TECHNICAL

REPORT

0462

The Zero-Inflated Negative Binomial Regression Model With Correction for Misclassification: An Example in Caries Research

Samuel M. Mwalili, Emmanuel Lesaffre and Dominique Declerck

IAP STATISTICS

NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Abstract: The dmft-index is a popular measure in caries research, but its distribution is complex. Lewsey and Thomson[1] show that the zero-inflated negative binomial (ZINB) distribution gives an excellent fit. We fitted this distribution to the dmft-index of 4468 seven-year old children from the Signal-Tandmobiel® study. However, several dental examiners were involved in scoring caries experience. This necessitated to correct for possible misclassification. We illustrate how a non-differential misclassification process for each examiner can lead to differential misclassification overall.

1 Introduction

The dmft-index is a commonly used index in oral health studies. It measures the degree of caries experience of a subject in the primary dentition and is defined as the sum of the number of decayed (d), missing due to caries (m) and filled (f) deciduous teeth. The dmft-index ranges from 0 (no caries experience) to 20 (all teeth affected). In this paper we will look at the dmft-index of 4468 Flemish children from the Signal-Tandmobiel® (STM) study.[2]

In a previous paper,[3] an ordinal logistic regression model on a categorized dmft-index of the first year's STM data revealed an East-West gradient for caries experience in Flanders. However, the dental examiners operated in different geographical areas. Further, from validation data obtained from the calibration exercises, it was observed that some examiners over-(under) scored caries experience compared to the benchmark examiner (third author). Therefore, it was questioned whether the observed East-West gradient was a result of misclassification of caries experience of the sixteen dental examiners vis-a-vis the benchmark examiner. Consequently, a correction to the logistic model for examiners' misclassification was applied, assuming that the misclassification is non-differential (conditional on the true response the distribution of the misclassified response is independent of the covariates) for each examiner. After correction, the East-West gradient remained significant.

In this paper, however, our attention is focused on fitting the original dmft-index, i.e. the observed counts, since the dentists are more comfortable with the classical score. Correcting for misclassification of the dmft-index presented extra problems as not all levels of the dmft-index were observed by the benchmark examiner in the calibration exercise. Also, the fact that multiple examiners were involved complicated the correction for misclassification. Furthermore, we show how differential misclassification can arise from non-differential misclassification when multiple examiners are involved.

The distribution of the dmft-index is overdispersed with respect to a Poisson distribution, namely $\mathrm{SD}^2(\mathrm{dmft})/\overline{\mathrm{dmft}} = 3.53$, and has an excess of zeroes, implying that a "zero-inflated" model is a good candidate for this kind of data. Böhning *et al.*[4] suggested the zero-inflated Poisson (ZIP) distribution[5] to model the DMFT-index and concluded that it gives a reasonable fit to the observed distribution. The ZIP distribution is a mixture distribution assigning a mass of p to "extra" zeroes and a mass of $(1-p)$ to a Poisson distribution. When covariates are involved, the ZIP distribution gives rise to a ZIP regression model. A useful feature of the ZIP regression model is that the effect of the covariates can be assessed simultaneously in the extra zeroes and the Poisson component of the model. However, a ZIP model is not appropriate when the non-zero part of the distribution is overdispersed with respect to a Poisson distribution. Lewsey and Thomson[1] instead suggested to replace the Poisson part by a negative binomial distribution to fit caries experience data giving rise to the zero-inflated negative binomial (ZINB) distribution. The negative binomial part assumes that the Poisson mean is a random variable following a gamma distribution. In this paper we will focus on the ZINB regression model.

In Section 2, we describe the distribution of the observed dmft-index of the first year's data of the STM study. The ZINB regression model is described in Section 3 and applied to the dmft-index of the STM study. In Section 4 we describe how to correct for misclassification in the case of count responses and multiple examiners. In Section 5 we present the ZINB regression model corrected for misclassification and apply our approach to the STM study. A discussion of our results is given in Section 6.

2 The Signal-Tandmobiel® study

The STM study involves a sample of 4468 children, representative for Flemish children (7% of children born in 1989); The children were first examined in 1996 and annually thereafter for 6 years. Here we have taken the data of the first year of the study, hence the data of seven-year old children are examined in this paper. However, due to the practical organisation of sampling, the age of the children actually varied from 6.12 years to 8.09 years. Besides the oral health information, data were collected on dietary habits and oral hygiene behaviour. For a more detailed description of the STM study we refer to Vanobbergen *et al.*[2]

As can be seen in Figure 1, the distribution of the dmft-index is markedly skewed, with the majority of the children having a low score for caries experience and a minority with a high score. About 44% ($n = 1913$) of 7-year-old children presented without any sign of caries experience. Further, from Figure 1 it is clear that the estimated Poisson distribution does not fit the observed distribution of the dmft-index well.

We are interested in establishing the determinants of caries experience. We considered: age (years), gender (girl = 1), the geographical location (in terms of the $x-$ and $y-$ coordinates) of the school that the child attends, age at start of brushing (years), use of systematic fluoride supplements (regular use = 1), daily consumption of sugar containing drinks between meals (yes = 1), intake of in-between-meals (= 1 if greater than 2, 0 otherwise), and frequency of brushing (= 1 if less than twice a day, 0 otherwise). Vanobbergen *et al.*[6] considered a logistic regression of the dichotomised (dmft = 0 versus dmft $\neq$ 0) response on a similar set of risk indicators, but instead of the $x-$ and $y-$ coordinates they used 4 dichotomous variables to indicate the five Flemish provinces to which the school of the child belongs.

Table 1 shows the results of fitting a Poisson regression model to the dmft-index. The effect of the covariates on the degree of caries experience are the same as those found in the

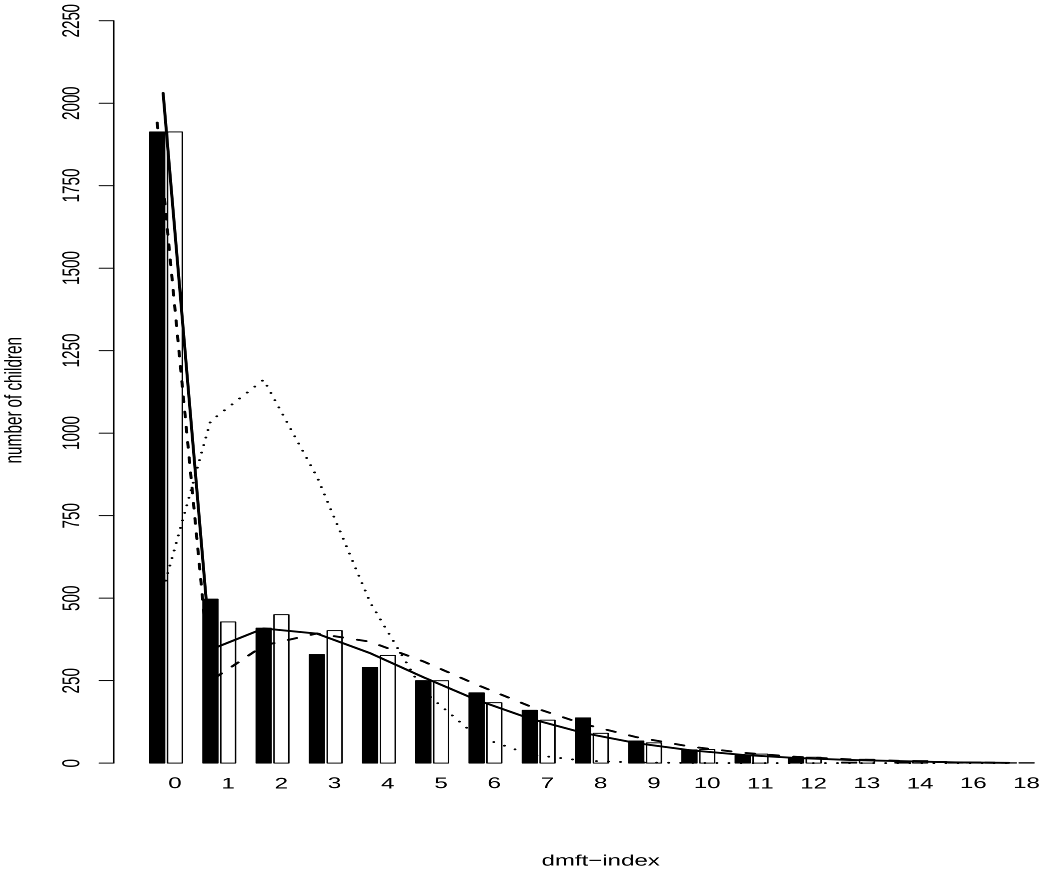

FIGURE 1. Signal-Tandmobiel® study: Distribution of the dmft-index among 7-year old Flemish children, ■ observed, □ fitted from ZINB model; the dotted line shows the fit of the Poisson model, the broken line shows the fit of the pooled corrected ZINB model and the solid line shows the fit of the examiner-specific corrected ZINB model.

logistic model of Vanobbergen *et al.*[6] All variables except for gender affect the degree of caries experience. More importantly, a significant East-West gradient in caries experience is observed, with higher levels in Limburg (Eastern province of Flanders).

3 The zero-inflated negative binomial model applied to the Signal-Tandmobiel® study

The ZINB distribution is a mixture distribution assigning a mass of p to "extra" zeroes and a mass of $(1 - p)$ to a negative binomial distribution, where $0 \leq p \leq 1$. Note that the negative binomial distribution is a continuous mixture of Poisson distributions, which allows the Poisson mean λ to be gamma distributed. More specifically, the negative binomial

TABLE 1. Signal-Tandmobiel® study: Maximum likelihood estimates of the multiple Poisson regression model fitted to the dmft-index.

Parameter	Estimate(SE)	95% CI	
Intercept	0.344(0.040)	0.265	0.423
x-ordinate	0.134(0.012)	0.110	0.157
y-ordinate	−0.034(0.012)	−0.059	−0.010
Gender (girl)	0.035(0.024)	−0.011	0.082
Age (years)	0.188(0.029)	0.130	0.246
Brushing frequency (< 2)	0.073(0.033)	0.009	0.138
Age start brushing (years)	0.102(0.011)	0.081	0.123
Fluoride supplement (yes)	−0.258(0.025)	−0.306	−0.210
Sugary drinks (yes)	0.292(0.026)	0.242	0.342
Between meals (> 2)	0.115(0.025)	0.066	0.164

distribution is given by

$$Pr(Y=y)=\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}\left(\frac{\tau}{\lambda+\tau}\right)^{\tau}\left(\frac{\lambda}{\lambda+\tau}\right)^{y},\ y=0,1,\cdots;\ \lambda,\tau>0, \tag{1}$$

where $\lambda = E(Y)$, τ is a shape parameter which quantifies the amount of overdispersion, and Y is the response variable of interest. The variance of Y is $\lambda + \lambda^2/\tau$. Note that, a negative binomial distribution approaches a Poisson distribution when τ tends to ∞ (no overdispersion). A ZINB distribution arises as a mixture of a negative binomial and a distribution degenerated at zero, and is given by

$$\Pr(Y=y)\begin{cases} p+(1-p)(1+\lambda/\tau)^{-\tau}, & y=0,\\ (1-p)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}(1+\lambda/\tau)^{-\tau}(1+\tau/\lambda)^{-y} & y=1,2,\cdots. \end{cases} \tag{2}$$

The mean and variance of the ZINB distribution are $\mathrm{E}(Y) = (1-p)\lambda$ and $\mathrm{var}(Y) = (1-p)\lambda(1+p\lambda+\lambda/\tau)$, respectively. Observe that this distribution approaches the ZIP and the negative binomial distribution as $\tau \to \infty$ and $p \to 0$, respectively. If both $1/\tau$ and $p \approx 0$ then the ZINB distribution reduces to the Poisson distribution.

The group of caries-free children, i.e. with dmft = 0, can be thought of as consisting of two subgroups. The first subgroup corresponds to children who (practically) cannot have decayed teeth. The assumption here is that some characteristics (genetic, social, environmental etc) of the child protect it from having caries. The second subgroup consists of children prone to caries development. Note that this is only a convenient explanation to justify the use of the ZINB model. Indeed, the ZINB distribution can also arise from Bernoulli trials with non-equal success probabilities.[7,8] The overdispersed data are characterized by “excess zeroes”, “excess large outcomes”, or both. The ZINB model therefore accounts for these “excess zeroes” and also for the extra heterogeneity in the positive outcome.

The ZINB regression model relates p and λ to covariates, i.e.

$$\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \text{ and } \operatorname{logit}(p_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}, \ (i = 1, \cdots, n) \tag{3}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are d- and q-dimensional vectors of covariates pertaining to the ith child, and with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the corresponding vector of regression coefficients, respectively. The ZINB (minus) log-likelihood given the observed data is

$$\begin{aligned}\mathcal{L}_z(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) \ = \ & \sum_{i=1}^{n} \left(1 + e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}\right) + \\ & \sum_{i:y_i=0} \log\left(e^{\boldsymbol{z}_i'\boldsymbol{\gamma}} + \left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + \tau}{\tau}\right)^{-\tau}\right) + \\ & \sum_{i:y_i>0} \left(\tau \, \log(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + \tau}{\tau}) + y_i \, \log(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}} \tau)\right) + \\ & \sum_{i:y_i>0} \Big(\log\Gamma(\tau) + \log\Gamma(1 + y_i) - \log\Gamma(\tau + y_i)\Big),\end{aligned} \tag{4}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$.

Parameter estimation can be carried out by the BFGS algorithm as described in Nocedal and Wright (pp. 193–201).[9] This technique is a quasi-Newton optimization method implemented in the *optim* *R*-software package. The optimization also requires the first derivatives of the ZINB model (Appendix A.1).

The ZINB model was fitted to the STM study without covariates. Figure 1 shows clearly that the ZINB distribution fits the distribution of the observed dmft-index nearly perfectly.

The results of fitting the ZINB regression model to the STM study are shown in Table 2. The East-West gradient ($x-$ordinate) is significant in both parts of the ZINB regression model, implying that the degree of caries experience increases from West to East of Flanders while the excess of the caries-free children decreases. The consumption of sugar containing drinks is also significant in both parts of the model, implying that the degree of caries experience is higher while the excess of caries-free children is lower for the children who took sugary drinks between meals.

Except for gender and the $y-$ordinate, the other covariates are significant (only) in the zero-inflated part. The negative regression coefficient for age implies that the older the children the lower the probability of caries-free children. Further, the negative regression coefficient for age at the start of brushing implies that the older the children start brushing their teeth the lower the probability of being caries-free. In addition, the children who brushed teeth irregularly and those who took more than two in-between-meals have a lower probability of being caries free. On the other hand, the use of systematic fluoride supplements increased the chances of being caries free.

The results of the ZINB regression model are somewhat more informative than of the Poisson regression or the logistic model of Vanobbergen *et al.*[6] Indeed, here we show clearly

TABLE 2. Signal-Tandmobiel® study: Maximum likelihood parameter estimates of the multiple ZINB regression model fitted to the dmft-index.

	Negative binomial part			Zero-inflated part		
Parameter	Estimate(SE)	95% CI		Estimate(SE)	95% CI	
Intercept	1.045(0.070)	0.907	1.183	0.146(0.148)	−0.144	0.437
x-ordinate	0.068(0.020)	0.029	0.107	−0.193(0.045)	−0.281	−0.105
y-ordinate	−0.033(0.022)	−0.075	0.010	0.014(0.046)	−0.076	0.103
Gender (girl)	0.034(0.039)	−0.043	0.112	−0.010(0.087)	−0.180	0.161
Age (years)	0.061(0.050)	−0.037	0.159	−0.346(0.109)	−0.560	−0.133
Brushing frequency (< 2)	−0.008(0.054)	−0.114	0.097	−0.281(0.136)	−0.547	−0.015
Age start brushing (years)	0.029(0.018)	−0.007	0.065	−0.227(0.044)	−0.313	−0.141
Systematic fluoride (yes)	−0.081(0.041)	−0.161	0.000	0.480(0.088)	0.308	0.652
Sugary drinks (yes)	0.196(0.042)	0.113	0.279	−0.271(0.089)	−0.446	−0.095
Between meals (> 2)	0.038(0.042)	−0.044	0.120	−0.223(0.096)	−0.410	−0.035
log(tau)	1.031(0.090)	0.854	1.208			

that many covariates are affecting more the probability of being caries-free and not so much the degree of caries experience.

4 Adjustment for misclassification applied to count models

4.1 Introduction

The above modelling does not take into account that the scores of the dental examiners are possibly corrupted. Indeed, at the end of the three calibration exercises the sensitivity and specificity of each dental examiner vis-a-vis a benchmark examiner (third author) was determined. There was quite some residual misclassification which needed to be taken into account in the modelling exercise.

For an excellent review on both measurement error and misclassification we refer to Gustafson.[10] Mwalili *et al.*[3] looked at ordinal responses subject to misclassification in a Bayesian context. Here a technique for adjusting for misclassification of counts (as response) is proposed in a likelihood and a Bayesian context.

4.2 Count models corrected for misclassification

Let $Y^* = (y_1^*, \cdots, y_n^*)'$ be the vector of the observed error-corrupted counts arising from the scoring of one dental examiner. If Y is the vector of the true unobservable responses obtained by a gold standard and $\boldsymbol{x}$ a vector of covariates and $\boldsymbol{\beta}$ the associated regression coefficients then

$$\Pr(Y^* = y^*|\boldsymbol{x}, \boldsymbol{\beta}) \ = \sum_y \Pr(Y^* = y^*|Y = y, \boldsymbol{x})\Pr(Y = y|\boldsymbol{x}, \boldsymbol{\beta}). \tag{5}$$

Expression (5) consists of (a) the misclassification model for Y^* given the true response and covariates; and (b) the underlying main model of interest. When misclassification is non-differential, the covariates provide no information about Y^* over and above what is provided by Y, so that (5) becomes

$$\Pr(Y^* = y^*|\boldsymbol{x}, \boldsymbol{\beta}) \ = \sum_y \pi_{y^*|y}\Pr(Y = y|\boldsymbol{x}, \boldsymbol{\beta}), \tag{6}$$

with $\pi_{y^*|y} = \Pr(Y^* = y^*|\boldsymbol{Y} = y)$ the misclassification probability of observing y^* instead of y. The misclassification probabilities can be collected into a $m' \times m'$ matrix $\Pi = (\pi_{y^*|y})$, where $m' = m + 1$, with m the maximal value of Y. Expression (6) can be applied to any misclassified count data distribution by replacing $\Pr(Y = y|\cdot)$ with the corresponding distribution.

When multiple examiners are involved we could either assume that (a) the misclassification matrix Π is the same for all examiners or that it is pooled over the examiners; or (b) the misclassification matrix varies with examiner. In the latter case expression (6) depends on the examiner-specific misclassification matrix. Observe that in some analyses non-differential misclassification for each examiner can result in differential misclassification overall. Namely, in many large-scale surveys the dental examiners are active in a restricted geographical area. If $\boldsymbol{x}$ contains a covariate indicating the geographical location of the dental examiner, then if misclassification depends on the examiner, misclassification will also indirectly depend on the the geographical covariate. Hence we deal with differential misclassification overall.

4.3 Misclassification model

To correct for misclassification, we need the probabilities Π. In practice, they are estimated using validation data and in caries research these data can be obtained from calibration exercises. A calibration exercise is a small sub-study in which subjects with a variety of pathologies but also caries-free subjects are scored by each dental examiner and a gold standard (or at least a benchmark examiner). The objective of the calibration exercise is to assess the agreement between the examiners and the benchmark examiner.

First, we consider the case of a single examiner. Let V^* denote the examiner score subject to error and V the true score that is given by a benchmark examiner obtained from the validation study. For each subject the dmft-index of $V^* = a$ given $V = b$, falls into one of m' categories. This gives rise to a $m' \times m'$ matrix $\boldsymbol{M} = (m_{ab})$ where m_{ab} is the observed frequency of the classification $(V^* = a, V = b)$. We assume that the bth column of $\boldsymbol{M}$, i.e. $\boldsymbol{m}_b$, follows a multinomial distribution:

$$\boldsymbol{m}_b \sim \text{Multinomial}\big(m_{+b}, \Pi_b\big), \tag{7}$$

where $\Pi_b = (\pi_{0|b}, \pi_{1|b}, \cdots, \pi_{m|b})'$, $\pi_{a|b} = \Pr(V^* = a|V = b)$ with $\sum_a \pi_{a|b} = 1$, and $m_{+b} = \sum_{a=0}^m m_{ab}$.

The multinomial estimate of $\pi_{a|b}$, i.e. $\hat{\pi}_{a|b} = \frac{m_{a|b}}{\sum_a m_{a|b}}$ is one possibility to estimate the misclassification probabilities. However, for a sparse table $\boldsymbol{M}$ the multinomial estimates $\hat{\pi}_{a|b}$ are either determined with high variability or do not exist, say when the benchmark examiner did not score 'b' in the validation data. Observe that for count data the misclassification table is often sparse. Clearly some modelling of the misclassification probabilities is needed to overcome this problem. Albert[11] suggested for ordinal scores the following model for conditional misclassification probabilities

$$\pi_{a|b} = \begin{cases} \frac{1}{1+\sum\limits_{c\neq b} \mathcal{G}(c|b)} & \text{if } a = b, \\ \frac{\mathcal{G}(a|b)}{1+\sum\limits_{c\neq b} \mathcal{G}(c|b)} & \text{if } a \neq b, \end{cases} \tag{8}$$

with $\mathcal{G}(a|b)$ being a positive-valued function of a given b. This model can also be used for count data.

Therefore, we suggest to model the misclassification probabilities for count data in a parsimonious way as follows:

$$\log \mathcal{G}(a|b) = \alpha_0 \text{ or} \tag{9}$$

$$\log \mathcal{G}(a|b) = \alpha_0 + \alpha_1|a - b|, \tag{10}$$

which gives symmetric misclassification probabilities with one and two parameters, respectively. The symmetric misclassification models (9) and (10) will be referred below to as the symmetric $1p$ (one parameter) and the symmetric $2p$ (two parameter) model, respectively. Further, the parameterization

$$\log \mathcal{G}(a|b) = \alpha_0 + \alpha_1(a - b)\mathrm{I}(a > b) + \alpha_2(b - a)\mathrm{I}(a < b) \tag{11}$$

allows for asymmetric misclassification probabilities, where $\mathrm{I}(x)$ is an indicator function: $\mathrm{I}(x) = 1$ if x is true and 0 otherwise. Expressions (9), (10) and (11) imply a reduction of the parameters to estimate from m to 1, 2 and 3, respectively for each vector Π_b. Further, even when the bth column of $\boldsymbol{M}$ contains only zeros, expressions (9), (10) and (11) allow to estimate $\pi_{a|b}$. The log-likelihood of the misclassification model given the validation data $\boldsymbol{M}$ is

$$\mathcal{L}(\Pi; \boldsymbol{M}) = -\sum_a \sum_b m_{ab} \log \pi_{a|b}, \tag{12}$$

where $\pi_{a|b}$ is given by expression (8).

Given the analytical first derivatives of the symmetric and the asymmetric misclassification model (Appendix A.2), maximum likelihood estimation is straightforward and can be carried out using the BFGS algorithm.[9] For the Bayesian analysis of the misclassification model the software WinBUGS can be used as in Mwalili *et al.*[3] When multiple examiners

TABLE 3. Signal-Tandmobiel® pooled validation data: The parameters estimates of both symmetric and asymmetric misclassification model from likelihood and Bayesian (WinBUGS) approach.

		Likelihood approach				Bayesian approach			
Model		Estimate(SE)	95% CI		AIC†	Mean(SD)	95% CR‡		DIC†
Symmetric $1p$	$\hat{\alpha}_0$	$-4.416(0.119)$	-4.65	-4.18	964	$-4.418(0.119)$	-4.66	-4.19	524
Symmetric $2p$	$\hat{\alpha}_0$	$-0.806(0.240)$	-1.28	-0.34	635	$-0.792(0.246)$	-1.28	-0.29	195
	$\hat{\alpha}_1$	$-1.253(0.153)$	-1.55	-0.95		$-1.270(0.157)$	-1.61	-0.98	
Asymmetric	$\hat{\alpha}_0$	$-0.483(0.259)$	-0.99	0.02	556	$-0.466(0.256)$	-0.99	0.03	115
	$\hat{\alpha}_1$	$-2.153(0.253)$	-2.65	-1.66		$-2.190(0.255)$	-2.69	-1.71	
	$\hat{\alpha}_2$	$-0.676(0.153)$	-0.98	-0.38		$-0.696(0.154)$	-1.01	-0.41	

†AIC – Akaike Information Criterion; DIC – Deviance Information Criterion.

‡CR – Credibility Interval.

TABLE 4. Signal-Tandmobiel® examiner-specific validation data: the selected misclassification model for the 16 dental examiners.†

Examiner	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Misc model	A	S-1	S-2	A	A	S-1	A	S-2	S-1	S-2	A	S-1	A	A	S-2	S-1

†A = asymmetric; S-1 = symmetric $1p$; S-2 = symmetric $2p$.

are involved, the total likelihood corresponding to the validation study is the sum of the expressions (12) each corresponding to an examiner.

We corrected for misclassification in the global sense (pooled over all examiners) and in an examiner-specific manner. The asymmetric misclassification structure seemed appropriate for the global correction (see Table 3). For the examiner-specific correction in most cases the asymmetric misclassification model seemed best; see Table 4 for the choice of the misclassification models for each examiner separately.

5 The corrected ZINB applied to the oral health study

The corrected ZINB model is obtained by replacing $\Pr(Y = y|\boldsymbol{x}, \boldsymbol{\beta})$ in expression (6) by the ZINB distribution. Therefore, the total likelihood of the *corrected* zero-inflated negative binomial (CZINB) distribution adjusted for misclassification in a pooled manner or when only one examiner is involved, is

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\beta},\boldsymbol{\gamma},\tau;\boldsymbol{y}*,\boldsymbol{X},\boldsymbol{Z},\Pi) \;=\; & -\sum_{i=1}^{n}\log\left\{ \frac{e^{\boldsymbol{z}_i'\boldsymbol{\gamma}} + \;\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}+\tau}{\tau}\Big)^{-\tau}}{1+e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}}\pi_{(y_i^*|0)} \;+\right. \\ & \left.\sum_{y=1}^{m}\left(\frac{\Gamma(\tau+y)\;\; 1+e^{\boldsymbol{z}_i'\boldsymbol{\gamma}} \quad \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}+\tau}{\tau}\Big)^{-\tau}}{\Gamma(\tau)\,\Gamma(1+y)\;\; 1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}\,\tau^{\;y}}\pi_{(y_i^*|y)}\right)\right\}.\end{aligned} \tag{13}$$

When multiple examiners are involved, $\mathcal{L}_c$ is a sum of the examiner-specific log-likelihood contributions. In order to account for the uncertainty of the misclassification probabilities, we maximized log-likelihood (13) and misclassification log-likelihood (12) jointly, i.e. the sum of the two log-likelihoods is maximized. In a Bayesian way the uncertainty of Π is taken into account by sampling simultaneously from the posterior distribution of the main and of the validation study. This can be done using WinBUGS as exemplified by Mwalili *et al.*[3] for the ordinal logistic regression.

The ZINB model was corrected in a global manner and in an examiner-specific way. The former approach will be referred to as 'pooled correction' while the latter as 'examiner-specific correction'. As can be seen in Figure 1 the two correction mechanisms do not give very different results for the model without covariates. The proportion of caries-free children from the pooled correction and from the examiner-specific correction are about 45% and 47% respectively, which is a slight increase over the observed 44%.

The results of fitting the corrected ZINB regression model to the dmft-index of the STM data are shown in Table 4. For the pooled correction, the East-West gradient ($x-$ordinate) remains significant in both parts of the corrected ZINB model. In contrast, for the examiner-specific correction the East-West gradient vanishes in the negative binomial part. The discrepancy between the two corrections is explained by the fact that the pooled correction ignores the fact that the misclassification model has become differential.

The ZINB regression model together with the examiner-specific correction shows more overdispersion ($\hat{\tau} = 3.9$) than that of the pooled correction ($\hat{\tau} = 5.7$). Indeed, the variance of the ZINB distribution is equal to $(1-\hat{p})\hat{\lambda}(1+\hat{p}\hat{\lambda}+\hat{\lambda}/\hat{\tau})]$. Thus, the examiner-specific correction preserves the negative binomial structure of the dmft-index more than the pooled correction. The Bayesian results are not shown as they are practically equal to results from the likelihood analysis.

6 Discussion

Another possible way to account for examiners' misclassification is to include a dummy variable for each examiner in the regression model as covariate. This approach does not need any validation study, and hence looks attractive. However, this approach has at least two drawbacks: (a) by including a dummy variable in the model one performs a correction, but not necessarily the correct one; and (b) inclusion of the dummy variable can only correct for bias and not for variability.

TABLE 5. Signal-Tandmobiel® study: Maximum likelihood estimates of the corrected ZINB regression model fitted to the dmft-index in a pooled and an examiner-specific way.

	Pooled correction			Examiner-specific correction		
Parameter	Estimate(SE)	95% CI		Estimate(SE)	95% CI	
				Negative binomial part		
Intercept	1.287(0.068)	1.154	1.421	1.172(0.074)	1.027	1.317
x-ordinate	0.054(0.019)	0.017	0.090	0.041(0.022)	−0.001	0.083
y-ordinate	−0.029(0.020)	−0.068	0.011	−0.017(0.023)	−0.062	0.029
Gender (girl)	0.033(0.037)	−0.039	0.105	0.044(0.042)	−0.038	0.126
Age (years)	0.013(0.047)	−0.079	0.105	−0.014(0.055)	−0.122	0.094
Brushing frequency (< 2)	−0.016(0.049)	−0.113	0.080	−0.003(0.055)	−0.112	0.106
Age start brushing (years)	0.013(0.017)	−0.020	0.047	0.016(0.019)	−0.022	0.054
Fluoride supplement (yes)	−0.026(0.039)	−0.102	0.049	−0.044(0.043)	−0.129	0.041
Sugary drinks (yes)	0.158(0.040)	0.080	0.236	0.181(0.045)	0.094	0.269
Between meals (> 2)	0.024(0.038)	−0.051	0.100	0.009(0.044)	−0.077	0.095
				Zero inflated part		
Intercept	0.427(0.160)	0.113	0.741	0.355(0.159)	0.043	0.667
x-ordinate	−0.213(0.048)	−0.308	−0.119	−0.205(0.050)	−0.303	−0.107
y-ordinate	0.018(0.048)	−0.077	0.113	0.045(0.050)	−0.052	0.142
Gender (girl)	−0.007(0.092)	−0.187	0.173	−0.011(0.094)	−0.194	−0.173
Age (years)	−0.455(0.116)	−0.683	−0.228	−0.394(0.118)	−0.626	−0.162
Brushing frequency (< 1)	−0.293(0.141)	−0.569	−0.017	−0.301(0.143)	−0.582	−0.021
Age start brushing (years)	−0.255(0.046)	−0.346	−0.165	−0.238(0.046)	−0.329	−0.148
Fluoride supplement (yes)	0.582(0.093)	0.400	0.764	0.553(0.094)	0.369	0.737
Sugary drinks (yes)	−0.327(0.095)	−0.514	−0.141	−0.305(0.096)	−0.494	−0.117
Between-meals (> 2)	−0.248(0.100)	−0.444	−0.051	−0.226(0.102)	−0.426	−0.026
log(tau)	1.737(0.144)	1.455	2.019	1.370(0.117)	1.141	1.598

Clearly our approach can be applied to all count data models like the Poisson, generalized Poisson, zero-inflated Poisson Negative binomial, finite mixture of Poisson distribution, etc. Hence, our proposal is quite general.

Parameter estimation in the likelihood approach was done using *R*-software calling C++ routines for fast computation of the likelihood and the first derivative. On the other hand the parameter estimation in the Bayesian approach was done using WinBUGS calling a faster WBDev (WinBUGS Development) routine. Both the *R*-software and WinBUGS program are available from the first author.

Finally, Gustafson[10] showed how a non-differential measurement error on a continuous variable can lead to a differential misclassification model for the dichotomized continuous variable. In this paper we give another example of how a non-differential misclassification process can turn itself into a differential process.

Acknowledgements

The study was financially supported by Research Grant OT/00/35, Catholic University Leuven; data collection was supported by Unilever, Belgium. The Signal-Tandmobiel® project comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care). We also acknowledge support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs.

Appendix A First order derivatives

Appendix A.1 ZINB regression model

The first order derivative of $\mathcal{L}_z$ (expression (4)) with respect to the $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau)'$-parameters:

$$\frac{\delta \mathcal{L}_z}{\delta \beta_j} = \begin{cases} \sum_{i=1}^{n} \left(\frac{\lambda_i \, \tau}{(\lambda_i+\tau)\,(1+q_i\, r_i)} \boldsymbol{x}_{ij} \right), & y = 0, \\ \sum_{i=1}^{n} \left(1 - \frac{\tau+y}{\lambda_i+\tau} \right) \tau\, \boldsymbol{x}_{ij} & y > 0. \end{cases}$$

$$\frac{\delta \mathcal{L}_z}{\delta \gamma_j} = \begin{cases} \sum_{i=1}^{n} \left(- \left(\frac{1}{1+e^{\boldsymbol{z}'_i \boldsymbol{\gamma}}} \right) + \frac{1}{1+q_i\, r_i} \boldsymbol{z}_{ij} \right), & y = 0, \\ \sum_{i=1}^{n} \left(\frac{q_i}{1+q_i} z_{ij} \right), & y > 0. \end{cases}$$

$$\frac{\delta \mathcal{L}_z}{\delta \log \tau} = \begin{cases} \sum_{i=1}^{n} \left(\frac{-\lambda_i+\lambda_i \log r_i}{(\lambda_i+\tau)\,(1+q_i\, r_i)} \tau \right) & y = 0, \\ \sum_{i=1}^{n} \left(-1 + \frac{\tau+y}{\lambda_i+\tau} + \log(\frac{\lambda_i+\tau}{\tau}) + \log\Gamma'(\tau) - \log\Gamma'(\tau+y) \right) \tau, & y > 0, \end{cases},$$

where $\lambda_i = e^{\boldsymbol{x}'_i \boldsymbol{\beta}}$, $r_i = \left(\frac{\lambda_i+\tau}{\tau}\right)^{\tau}$, $q_i = e^{\boldsymbol{z}'_i \boldsymbol{\gamma}}$ and $\log\Gamma'(x) = \frac{\delta\, log\, \Gamma(x)}{\delta\, x}$.

Appendix A.2 Misclassification model

The first order derivative of $\mathcal{L}_m$ (expression (12)) with respect to the $\boldsymbol{\alpha} = (\alpha_0, \alpha_0, \alpha_0)'$-parameters for:

(a) Symmetric misclassification:

(i) Symmetric 1p

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_0} = -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1\Big) \Big(1 + \sum_{c \neq b} \mathcal{G}(c|b)\Big) + 1}{1 + \sum_{c \neq b} \mathcal{G}(c|b)} \right).$$

(ii) Symmetric $2p$

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_0} = -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1\Big) \left(1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)\right) + 1}{1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)} \right).$$

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_1} = -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)|a-b| \left(1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)\right) - \sum\limits_{c \neq b} |c-b| \mathcal{G}(c|b)}{1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)} \right).$$

(b) Asymmetric misclassification:

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_0} = -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1\Big) \left(1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)\right) + 1}{1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)} \right).$$

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_1} = -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(a-b)\mathrm{I}(a > b) \left(1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)\right) - \sum\limits_{c \neq b} (c-b)\mathrm{I}(c > b) \mathcal{G}(c|b)}{1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)} \right).$$

$$\frac{\delta \mathcal{L}_m}{\delta \alpha_2} = -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(b-a)\mathrm{I}(a < b) \left(1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)\right) - \sum\limits_{c \neq b} (b-c)\mathrm{I}(c < b) \mathcal{G}(c|b)}{1 + \sum\limits_{c \neq b} \mathcal{G}(c|b)} \right).$$

Appendix A.3 Corrected ZINB regression model

The first order derivative of $\mathcal{L}_c$ (expression (13)) with respect to the $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau)'$-parameters:

$$\begin{aligned} \frac{\delta \mathcal{L}_c}{\delta \beta_j} = & -\sum_{i=1}^{n} \frac{\lambda_i \left(\frac{\lambda_i + \tau}{\tau}\right)^{-1-\tau} \pi_{(y_i^*|0)}}{(1+q_i)\, D_i} \boldsymbol{x}_{ij} \\ & -\sum_{i=1}^{n} \left(\sum_{y=1}^{m} \frac{\left(\frac{\lambda_i + \tau}{\tau}\right)^{-1-\tau} \boldsymbol{x}_{ij}\, \Gamma(\tau + y)\, (\lambda_i - y)\, \pi_{(y_i^*|y)}}{(1+q_i)\, (1 + \tau/\lambda_i)^y\, \Gamma(\tau)\, \Gamma(1+y) D_i} \boldsymbol{x}_{ij} \right), \end{aligned}$$

$$
\begin{aligned}
\frac{\delta \mathcal{L}_c}{\delta \gamma_j} &= -\sum_{i=1}^{n} \frac{q_i\ \left(\left(\frac{\lambda_i+\tau}{\tau}\right)^{\tau} - 1\right)\ \boldsymbol{z}_{ij}\, \pi_{(y_i^*|0)}}{(1+q_i)^2\ \left(\frac{\lambda_i+\tau}{\tau}\right)^{\tau} D_i} \\
&\quad + \sum_{i=1}^{n} \left(\sum_{y=1}^{m} \frac{q_i\, \boldsymbol{z}_{ij}\, \Gamma(\tau+y)\, \pi_{(y_i^*|y)}}{(1+q_i)^2\ \left(\frac{\lambda_i+\tau}{\tau}\right)^{\tau} (1+\lambda_i\, \tau)^{y}\, \Gamma(\tau)\, \Gamma(1+y) D_i} \right), \\
\frac{\delta \mathcal{L}_c}{\delta \log \tau_j} &= \sum_{i=1}^{n} \frac{\left(\frac{\lambda_i+\tau}{\tau}\right)^{-1-\tau}\ \left(-\lambda_i + (\lambda_i+\tau)\ \log(\frac{\lambda_i+\tau}{\tau})\right)\ \pi_{(y_i^*|0)}}{(1+q_i)\ D_i} \\
&\quad - \sum_{i=1}^{n} \left(\sum_{y=1}^{m} \frac{\left(\frac{\lambda_i+\tau}{\tau}\right)^{-1-\tau} \Gamma(\tau+y)\, \pi_{(y_i^*|y)} \triangle_i}{(1+q_i)\ \tau\, (1+\lambda_i\,)^{y}\, \Gamma(\tau)\, \Gamma(1+y) D_i} \right),
\end{aligned}
$$

where $\lambda_i = e^{\boldsymbol{x}'_i \boldsymbol{\beta}}$, $q_i = e^{\boldsymbol{z}'_i \boldsymbol{\gamma}}$,

$\triangle_i \ = \ \left(\lambda_i - (\lambda_i+\tau)\ \log(\frac{\lambda_i+\tau}{\tau}) + (\lambda_i+\tau)\ (-\log\Gamma'(\tau) + \log\Gamma'(\tau+y)) - y\right)$, and D_i is the ZINB likelihood of the ith individual.

References

1 Lewsey JD, Thomson WM. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dentistry and Oral Epidemiology 2004;32:183–189.

2 Vanobbergen J, Martens L, Lesaffre E, Declerck D. The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. European Journal of Paediatric Dentistry 2000;2:87–96.

3 Mwalili S, Lesaffre E, Declerck D. A Bayesian Ordinal Logistic Regression Model to Correct for Inter-observer Measurement Error in a Geographical Oral Health Study. Journal of the Royal Statistical Society, Series C 2005;1:77–93.

4 Böhning D, Dietz E, Schlattman P, Mendonça L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society, Series A 1999;162:195–209.

5 Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992;34:1–14.

6 Vanobbergen J, Martens L, Lesaffre E, Bogaerts K, Declerck D. Assessing the indicators for dental caries in the primary dentition. Community Dentistry and Oral Epidemiology 2001;29:424–434.

7 Lord D, Washington SP, Ivan JN. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 2005;37:35–46.

8 Carrivick PJW, Lee AH, Yau KKW. Zero-inflated Poisson modeling to evaluate occupational safety interventions. Safety Science 2003;41:53–63.

9 Nocendal J, Wright SJ. Numerical Optimization. New York: Springer-Verlag; 1999.

10 Gustafson P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Chapman & Hall, New York; 2004.

11 Albert PS, Hunsberger A S, Biro FM. Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. Journal of the American Statistical Association 1997;92:1304–1311.