate $u_i \sim \mathrm{U}(0,1)$, (iv) for $v_i = r$ define $y_i = 1$ if $0 < u_i \leq \lambda_{1r}$, and define $y_i = s$ with $s = 2, \ldots, J$ if $\sum_{j=1}^{s-1} \lambda_{jr} < u_i \leq \sum_{j=1}^{s} \lambda_{jr}$, where λ_{sr} are the elements of the matrix of misclassification probabilities. For the misclassification probabilities two cases have been considered assuming that the misclassification is assigned to upper or lower categories. For $J = 3$ it is

$$\boldsymbol{\lambda}_{upper} = \begin{pmatrix} 0.75 & 0 & 0 \\ 0.25 & 0.75 & 0 \\ 0 & 0.25 & 1 \end{pmatrix} \quad \text{or} \quad \boldsymbol{\lambda}_{lower} = \begin{pmatrix} 1 & 0.25 & 0 \\ 0 & 0.75 & 0.25 \\ 0 & 0 & 0.75 \end{pmatrix} ,$$

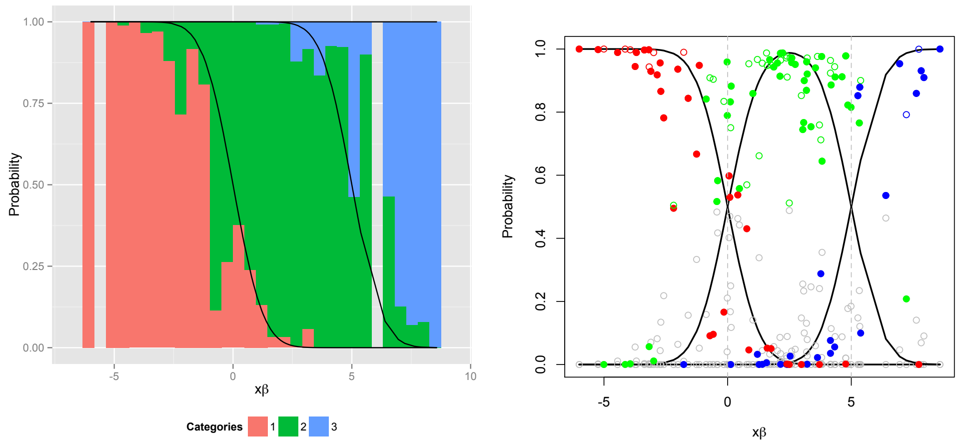

Figure 1. Dataset with ordinal misclassified data. The black lines represent the theoretical probabilities. The first graph shows the three shaded areas of the stacked bar chart that represent the empirical probabilities. The second graph shows the misclassified data: ∘ (empty dots) denote the probabilities p_{ir}, and • (filled dots) denote the category of each observation y_i.

and for $J = 4$ it is

$$\boldsymbol{\lambda}_{upper} = \begin{pmatrix} 0.70 & 0 & 0 & 0 \\ 0.25 & 0.70 & 0 & 0 \\ 0.05 & 0.25 & 0.75 & 0 \\ 0 & 0.05 & 0.25 & 1 \end{pmatrix} \quad \text{or} \quad \boldsymbol{\lambda}_{lower} = \begin{pmatrix} 1 & 0.25 & 0.05 & 0 \\ 0 & 0.75 & 0.25 & 0.05 \\ 0 & 0 & 0.70 & 0.25 \\ 0 & 0 & 0 & 0.70 \end{pmatrix}.$$

Note that adjacent categories are more likely to be misclassified.

Figure 1 shows a randomly chosen data set. In both graphics, the black lines represent the true probabilities. The first graph shows the three shaded areas of the stacked bar chart that represent the empirical probabilities. It is evident that there exists misclassification because the empirical probabilities and the true probabilities are different. The second graph shows the misclassified data. In this graph, there are data whose highest probability are, for example, the category 1 (drawn as empty red dots), but these are classified as category 2 (drawn as filled green dots) or category 3 (drawn as filled blue dots).

When considering measurement error models, it is usually needed to have validation data, prior information or impose some assumptions [4]. This allows to identify the estimates of the misclassification parameters and to achieve convergence of the Markov chain. In this case, informative prior distributions must be considered. Specifically, initial information is introduced according to the misclassification proportions that have been considered for the simulated data, that is, $\boldsymbol{\lambda}_r \sim \text{Dirichlet}(\boldsymbol{a}_r)$ where $\boldsymbol{a}_r$ is the rth column of the matrix of misclassification probabilities multiplied by 10.

The MCMC algorithms have been implemented in R. The standard ordinal models that have been used are the probit and logit models defined by [1] and [15], respectively. A total of 15,000 iterations have been generated for each model. Then, 5,000 iterations were taken as burn-in and one out of 5 values have been saved (thinning equal to 5). With these specifications the chains seem to have converged.

In order to avoid that the results depend on a single simulation, the experiment has been replicated 100 times. The same covariate set, parameters and specifications are used, but data are generated and randomly misclassified at each time (steps (i)-(iv)).

Tables 1 and 2 show the posterior estimations of the regression parameters and the three goodness-of-fit criteria. Note that TVD criterion values for the models considering misclassification are much smaller than the ones for the standard models. This is translated into better estimations, i.e., the estimations from the models addressing misclassification are less biased than the ones from the models that do not consider misclassification.

Table 1. Simulated data ($n = 300$, $J = 3$): Estimated means (SD) for the parameters and the goodness-of-fit criteria.

Dataset	Parameter or Criterion	Upper misclassification Probit	Upper misclassification Probit Mis	Lower misclassification Probit	Lower misclassification Probit Mis
	$\beta_1 = -2$	-1.172 (0.257)	-2.348 (0.529)	-1.893 (0.275)	-2.111 (0.409)
Probit	$\beta_2 = 3$	2.306 (0.317)	3.303 (0.569)	2.289 (0.350)	3.282 (0.527)
Mis	$\beta_3 = 3$	2.367 (0.334)	3.382 (0.602)	2.306 (0.331)	3.341 (0.674)
	$\kappa_2 = 2$	1.596 (0.126)	2.218 (0.467)	1.573 (0.113)	2.262 (0.410)
	DIC	484.908 (20.864)	483.561 (20.878)	502.996 (21.196)	501.767 (20.685)
	TVD	85.385 (13.874)	43.366 (26.940)	95.037 (13.418)	45.146 (21.817)
	S	292.972 (12.566)	288.893 (16.006)	304.018 (12.872)	293.916 (15.209)
Dataset	Parameter or Criterion	Upper misclassification Logit	Upper misclassification Logit Mis	Lower misclassification Logit	Lower misclassification Logit Mis
	$\beta_1 = -2$	-1.193 (0.365)	-2.326 (0.614)	-2.112 (0.402)	-1.884 (0.666)
Logit	$\beta_2 = 3$	2.599 (0.447)	3.266 (0.686)	2.571 (0.437)	3.207 (0.653)
Mis	$\beta_3 = 3$	2.648 (0.564)	3.356 (0.793)	2.576 (0.528)	3.240 (0.792)
	$\kappa_2 = 2$	1.909 (0.172)	2.315 (0.554)	1.887 (0.190)	2.326 (0.798)
	DIC	564.785 (20.120)	562.970 (20.725)	580.099 (18.839)	579.910 (19.116)
	TVD	66.441 (15.456)	53.670 (28.542)	72.603 (14.625)	51.520 (30.778)
	S	335.006 (12.616)	335.348 (13.599)	344.800 (11.811)	345.228 (11.613)

Table 2. Simulated data ($n = 500$, $J = 4$): Estimated means (SD) for the parameters and the goodness-of-fit criteria.

Dataset	Parameter or Criterion	Upper misclassification Probit	Upper misclassification Probit Mis	Lower misclassification Probit	Lower misclassification Probit Mis
	$\beta_1 = -2$	-0.951 (0.285)	-2.280 (0.428)	-1.686 (0.331)	-1.908 (1.131)
Probit	$\beta_2 = 3$	2.180 (0.338)	3.240 (0.436)	1.972 (0.396)	3.190 (0.564)
Mis	$\beta_3 = 3$	2.186 (0.324)	3.281 (0.507)	2.003 (0.347)	3.256 (0.545)
	$\beta_4 = 3$	2.205 (0.315)	3.312 (0.479)	1.998 (0.329)	3.231 (0.535)
	$\kappa_2 = 2$	1.549 (0.141)	2.215 (0.323)	1.375 (0.109)	2.333 (1.111)
	$\kappa_3 = 4$	3.021 (0.184)	4.306 (0.429)	2.810 (0.187)	4.479 (1.248)
	DIC	912.413 (39.208)	903.654 (36.247)	1033.199 (50.375)	1022.597 (46.975)
	TVD	176.448 (16.638)	74.358 (39.463)	216.136 (17.570)	82.400 (37.979)
	S	527.576 (20.241)	514.938 (21.616)	589.383 (25.009)	550.077 (25.601)
Dataset	Parameter or Criterion	Upper misclassification Logit	Upper misclassification Logit Mis	Lower misclassification Logit	Lower misclassification Logit Mis
	$\beta_1 = -2$	-1.055 (0.367)	-2.273 (0.462)	-2.034 (0.371)	-1.995 (0.858)
Logit	$\beta_2 = 3$	2.602 (0.426)	3.244 (0.559)	2.415 (0.428)	3.219 (0.701)
Mis	$\beta_3 = 3$	2.591 (0.432)	3.218 (0.522)	2.435 (0.439)	3.229 (0.663)
	$\beta_4 = 3$	2.612 (0.464)	3.247 (0.633)	2.446 (0.457)	3.250 (0.833)
	$\kappa_2 = 2$	1.895 (0.179)	2.195 (0.365)	1.730 (0.161)	2.167 (0.745)
	$\kappa_3 = 4$	3.681 (0.229)	4.412 (0.683)	3.554 (0.214)	4.348 (0.757)
	DIC	1094.323 (34.386)	1092.900 (33.579)	1190.164 (35.821)	1188.822 (36.688)
	TVD	126.221 (20.768)	82.119 (48.678)	153.598 (19.295)	90.235 (44.171)
	S	604.093 (17.459)	606.154 (17.154)	654.239 (17.485)	643.964 (21.428)

In general, the standard deviations for the regression parameters of the misclassification models are larger than the ones of the standard models (that do not consider misclassification), see e.g., [4]. This happens because there are additional parameters and, therefore, the models addressing misclassifications become more complex to estimate. In spite of their complexity, they perform better. Note that in some cases of this simulation-based example, biases around 10-15% have been obtained for the misclassification models (less than the ones obtained for those models that do not consider misclassification), but it should be noted that the considered misclassification parameters ranged from 0.05 to 0.25, which included a large amount of uncertainty to be addressed.

Tables 3 and 4 show the posterior estimations for the misclassification parameters. The correct information provided for the misclassification parameters allows the proposed approaches to properly recover the misclassification parameters in the considered scenarios.

Table 3. Simulated data ($n = 300$, $J = 3$): Estimated means (SD) for the misclassification parameters.

Parameter	Upper misclassification		Lower misclassification	
	Probit Mis	Logit Mis	Probit Mis	Logit Mis
λ_{11}	0.7356 (0.0585)	0.7394 (0.0579)	0.9992 (0.0011)	1.0000 (0.0000)
λ_{21}	0.2639 (0.0586)	0.2598 (0.0579)	0.0006 (0.0011)	0.0000 (0.0000)
λ_{31}	0.0005 (0.0008)	0.0008 (0.0011)	0.0002 (0.0004)	0.0000 (0.0000)
λ_{12}	0.0011 (0.0037)	0.0019 (0.0074)	0.2647 (0.0597)	0.2891 (0.0670)
λ_{22}	0.7342 (0.0579)	0.7130 (0.0491)	0.7344 (0.0600)	0.7109 (0.0670)
λ_{32}	0.2647 (0.0588)	0.2850 (0.0498)	0.0009 (0.0026)	0.0000 (0.0000)
λ_{13}	0.0003 (0.0011)	0.0010 (0.0044)	0.0006 (0.0014)	0.0000 (0.0001)
λ_{23}	0.0005 (0.0005)	0.0004 (0.0003)	0.2506 (0.0565)	0.2551 (0.0541)
λ_{33}	0.9992 (0.0012)	0.9986 (0.0043)	0.7487 (0.0565)	0.7449 (0.0541)

Table 4. Simulated data ($n = 500$, $J = 4$): Estimated means (SD) for the misclassification parameters.

Parameter	Upper misclassification		Lower misclassification	
	Probit Mis	Logit Mis	Probit Mis	Logit Mis
λ_{11}	0.6972 (0.0571)	0.6983 (0.0587)	0.9998 (0.0005)	0.9999 (0.0002)
λ_{21}	0.2575 (0.0599)	0.2558 (0.0596)	0.0000 (0.0000)	0.0000 (0.0001)
λ_{31}	0.0452 (0.0269)	0.0459 (0.0167)	0.0001 (0.0005)	0.0001 (0.0002)
λ_{41}	0.0001 (0.0002)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0001)
λ_{12}	0.0000 (0.0001)	0.0000 (0.0001)	0.2552 (0.0703)	0.2692 (0.0582)
λ_{22}	0.6784 (0.0702)	0.6760 (0.0582)	0.7448 (0.0704)	0.7307 (0.0582)
λ_{32}	0.2672 (0.0624)	0.2690 (0.0538)	0.0001 (0.0002)	0.0001 (0.0002)
λ_{42}	0.0544 (0.0304)	0.0550 (0.0302)	0.0000 (0.0000)	0.0000 (0.0000)
λ_{13}	0.0000 (0.0000)	0.0000 (0.0000)	0.0493 (0.0264)	0.0519 (0.0277)
λ_{23}	0.0001 (0.0002)	0.0000 (0.0000)	0.2593 (0.0569)	0.2728 (0.0548)
λ_{33}	0.7403 (0.0564)	0.7216 (0.0622)	0.6914 (0.0686)	0.6752 (0.0564)
λ_{43}	0.2596 (0.0564)	0.2784 (0.0622)	0.0000 (0.0001)	0.0001 (0.0006)
λ_{14}	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0002)	0.0002 (0.0017)
λ_{24}	0.0000 (0.0000)	0.0000 (0.0000)	0.0462 (0.0260)	0.0484 (0.0246)
λ_{34}	0.0001 (0.0004)	0.0000 (0.0001)	0.2548 (0.0649)	0.2622 (0.0631)
λ_{44}	0.9999 (0.0004)	1.0000 (0.0001)	0.6990 (0.0679)	0.6892 (0.0646)

The obtained results show that the proposed models considering misclassification perform better than the models that do not consider it in these simulated scenarios. Next, the proposed approach will be applied to data coming from a real application.

6. The analysis of the Signal-Tandmobiel® data

The proposed methodology uses the misclassification information provided by the validation dataset and relates it to the main dataset from the study to arrive at a posterior predictive distribution that is used to estimate probabilities of the levels of CE degree. The interest of the present analysis is to evaluate the misclassification probabilities of the levels of CE degree, and to address the influence of oral hygiene and geographical information on the levels of CE degree.

The ordinal outcome y is the level of CE degree. The covariates considered in the model were the following: gender, age, frequency of brushing, plaque index proscimal surfaces, plaque index occlusal surfaces, and geographical location (represented by the standardized (x, y) coordinate of the municipality of the school to which the child belongs).

In order to illustrate de applicability of the proposed methods, three different models for both probit and logit link functions have been considered for the main dataset. The first models are the ordinal probit and logit regression ones (Probit-Standard and Logistic-Standard), i.e. the standard models without considering misclassification. In the second models (Probit-Validation and Logistic-Validation), the validation dataset has been used to estimate the misclassification probabilities $\boldsymbol{\lambda}$, and afterwards, the regression parameters $\boldsymbol{\beta}$ and the cutpoints $\boldsymbol{\kappa}$ of the ordinal regression models have been estimated for the main dataset. The regression parameters and cutpoints are estimated by using the full conditional distributions of the algorithm proposed in Section 4.2. Specifically, the algorithms consist of choosing initial values $\boldsymbol{z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\kappa}^{(0)}$ and $\boldsymbol{c}^{(0)}$, and iteratively sampling $\boldsymbol{z}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\kappa}^{(t)}$ and $\boldsymbol{c}^{(t)}$ from the following full conditional distributions. For the probit link they are sampled from the algorithm defined in Appendix A.1, whereas for the logit link, they are sampled from the algorithm defined in Appendix A.2. Note that, in these cases, the matrix of misclassification probabilities remain fixed as the estimations obtained from the validation dataset. Finally, the third models (Probit-Misclassification and Logistic-Misclassification) are the algorithms proposed in Appendices A.1 and A.2. In these cases, the validation dataset has been used to elicit the prior distribution for the misclassification parameters $\boldsymbol{\lambda}$, and then the algorithms are applied to the main dataset.

Note that the validation dataset has been used in two different ways. In the second models, the validation dataset is used to compute the misclassification probabilities. This is common when there exists a validation dataset, because the scores from the examiners and from the benchmark are available. Therefore, the misclassification probabilities can be estimated from the validation dataset and the regression parameters can be estimated by using a simplified version of the MCMC method that uses the estimated misclassification probabilities as fixed values. In the third models, the validation dataset has been used to construct a prior distribution for the misclassification parameters. However, historical data and/or experts' information can be also considered to elicit the prior distribution for the misclassification parameters in many contexts.

The way how the validation dataset has been used is as follows. Let y^{exa} and y^{ben} be the scores of the examiners and the benchmark in the validation dataset, respectively. The hierarchical model $y^{exa}|y^{ben} = r \sim \text{Multinomial}(1, \boldsymbol{\lambda}_r)$, $\boldsymbol{\lambda}_r \sim \text{Dirichlet}(\boldsymbol{a}_r)$, and $a_{rs} \sim \text{Gamma}(0.01, 0.01)$, for $r, s = 1, \ldots, 4$, allows to estimate the posterior distributions of the misclassification probabilities in the validation dataset. Then, the posterior estimations (mean and standard deviation) are given by

$$\widehat{\boldsymbol{\lambda}} = \begin{pmatrix} 0.880\ (0.019) & 0.198\ (0.039) & 0.060\ (0.038) & 0 \\ 0.120\ (0.019) & 0.702\ (0.045) & 0.249\ (0.074) & 0 \\ 0 & 0.099\ (0.029) & 0.691\ (0.074) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

and the marginal posterior distributions are

$$\begin{aligned} \boldsymbol{\lambda}_1 &\sim \text{Dirichlet}(25.198,\ 3.584,\ 0,\ 0), \\ \boldsymbol{\lambda}_2 &\sim \text{Dirichlet}(6.846,\ 23.868,\ 3.511,\ 0), \\ \boldsymbol{\lambda}_3 &\sim \text{Dirichlet}(2.187,\ 8.954,\ 25.129,\ 0), \\ \boldsymbol{\lambda}_4 &\sim \text{Dirichlet}(0,\ 0,\ 0,\ 1). \end{aligned} \tag{4}$$

From the estimation of the misclassification probabilities it is evident that adjacent categories are related, in the sense that the probability of misclassification in an adjacent category is higher than the one of misclassification in a non adjacent category. The misclassification probabilities obtained from the validation dataset can be used to correct for misclassification. For the Probit-Validation and Logistic-Validation models, the estimated misclassification probabilities $\widehat{\boldsymbol{\lambda}}$ are given in (3). For the Probit-Misclassification and Logistic-Misclassification models, the distributions (4) are used as the prior distributions of $\boldsymbol{\lambda}$.

The estimated parameters obtained with the probit and logit models are summarized in Table 5. The posterior means, standard deviations (SD), and the 95% highest posterior density (HPD) intervals are represented.

Table 5. Summary of the posterior estimates for the parameters of the ST data by using several methods.

Parameter	**Mean (SD)**	**95% HDP interval**	**Mean (SD)**	**95% HDP interval**
	Probit-Standard		**Logistic-Standard**	
Intercept	-2.758 (0.737)	(-4.209,-1.290)	-4.501 (1.380)	(-7.074,-1.695)
Gender (girl)	0.201 (0.045)	(0.111,0.293)	0.316 (0.077)	(0.161,0.466)
Age	0.219 (0.062)	(-0.101,0.341)	0.369 (0.111)	(0.134,0.565)
Brushing	-0.093 (0.016)	(-0.126,-0.062)	-0.151 (0.028)	(-0.203,-0.092)
Proscimal	0.276 (0.047)	(0.189,0.370)	0.469 (0.077)	(0.314,0.621)
Occlusal	0.364 (0.112)	(0.141,0.587)	0.585 (0.204)	(0.225,1.012)
x-coordinate	0.002 (0.001)	(0.001,0.003)	0.003 (0.001)	(0.002,0.005)
y-coordinate	-0.001 (0.001)	(-0.003,0.002)	-0.001 (0.002)	(-0.005,0.003)
κ_2	0.683 (0.024)	(0.638,0.732)	1.149 (0.043)	(1.065,1.226)
κ_3	2.061 (0.072)	(1.930,2.214)	4.046 (0.173)	(3.720,4.383)
	Probit-Validation		**Logistic-Validation**	
Intercept	-3.507 (0.902)	(-5.299,-1.713)	-5.873 (1.536)	(-8.923,-2.922)
Gender (girl)	0.260 (0.057)	(0.148,0.371)	0.438 (0.099)	(0.238,0.631)
Age	0.268 (0.076)	(0.118,0.417)	0.458 (0.130)	(0.216,0.720)
Brushing	-0.115 (0.020)	(-0.154,-0.075)	-0.196 (0.034)	(-0.262,-0.128)
Proscimal	0.322 (0.057)	(0.207,0.427)	0.568 (0.101)	(0.378,0.769)
Occlusal	0.468 (0.134)	(0.202,0.725)	0.794 (0.226)	(0.351,1.231)
x-coordinate	0.003 (0.001)	(0.002,0.004)	0.005 (0.001)	(0.003,0.006)
y-coordinate	-0.001 (0.002)	(-0.003,0.003)	-0.001 (0.003)	(-0.006,0.004)
κ_2	0.375 (0.045)	(0.286,0.467)	0.624 (0.076)	(0.476,0.771)
κ_3	1.997 (0.073)	(1.857,2.142)	3.949 (0.181)	(3.603,4.328)
	Probit-Misclassification		**Logistic-Misclassification**	
Intercept	-3.458 (0.969)	(-5.431,-1.605)	-6.257 (1.730)	(-9.610,-2.919)
Gender (girl)	0.259 (0.058)	(0.150,0.374)	0.464 (0.099)	(0.278,0.664)
Age	0.269 (0.079)	(0.116,0.431)	0.487 (0.141)	(0.213,0.757)
Brushing	-0.114 (0.020)	(-0.153,-0.075)	-0.201 (0.036)	(-0.270,-0.130)
Proscimal	0.327 (0.059)	(0.211,0.444)	0.566 (0.108)	(0.359,0.776)
Occlusal	0.460 (0.139)	(0.186,0.731)	0.849 (0.244)	(0.366,1.326)
x-coordinate	0.003 (0.001)	(0.001,0.004)	0.005 (0.001)	(0.003,0.007)
y-coordinate	-0.001 (0.002)	(-0.003,0.003)	-0.001 (0.003)	(-0.007,0.004)
κ_2	0.384 (0.197)	(0.033,0.807)	0.470 (0.284)	(0.038,1.067)
κ_3	2.029 (0.136)	(1.761,2.298)	3.880 (0.227)	(3.434,4.341)

The standard deviations estimated with the models considering misclassification are larger than those obtained with the standard models due to the inclusion of more parameters. Therefore, the 95% credible intervals for the models considering misclassification are wider than the ones that do not consider it. Also, as expected, the estimations from misclassification-based models (Logistic-Misclassification and Probit-Misclassification) are closer to the ones from the validation-based models (Logistic-Validation and Probit-Validation) than they are to the standard models (Logistic-Standard and Probit-Standard). Once the validation dataset has shown that the study has misclassification errors, the standard models should not be used. The purpose of including the parameter estimation of the standard models in the experimental results is purely illustrative, i.e., to show that the regression parameters go far from the ones corrected by misclassification.

Some model selection criteria have been used in real data application for misclassification (see, e.g., [16], [31] and [34]). DIC and S criteria have been used in this application for the four models that consider misclassifications. TVD criterion cannot be applied in this context, since the real misclassification probabilities are not known as it happened in the simulation-based example. Standard models have been discarded from this comparison, since misclassifications have been proved to exist and no correction has been performed with those models.

Table 6 presents the values of DIC and S criteria for these models considering misclassification. When comparing logit and probit models, DICs show that models based on logit link provide similar results than the probit-based models, whereas the logit-based models are preferred based on the S criteria.

Table 6. DIC and S values for the estimated models considering misclassification.

	Probit-Validation	**Logistic-Validation**
DIC	5473.849	5474.976
S	3004.884	2997.624
	Probit-Misclassification	**Logistic-Misclassification**
DIC	5474.721	5475.128
S	3032.063	2969.581

Finally, the practical conclusions obtained from regression parameters is presented. Positive regression coefficients reflect higher probabilities of CE lesion severity compared to the reference level for categorical covariates. For the variable gender, the category of boys was taken as the reference. The girls have higher probability of having CE than the boys. The reason is that the permanent teeth emerge earlier with girls than with boys, and hence teeth of girls are longer at risk at the same age as those of boys. The probability of CE lesion severity increases as the age of children increases, which is a biologically expected result due to the fact that CE is a progressive illness. The regression coefficient of brushing frequency is negative, indicating that the brushing frequency is a protection factor against CE. The regression coefficients of plaque index on both proximal and occlusal surfaces are positive, indicating that high values of the corresponding covariates are associated with high probabilities of having high levels of CE lesion. Moreover, there was a significant effect of the x-coordinate, but not of the y-coordinate of the school geographical location. These results indicating which covariables are protective or risk factors matches with the one obtained by [32] and [20] for different models with the same data.

7. Conclusion

A Bayesian approach to polychotomous response data that are subject to misclassification has been proposed and discussed in this paper. The idea of using a data augmentation framework has been exploited to derive MCMC algorithms. This model has been explored for ordered categories in the response variable by using both probit and logit link functions. Besides, the proposed approach can be extended using other link functions.

The applicability of the proposed approach has been illustrated through a simulated example that shows their good performance when compared with models that do not consider misclassification. A longitudinal oral health study conducted in Flanders (Belgium), the Signal-Tandmobiel® study, has been analyzed. The main advantage of the proposed model is provided better estimations than the standard ones. Through the simulated example we have shown that, when data are misclassified, the estimates from models that do not consider misclassification are biased, and that the estimates from models considering misclassification are closer to the real ones. Moreover, by using latent variables and considering prior information it is possible to update misclassification probabilities. Therefore, when ordinal data are subjected to misclassification, it is highly recommended to consider model that take into account this fact.

Although the approach has been explored in the case of ordered categories, it also can be adapted for unordered categories as follows. The data augmentation scheme provided in Section 3 is firstly considered. The latent variables $\boldsymbol{c}$ and the misclassification probabilities $\boldsymbol{\lambda}$ are introduced in the nominal response data model. The Gibbs sampling described in Section 4.2 is used to sample $\boldsymbol{c}$ and $\boldsymbol{\lambda}$ from the full conditional posterior distributions (1) and (2), respectively. Then, the latent vector $\boldsymbol{v}$ is obtained from $\boldsymbol{c}$, which correspond to the true classifications, where $v_i = r$ if $c^i_{+r} = 1$. Finally, the outcomes $\boldsymbol{Y}$ are replaced by the latent vector $\boldsymbol{v}$ to obtain the probabilities $\boldsymbol{p}$. The regression parameters are estimated by using other algorithms (see, for example, [17]). Moreover, the approach can also be extended to other link functions. The potential applicability of this approach to many fields of knowledge makes this proposal appealing.

Acknowledgements

The Signal-Tandmobiel® study comprises the following partners: D. Declerck (Dental School, Katholieke Universiteit Leuven), L. Martens (Dental School, University of Ghent), J. Vanobbergen (Dental School, University of Ghent), P. Bottenberg (Dental School, University of Brussels), E. Lesaffre (L-BioStat, Katholieke Universiteit Leuven), and K. Hoppenbrouwers (Youth Health Department, Katholieke Universiteit Leuven, and Flemish Association for Youth Health Care).

Funding

This research has been supported by *Ministerio de Economía y Competitividad*, Spain (Projects MTM2014-56949-C3-3-R and MTM2017-86875-C3-2-R), *Gobierno de Extremadura*, Spain (Project GRU18108), and *European Union* (European Regional Development Funds). First author has also been supported by *Sociedad Matemática Mexicana* and *Fundación Sofía Kovalévskaia*.

References

[1] J. Albert and S. Chib, *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association 88 (1993), pp. 669–679.

[2] P.S. Albert, S.A. Hunsberger, and F.M. Biro, *Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation*, Journal of the American Statistical Association 92 (1997), pp. 1304–1311.

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

[4] J.P. Buonaccorsi, *Measurement Error*, Chapman & Hall/CRC, London, 2010.

[5] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed., Chapman & Hall/CRC, Boca Raton, Florida, 2006.

[6] M.H. Chen, J.G. Ibrahim, and Q.M. Shao, *Power prior distributions for generalized linear models*, Journal of Statistical Planning and Inference 84 (2000), pp. 121–137.

[7] P. Congdon, *Bayesian Models for Categorical Data*, John Wiley & Sons, Chichester, 2005.

[8] P. Congdon, *Bayesian Statistical Modelling*, 2nd ed., John Wiley & Sons, Chichester, 2006.

[9] M.K. Cowles, *Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models*, Statistics and Computing 6 (1996), pp. 101–111.

[10] C. Czado, A. Heyn, and G. Müller, *Modeling individual migraine severity with autoregressive ordered probit models*, Statistical Methods and Applications 20 (2011), pp. 101–121.

[11] K.T. Fang, Z. Geng, and G.L. Tian, *Statistical inference for truncated Dirichlet distribution and its application in misclassification*, Biometrical Journal 42 (2000), pp. 1053–1068.

[12] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.

[13] P. Gustafson, *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*, Chapman and Hall, Boca Raton, Florida, 2003.

[14] ICDAS, *Criteria Manual. International Caries Detection and Assessment System (ICDAS II)*, International Caries Detection and Assessment System (ICDAS) Coordinating Committee, 2005.

[15] V.E. Johnson and J.H. Albert, *Ordinal Data Modeling*, Springer, New York, 1999.

[16] S. Luo, X. Su, S.M. DeSantis, X. Huang, M. Yi, and K.K. Hunt, *Joint model for a diagnostic test without a gold standard in the presence of a dependent terminal event*, Statistics in Medicine 33 (2014), pp. 2554–2566.

[17] R.E. McCulloch, N.G. Polson, and P.E. Rossi, *A Bayesian analysis of the multinomial probit model with fully identified parameters*, Journal of Econometrics 99 (2000), pp. 173–193.

[18] A. McGlothlin, J.D. Stamey, and J.W.J. Seaman, *Binary regression with misclassified response and covariate subject to measurement error: a Bayesian approach*, Biometrical Journal 50 (2008), pp. 123–134.

[19] P. McInturff, W.O. Johnson, D. Cowling, and I.A. Gardner, *Modelling risk when binary outcomes are subject to error*, Statistics in Medicine 23 (2004), pp. 1095–1109.

[20] S.M. Mwalili, E. Lesaffre, and D. Declerck, *A Bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study*, Journal of the Royal Statistical Society: Series C, Applied Statistics 54 (2005), pp. 77–93.

[21] L. Naranjo, J. Martín, C.J. Pérez, and M.J. Rufo, *Addressing misclassification for binary data: probit and t-link regressions*, Journal of Statistical Computation and Simulation 84 (2014), pp. 2187–2213.

[22] C.D. Paulino, G. Silva, and J.A. Achcar, *Bayesian analysis of correlated misclassified binary data*, Computational Statistics and Data Analysis 49 (2005), pp. 1120–1131.

[23] C.D. Paulino, P. Soares, and J. Neuhaus, *Binomial regression with misclassification*, Biometrics 59 (2003), pp. 670–675.

[24] N.B. Pitts, D.J. Evans, and C.M. Pine, *British association for the study of community dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97*, Community Dent Health 14 (1997), pp. 6–9.

[25] W.Y. Poon and H. Wang, *Bayesian analysis of multivariate probit models with surrogate outcome data*, Psychometrika 75 (2010), pp. 498–520.

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

[27] S. Roy, *Analysis of ordered probit model with surrogate response data and measurement error in covariates*, Communications in Statistics - Theory and Methods 45 (2016), pp. 2665–2678.

[28] S. Roy and T. Banerjee, *Analysis of misclassified correlated binary data using a multivariate probit model when the covariates are subject to measurement error*, Biometrical Journal 51 (2009), pp. 420–432.

[29] D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde, *Bayesian measures of model complexity and fit*, Journal of the Royal Statistical Society, Series B 64 (2002), pp. 583–639.
[30] T. Swartz, Y. Haitovsky, A. Vexler, and T. Yang, *Bayesian identifiability and misclassification in multinomial data*, The Canadian Journal of Statistics 32 (2004), pp. 285–302.
[31] L. Tang, R.H. Lyles, C.C. King, D.D. Celentano, and Y. Lo, *Binary regression with differentially misclassified response and exposure variables*, Statistics in Medicine 34 (2015), pp. 1605–1620.
[32] J. Vanobbergen, L. Martens, E. Lesaffre, and D. Declerck, *The Signal-Tandmobiel® project, a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results*, European Journal of Paediatric Dentistry 2 (2000), pp. 87–96.
[33] S. Wacholder, B. Armstrong, and P. Hartge, *Validation studies using an alloyed gold standard*, American Journal of Epidemiology 137 (1993), pp. 1251–1258.
[34] J. Zhang, S.R. Cole, D.B. Richardson, and H. Chu, *A bayesian approach to strengthen inference for case-control studies with multiple error-prone exposure assessments*, Statistics in Medicine 32 (2013), pp. 4426–4437.

Appendix A. Steps within the Gibbs sampling

In order to generate $\boldsymbol{\kappa}^{(t)}$, $\boldsymbol{z}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ at the t-th iteration of the Gibbs sampler defined in Section 4.2, the Metropolis-Hastings-within-Gibbs algorithm can be implemented as the following.

A.1 *Ordinal probit model*

Let $\Psi = \Phi$ be the cdf of a standard normal distribution, $\mathrm{N}(0,1)$. The Metropolis-Hastings-within-Gibbs algorithm is as follows:

(1a) Generate a candidate κ_r^{new} for $r = 2, \dots, J-1$, from a truncated normal distribution $\mathrm{N}\left(\kappa_r^{(t-1)}, \sigma_\kappa^2\right) I\left[\kappa_{r-1}^{(t)} < \kappa_r^{new} < \kappa_{r+1}^{(t-1)}\right]$, where σ_κ^2 is a value chosen to obtain an appropriate acceptance rate, e.g. $\sigma_\kappa = 0.4$.

(1b) Evaluate the acceptance probability for the vector of new cutpoints as $\alpha = \min(1, R)$ where

$$R = \prod_{r=2}^{J-1} \frac{\Phi\left((\kappa_{r+1}^{(t-1)} - \kappa_r^{(t-1)})/\sigma_\kappa\right) - \Phi\left((\kappa_{r-1}^{new} - \kappa_r^{(t-1)})/\sigma_\kappa\right)}{\Phi\left((\kappa_{r+1}^{new} - \kappa_r^{new})/\sigma_\kappa\right) - \Phi\left((\kappa_{r-1}^{(t-1)} - \kappa_r^{new})/\sigma_\kappa\right)}$$

$$\times \frac{\Phi\left(\kappa_{v_i}^{new} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right) - \Phi\left(\kappa_{v_i-1}^{new} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right)}{\Phi\left(\kappa_{v_i}^{(t-1)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right) - \Phi\left(\kappa_{v_i-1}^{(t-1)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right)}.$$

Note that $v_i = r$ is equivalent to $c_{+r}^i = 1$.

(1c) With probability α, set $\boldsymbol{\kappa}^{(t)} = \boldsymbol{\kappa}^{new}$ and generate $\boldsymbol{z}^{(t)} = (z_1^{(t)}, \dots, z_n^{(t)})$ from the truncated normal distribution

$$\mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\ ,\ 1\right) \times \sum_{r=1}^{J} I\left[\kappa_{r-1}^{(t)} < z_i < \kappa_r^{(t)}\right] I\left[c_{+r}^i = 1\right].$$

Otherwise, set $\boldsymbol{\kappa}^{(t)} = \boldsymbol{\kappa}^{(t-1)}$ and $\boldsymbol{z}^{(t)} = \boldsymbol{z}^{(t-1)}$.

(1d) Generate $\boldsymbol{\beta}^{(t)}$ from $\mathrm{N}_k\left((\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{z}\ ,\ (\boldsymbol{x}'\boldsymbol{x})^{-1}\right)$.

The final algorithm consists of choosing initial values $\boldsymbol{\kappa}^{(0)}$, $\boldsymbol{z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{c}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, and iteratively sampling $\boldsymbol{\kappa}^{(t)}$, $\boldsymbol{z}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{c}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ from the algorithm described in (1a)-(1d) and the full conditional distributions (1) and (2).

A.2 *Ordinal logit model*

Let Ψ be the cdf of a standard logistic distribution, $\mathrm{L}(0,1)$. The Metropolis-Hastings-within-Gibbs algorithm is as follows:

(2a) Generate a candidate κ_r^{new} for $r = 2, \ldots, J-1$, from a truncated logistic distribution $\mathrm{L}\left(\kappa_r^{(t-1)}, \sigma_\kappa\right) I\left[\kappa_{r-1}^{(t)} < \kappa_r^{new} < \kappa_{r+1}^{(t-1)}\right]$.

(2b) Evaluate the acceptance probability for the vector of new cutpoints as $\alpha_\kappa = \min(1, R_\kappa)$ where

$$R_\kappa = \prod_{r=2}^{J-1} \frac{\Psi\left((\kappa_{r+1}^{(t-1)} - \kappa_r^{(t-1)})/\sigma_\kappa\right) - \Psi\left((\kappa_{r-1}^{new} - \kappa_r^{(t-1)})/\sigma_\kappa\right)}{\Psi\left((\kappa_{r+1}^{new} - \kappa_r^{new})/\sigma_\kappa\right) - \Psi\left((\kappa_{r-1}^{(t-1)} - \kappa_r^{new})/\sigma_\kappa\right)}$$
$$\times \frac{\Psi\left(\kappa_{v_i}^{new} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right) - \Psi\left(\kappa_{v_i-1}^{new} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right)}{\Psi\left(\kappa_{v_i}^{(t-1)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right) - \Psi\left(\kappa_{v_i-1}^{(t-1)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right)}.$$

(2c) With probability α_κ, set $\boldsymbol{\kappa}^{(t)} = \boldsymbol{\kappa}^{new}$ and generate $\boldsymbol{z}^{(t)} = (z_1^{(t)}, \ldots, z_n^{(t)})$ from the truncated logistic distribution

$$\mathrm{L}\left(\boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\ ,\ 1\right) \times \sum_{r=1}^{J} I\left[\kappa_{r-1}^{(t)} < z_i < \kappa_r^{(t)}\right] I\left[c_{+r}^i = 1\right].$$

Otherwise, set $\boldsymbol{\kappa}^{(t)} = \boldsymbol{\kappa}^{(t-1)}$ and $\boldsymbol{z}^{(t)} = \boldsymbol{z}^{(t-1)}$.

(2d) Generate a candidate $\boldsymbol{\beta}^{new}$ from a normal distribution $\mathrm{N}_k\left(\boldsymbol{\beta}^{(t-1)}, (\boldsymbol{x}'\boldsymbol{x})^{-1}\right)$.

(2e) Evaluate the acceptance probability for the vector of regression parameters $\alpha_\beta = \min(1, R_\beta)$ where

$$R_\beta = \prod_{i}^{n} \frac{\left(z_i^{(t)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{new}\right)}{\left(z_i^{(t)} - \boldsymbol{x}_i'\boldsymbol{\beta}^{(t-1)}\right)}.$$

(2f) With probability α_β, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{new}$. Otherwise, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$.

The final algorithm consists of choosing initial values $\boldsymbol{\kappa}^{(0)}$, $\boldsymbol{z}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{c}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, and iteratively sampling $\boldsymbol{\kappa}^{(t)}$, $\boldsymbol{z}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{c}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$ from the algorithm described in (2a)-(2f) and the full conditional distributions (1) and (2).