Clin Oral Invest (2013) 17:1799–1805
DOI 10.1007/s00784-013-0993-4

REVIEW

# Correction for misclassification of caries experience in the absence of internal validation data

**T. Mutsvari · D. Declerck · E. Lesaffre**

Received: 25 November 2012 / Accepted: 23 April 2013 / Published online: 11 May 2013


**Abstract**
*Objectives* To quantify the effects of risk factors and/or determinants on disease occurrence, it is important that the risk factors as well as the variable that measures the disease outcome are recorded with the least error as possible. When investigating the factors that influence a binary outcome, a logistic regression model is often fitted under the assumption that the data are collected without error. However, most categorical outcomes (e.g., caries experience) are accompanied by misclassification and this needs to be accounted for. The aim of this research was to adjust for binary outcome misclassification using an external validation study when investigating factors influencing caries experience in schoolchildren.
*Materials and methods* Data from the Signal Tandmobiel® study were used. A total of 500 children from the main and 148 from the validation study were included in the analysis. Regression models (with several covariates) for sensitivity and specificity were used to adjust for misclassification in the main data.
*Results* The use of sensitivity and specificity modeled as functions of several covariates resulted in a better correction compared to using point estimates of sensitivity and specificity. Age, geographical location of the school to which the child belongs, dentition type, tooth type, and surface type were significantly associated with the prevalence of caries experience.
*Conclusions* Sensitivity and specificity calculated based on an external validation study may resemble those obtained from an internal study if conditioned on a rich set of covariates.
*Clinical relevance* Main data can be corrected for misclassification using information obtained from an external validation study when a rich set of covariates is recorded during calibration.

**Keywords** Misclassification correction · Caries experience · Validation

T. Mutsvari · E. Lesaffre (✉)
Leuven Biostatistics and Statistical Bioinformatics Centre (L-BioStat), KU Leuven, Kapucijnenvoer 35 Blok D, Box 7001, 3000 Leuven, Belgium
e-mail: emmanuel.lesaffre@med.kuleuven.be

D. Declerck
Department of Oral Health Sciences, KU Leuven, Kapucijnenvoer 7 Blok A, Box 7001, 3000 Leuven, Belgium

E. Lesaffre
Department of Biostatistics, Erasmus Medical Centre, Erasmus University Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands

## Introduction

Epidemiologic studies aim to explore associations between risk factors and/or determinants and disease occurrence. To quantify the effects of risk factors and/or determinants on disease occurrence, it is important that the risk factors as well as the variable that measures the disease outcome are recorded with the least error as possible. Often, the measurements obtained are noisy (error prone) versions of the true underlying variable of primary interest. When the variables under consideration are categorical, such error is termed misclassification error.

It is well documented that the process of scoring caries experience (CE) is challenging, thereby affecting the quality of the obtained data due to misclassification. A lot of effort has been devoted to improve the quality of CE recording by providing guidelines for caries surveys [1].

To investigate the factors that influence a binary CE outcome, a logistic regression model is often fitted under the assumption that the data are collected without error. When misclassification error is ignored in the analysis, then

this may lead to biased estimates, hence erroneous conclusions [2, 3]. A possible way of dealing with this is correcting for misclassification.

When correction for misclassification in the logistic regression is envisaged, one needs to understand better the underlying misclassification mechanism. In a previous analysis of the data set considered in this paper, misclassification was found to be influenced by several factors, i.e., the process is differential [2].

A vast amount of literature exists on correcting for misclassification in models for categorical data [4–6], and for this, a validation data set is most often needed to perform the correction. In a validation study, a small sample of subjects or study units is scored by both the benchmark scorer and the examiner(s) enabling the estimation of sensitivity (SE) and specificity (SP) of the examiners vis-à-vis the benchmark scorer. Validation data may be *internal* and nested within the main study, as is the case when a subset of the main study sample is reexamined by the benchmark scorer. This is also referred to as double sampling [7, 8]. Internal validation is regarded as the best possible way to assess misclassification. It seems plausible to assume that the misclassification mechanisms in the internal validation and main data are similar [9]. However, in large or complex surveys, internal validation data are challenging to obtain in practice due to several (practical) constraints, e.g., the need to doubly examine in different locations and at different time points, especially when several examiners are involved and spread over different geographical areas [10].

Alternatively, *external* validation data can be obtained from a separate independent sample from the main study. External validation data may be obtained in different ways: (a) the sample being a random sample from the population of interest, sampled in addition to the main data and assessed (or recorded) under identical conditions as the main data, (b) being a convenience sample (not a random sample from the population of interest) but scored under identical conditions as the main data, (c) being a convenience sample from the population of interest and assessed under (slightly) different conditions than the main data, (d) being a (random or nonrandom) sample taken from a different population examined under identical conditions, and (e) being a (random) sample taken from a different population and examined under different conditions. Note that one would expect the misclassification probabilities from the external validation type (a) to be least biased since they are based on a random sample from the same population as the main study and examined in identical conditions. For cases (b) to (e), unbiased misclassification probabilities cannot be guaranteed. Hence, the use of data obtained from such validation studies needs extra caution.

Since it is not always feasible to collect internal validation data, it is worth exploring approaches for using available external validation data. In fact, in most large-scale epidemiological surveys, inter-examiner agreement will be assessed using an external sample. In this paper, an approach to deal with this situation is illustrated on data (main and validation) obtained from the Signal Tandmobiel® (ST) study [11]. Since data in this CE survey were recorded at surface level, the multilevel structure (surfaces nested within teeth within mouths) is additionally needed to be taken into account.

The aim of this research was to illustrate an approach of using external validation data to correct for misclassification. Factors influencing CE in schoolchildren using a multilevel logistic model were investigated.

## Materials and methods

### Motivating data set

The Signal Tandmobiel® study is a longitudinal (1996–2001) oral health intervention project that took place in Flanders (north of Belgium). At the first examination, the average age of the children was 7.1 years (SD=0.4). Sixteen trained dentists (examiners) conducted annual oral health examinations on 4,468 children (2,315 boys and 2,153 girls) from 179 primary schools, after parental consent was obtained. Data on oral hygiene and dietary habits were obtained through structured questionnaires, completed by the parents. The children received a clinical examination based on the diagnostic criteria for caries prevalence surveys published by the British Association for the Study of Community Dentistry (BASCD) [12]. Caries experience was determined at surface level and expressed using the dmfs/DMFS score [13]. The clinical examinations took place in a mobile dental clinic, equipped with a standard dental chair and artificial dental light source. Recording was performed by a visual–tactile method, using a mouth mirror and a WHO/CPITN type E probe (Prima Instruments, Gloucester, UK). No radiographs were taken. More details of the ST study are given in Declerck et al. [11].

Out of the 4,468 children, a random sample of 500 children was selected to constitute the main data. We considered only the data collected at the last visit, i.e., year 2001, when the children were about 12 years old. In the present work, CE was binarized at surface level (0 if dmfs/DMFS=0 and 1 if dmfs/DMFS≠0).

Training sessions for scoring CE were organized parallel to the study and the scoring behavior (SE and SP) of each of the 16 dental examiners was compared to that of the benchmark (second author, DD). During the study period (1996–2001), three calibration exercises (training sessions) (1996, 1998, 2000) for scoring CE, involving 92, 32, and 24 children, respectively, were organized. The clinical assessment of CE was undertaken in a way identical to the one used in the main study, but the children included in these calibration exercises

were not sampled at random from the main study. Rather, a school was selected where a relatively high prevalence of caries experience could be expected in order to guarantee sufficient variation of pathology. Since there was not much difference in examiners' scoring behavior between the different calibration exercises, data were combined to form the external validation data set.

### Misclassification for binary data

Misclassification information was compiled in a 2×2 contingency table with the benchmark scores and examiners' scores. The entries in this misclassification table allow estimating the SE and the SP of the dental examiners vis-à-vis the benchmark scorer. SE and SP are statistical measures of the performance of a binary classification test. SE measures the proportion of actual positives which are correctly identified as such (e.g., the proportion of surfaces with CE that are identified as having the condition). SP measures the proportion of negatives which are correctly identified (e.g., the proportion of surfaces without CE that are identified as not having the condition). Hence, SE and SP measure the capability of the examiners in detecting the true prevalence of CE or noting the true absence of the disease. The higher the SE and SP, the better is the scoring behavior of the examiner.

### Multilevel logistic regression model

The methodology presented in this paper is explained for a binary outcome of CE of each surface, that is CE is 1 if the surface shows CE and 0 if not. Since the response is considered binary, a popular model to apply is logistic regression. However, this model assumes independence of CE responses obtained from different sites. However, the CE of surfaces from a same tooth is correlated since they are exposed to similar conditions. Similarly, teeth that belong to one mouth share common characteristics, resulting in a hierarchical structure of the data. Hence, a multilevel logistic regression was considered for further exploration.

A first analysis of the main data was undertaken without correcting the binary CE outcome variable for misclassification (also called, the naïve approach). This means that a multilevel logistic regression model was fitted as if there was no misclassification present. The covariates included in this model were gender (girls versus boys), age, geographical location (represented by the standardized $(x,y)$ coordinate of the municipality of the school to which the child belongs), dentition type (permanent versus deciduous), tooth type (canine, incisor, molar, and premolar), and surface type (distal, mesial, lingual, occlusal, and buccal). More technical details are given in Appendix.

The second analysis considered a multilevel logistic model, extended to account for misclassification using estimates of SE and SP. This information was obtained from the ST validation study and a non-differential misclassification correction approach was used in this second analysis. This correction is justified if the validation data are internal, i.e., a random sample of the main sample.

However, the children who participated in the ST calibration exercises were not sampled at random from the main study, and therefore, the obtained validation data are of the external type. This implies that the misclassification mechanism in the main and validation data are likely to be different. Therefore, a third analysis was undertaken considering the use of logistic regression models of SE and SP including a rich set of covariates. In this way, an attempt was made to bring the characteristics of the external validation data closer to those of an internal one. More details on this approach are given in Appendix. The estimates of the regression coefficients for SE and SP were imputed into the model for the main data in order to adjust for misclassification. The covariates included in the SE and SP models were gender (girls versus boys), dentition type (permanent versus deciduous), tooth type (canine, incisor, molar, and pre-molar), and surface type (distal, mesial, lingual, occlusal, and buccal). More details regarding the model for the main data are given in Appendix.

### Estimating the parameters

To estimate the parameters in the models mentioned above, a Bayesian approach was used since this is better suited for misclassification problems. In a Bayesian approach, prior knowledge about the parameters is combined with the observed data (likelihood) to yield the posterior distribution. From the posterior distribution, we obtain the estimates of the parameters (posterior mean or median) and the standard error of the estimate. Further details on the Bayesian approach can be found in Appendix.

## Results

Out of the 53,283 surfaces included in the main data set, 1,675 (3.14 %) showed CE. In the validation data, 519 (5.33 %) of 9,741 surfaces presented CE. Details on the performance of the dental examiners in the ST study have been reported elsewhere [14]. SE and SP values of the examiners ranged between 97–99 and 63–90 %, respectively.

Table 1 shows the estimates obtained from the logistic models of SE and SP fitted to the validation data set. These were used for differential misclassification correction of the main data (model 3). Since the multilevel data structure in the validation data set was not accounted for (this would complicate analysis additionally), further interpretation of these data was not undertaken. Table 2 shows the results of the three multilevel models for the main data, i.e., (a) no

**Table 1** Parameter estimates and standard deviations (SD) of the misclassification model (SE and SP) used for differential misclassification correction (pooled over all examiners)

| Parameter | Sensitivity Mean (SD) | Specificity Mean (SD) |
|---|---|---|
| Fixed effects | | |
| Intercept | −0.35 (0.50) | 7.12 (0.30) |
| Gender | | |
| Girls | 0.41 (0.14) | 0.29 (0.11) |
| Boys | – | – |
| Dentition type | | |
| Permanent | −0.26 (0.27) | 0.22 (0.12) |
| Deciduous | – | – |
| Tooth type | | |
| Canine | 0.27 (0.43) | −1.35 (0.40) |
| Molar | 1.74 (0.38) | −2.99 (0.34) |
| Premolar | 1.64 (0.51) | −0.25 (0.42) |
| Incisor | – | – |
| Surface type | | |
| Distal | 0.25 (0.32) | 0.22 (0.20) |
| Mesial | −0.52 (0.26) | 0.11 (0.19) |
| Lingual | −0.08 (0.26) | 0.14 (0.17) |
| Occlusal | 0.43 (0.22) | −0.99 (0.15) |
| Buccal | – | – |

correction, which means that the multilevel model was fitted to the main data without correcting for misclassification, (b) non-differential correction, which means that SE and SP used for correction in the main data did not depend on covariates, and (c) differential correction in which the proposed approach is applied, i.e., use of SE and SP conditioned on several covariates, i.e., expressed as a logistic regression model of several covariates. In this table, a positive estimate for a categorical variable reflects a higher prevalence of CE compared to the reference level. For a continuous variable, a positive estimate reflects an increase in the probability of CE with a unit increase in that variable on a log scale. Negative estimates reflect the opposite, i.e., a lower probability of presenting CE.

The parameter estimates obtained with differential correction are generally higher in absolute value than those with no correction (smallest) and non-differential correction. Also, the 95 % Bayesian confidence intervals for differential correction are often wider than for the no correction and non-differential ones. Under differential correction, a significant effect of the *x*-coordinate (0.50 [0.02; 1.00]) was noted, which was not the case in the two other models. The positive effect of tooth type (canine versus incisor) (0.32 [−0.65; 1.31]) under non-differential correction changed into a negative one (−0.69 [−2.05; 0.54]) under differential correction, although the effect was nonsignificant in both cases. Further, there was a significant effect of age in all models: no correction (1.27 [0.56; 2.04]), non-differential correction (1.67 [0.61; 2.68]), and differential correction (2.12 [0.95; 3.29]) with the probability of presenting CE increasing with an increase in age. Permanent teeth showed significantly less CE prevalence compared to deciduous teeth in all models: no correction (−2.12 [−2.38; −1.86]), non-differential correction (−2.79 [−3.20; −2.42]), and differential correction (−3.26 [−3.81; −2.76]). Regarding surface type, occlusal surfaces showed significantly more CE compared to buccal surfaces in all three models: no correction (3.64 [3.37; 3.90]), non-differential correction (4.65 [4.15; 5.20]), and differential correction (5.17 [4.60; 5.91]). As indicated by the random effects, the clustering effect was higher at mouth level than at tooth level.

## Discussion

The effects of misclassification have been well addressed in the literature. Misclassification has been studied for long and one of the general findings is that it introduces bias in the obtained results. A detailed discussion can be found in Neuhaus [3], Brenner and Savitz [15], and Wacholder et al. [16]. Little work has been done on evaluating the use of external validation studies to adjust for this type of misclassification. Analyses presented here revealed that ignoring misclassification can result in erroneous conclusions. Further, different misclassification adjustments result in different conclusions. Hence, the underlying type of misclassification should be well studied in order to do a proper adjustment. In the present study, the two correction approaches (no correction and non-differential) failed to depict a significant effect of the geographical location of the school which the child attends (*x*-coordinate), as was shown in a previous analysis [5].

A fundamental assumption in a validation study is that true scores are recorded by a gold standard, i.e., an approach (instrument or examiner) that is 100 % error free. However, in practice, the scores are often generated by a benchmark scorer, i.e., an experienced examiner who is assumed to be error free or nearly so. A benchmark scorer can also be referred to as an "alloyed gold standard." The question of how "alloyed" a benchmark can be in order to maintain valid statistical inferences is subjective. More discussion on this matter is given by Wacholder et al. [16]. Indeed, it is possible that methods correcting for the effect of misclassification that make use of information gathered by comparing to a benchmark scorer might introduce more bias than they are correcting. The benchmark scorer in the present study had a vast experience in CE surveys and was trained by a BASCD trainer in 1990. It is recommended that in longitudinal surveys at least once every 2 years, benchmark scorers or trainers from different districts should meet to undertake a training and calibration amongst themselves [12]. In the survey presented here, this was not possible due to logistic constraints.

**Table 2** Parameter estimates and 95 % credible intervals of the main model for cross-sectional data without correction and with non-differential and with differential correction

| Parameter | Model 1<br>No correction<br>Estimate [2.5 %; 97.5 %] | Model 2<br>Non-differential<br>Estimate [2.5 %; 97.5 %] | Model 3<br>Differential<br>Estimate [2.5 %; 97.5 %] |
|---|---|---|---|
| Fixed effects | | | |
| Intercept | −8.86 [−9.79; −8.12] | −10.86 [−12.31; −9.53] | −10.73 [−12.78; −9.23] |
| Gender | | | |
| Girls | −0.07 [−0.63; 0.48] | −0.18 [−0.85; 0.57] | −0.07 [−1.00; 0.87] |
| Boys | – | – | – |
| Age | 1.27 [0.56; 2.04] | 1.67 [0.61; 2.68] | 2.12 [0.95; 3.29] |
| Geographical location | | | |
| *x*-coordinate | 0.27 [−0.05; 0.59] | 0.37 [−0.01; 0.77] | 0.50 [0.02; 1.00] |
| *y*-coordinate | −0.18 [−0.46; 0.10] | −0.30 [−0.71; 0.14] | −0.35 [−0.84; 0.14] |
| Dentition type | | | |
| Permanent | −2.12 [−2.38; −1.86] | −2.79 [−3.20; −2.42] | −3.26 [−3.81; −2.76] |
| Deciduous | – | – | – |
| Tooth type | | | |
| Canine | −0.14 [−0.83; 0.56] | 0.32 [−0.65; 1.31] | −0.69 [−2.05; 0.54] |
| Molar | 3.98 [3.51; 4.60] | 5.46 [4.54; 6.43] | 3.99 [2.88; 5.09] |
| Premolar | −1.95 [−2.70; −1.18] | −2.14 [−2.80; −1.48] | −4.17 [−5.71; −2.84] |
| Incisor | – | – | – |
| Surface type | | | |
| Distal | 0.86 [0.58; 1.12] | 1.17 [0.81; 1.55] | 1.42 [0.89; 2.05] |
| Mesial | 1.55 [1.29; 1.80] | 2.13 [1.74; 2.55] | 2.96 [2.35; 3.82] |
| Lingual | 0.13 [−0.16; 0.40] | 0.19 [−0.20; 0.56] | 0.47 [−0.05; 0.99] |
| Occlusal | 3.64 [3.37; 3.90] | 4.65 [4.15; 5.20] | 5.17 [4.60; 5.91] |
| Buccal | – | – | – |
| Random effects | | | |
| $\sigma^2_{\text{mouth}}$ | 6.50 [5.20; 8.18] | 10.82 [8.12; 14.44] | 15.92 [11.97; 21.90] |
| $\sigma^2_{\text{tooth}}$ | 3.35 [2.86; 3.96] | 4.75 [3.42; 6.45] | 6.50 [4.97; 9.36] |
| $\sigma^2_{\text{examiner}}$ | 0.13 [0.0004; 0.72] | 0.27 [0.0001; 1.44] | 0.30 [0.002; 1.69] |

The easiest and probably the most convenient way to correct for misclassification is to use SE and SP estimates obtained from independent but comparable studies. In that case, the values of SE and SP are plugged in to the main model for correction. However, in this situation, the underlying assumption is that the estimates behave as if they were from an internal validation study. This may be a problematic strategy since the settings in the main and validation study are often different. Agbaje et al. [10] have shown that the misclassification obtained during field work and in a classical calibration situation is seldom similar due to differences between both settings. They also indicated that internal validation is almost impossible to attain in a large oral health survey, supporting the argument that we need to make use of external validation data and hence need to correct for misclassification in a different way. In this paper, a suggestion is made whereby the misclassification errors obtained from external validation data (which are usually available) are used to correct the main model.

The approach presented here is based on expressing SE and SP as a statistical model that depends on a rich set of covariates. Hence, the higher the number of (relevant) variables one can collect during validation exercises, the better the correction will be. From previous research, we have learned that the following variables are useful to record: dentition type, tooth type, surface type, examiner experience, and the position of the tooth in the mouth [10, 14]. It might be useful to collect following additional information: level of oral hygiene and type of restorations. The suggested idea of conditioning can be used in different kinds of studies, e.g., cross-sectional studies, longitudinal studies, case–control studies, etc. The approach suggested in this work considers a binary outcome. In principle, nothing changes in the approach when the outcome has more categories except that the correction procedures become more involved.

Other approaches to correct for misclassification have been suggested in literature. For example, when historical data or expert opinion is available about misclassification probabilities, prior information can be incorporated in a Bayesian framework to provide a prior distribution of the misclassification parameters [17]. Since the children in the validation study were selected from one single school, one

would possibly adjust for misclassification using the inverse probability selection weighting, a technique which is commonly used in survey sampling. The propensity score method is another approach that may be used when covariate information between groups differs [18].

In conclusion, in this research work, an external validation data set was used to correct for misclassification in the main data whereby SE and SP were modeled as a function of several covariates. This approach might be useful to make optimal use of available external validation data to correct for misclassification in the main model.

**Acknowledgments** This investigation was supported by Research Grant OT/05/60, KU Leuven; data collection was supported by Unilever, Belgium. The Signal Tandmobiel® project has the following partners: D. Declerck (Department of Oral Health Sciences, KU Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands and L-Biostat, KU Leuven), and K. Hoppenbrouwers (Youth Health Department, KU Leuven; Flemish Association for Youth Health Care).

**Conflict of interest** The authors declare no conflicts of interest.

## Appendix

### Multilevel model assuming error-free CE data

Let $Y_{\mathrm{M,\ stme}}$ be the true CE score of surface $s$, ($s$=1, …, $n_t$) nested in tooth $t$=1, …, $n_m$, which is nested in child/mouth $m$=1, …, $N_{\mathrm{M}}$ according to examiner $e$, ($e$=1, …, $n_e$) in the main study. The model uses $\pi_{\mathrm{stme}} = Pr\left(Y_{\mathrm{M,stme}} = 1|\beta, \mathbf{x}_{\mathrm{stme}},\ u_{\mathrm{m}},\ u_{\mathrm{tm}}, u_{\mathrm{e}}\right)$, which is the true conditional probability for CE on surface $s$ nested in tooth $t$ in mouth $m$ by examiner $e$. The multilevel logistic model for the true main data is then given by:

$$\mathrm{logit}(\pi_{\mathrm{stme}}) = \mathbf{x}_{\mathrm{stme}}^{T}\beta + u_{\mathrm{m}} + u_{\mathrm{tm}} + u_{\mathrm{e}}$$

where $\boldsymbol{x}_{\mathrm{stme}}$ represents the risk factors and/or determinants, $\boldsymbol{\beta}$ is a vector of regression coefficients and it quantifies the effect of the risk factors/or determinants. The quantities $u_{\mathrm{m}}$, $u_{\mathrm{tm}}$, and $u_{\mathrm{e}}$ are random intercepts at mouth, tooth, and examiner level and they are independently distributed with mean zero and variances $\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{tm}}^2,$ and $\sigma_{\mathrm{e}}^2$, at mouth, tooth (nested in mouth), and examiner level, respectively, i.e., $(u_{\mathrm{m}}, u_{\mathrm{tm}}, u_{\mathrm{e}}) \sim N(0, \boldsymbol{D})$ where $\mathbf{D} = \mathrm{diag}\left(\sigma_{\mathrm{m}}^2, \sigma_{\mathrm{tm}}^2, \sigma_{\mathrm{e}}^2\right)$. They take into account the clustering of teeth within mouths, surfaces within teeth, and an examiner recording many surfaces, respectively.

### Dealing with external validation data

There are two types of misclassification, i.e., differential and non-differential. Non-differential misclassification occurs when the misclassification does not depend on determinants [19, 20]. Differential misclassification occurs when misclassification is different in boys and girls.

If scoring in the main and validation studies is done by the same fallible dental examiners, then either the misclassification probabilities are the same in the two studies or the misclassification process is differential (misclassification depending on covariates), but given a rich set of covariates (e.g., gender, dentition type, tooth type, and surface type as for the ST validation study), they become the same in the two studies. If the misclassification probabilities between the two studies are equal, then the external validation data can be immediately used to correct for misclassification in the main model. However, if these misclassification probabilities do differ between the main data and validation data, then the misclassification process is differential and, conditional on a rich set of covariates, the scoring in the two data sets may become identical.

Consider an external validation data set as in the ST study. Assume $P_1$ and $P_2$ are the misclassification processes in the main study and validation study, respectively. Suppose that subjects are well characterized by a rich covariate vector, $\boldsymbol{z}$, then under the settings described above, our claim is that often $P_1\left(Y^*|Y,z\right) = P_2\left(Y^*|Y,z\right)$ when $P_1\left(Y^*|Y\right) \neq P_2\left(Y^*|Y\right)$. Inequality of the (assumed non-differential) misclassification process $P_1$ and $P_2$ occurs because $f_1(\boldsymbol{z}) \neq f_2(\boldsymbol{z})$ in $P_j\left(Y^*|Y\right) = \int P_j\left(Y^*|Y,\mathbf{z}\right) f_j(\mathbf{z})d(\mathbf{z})$ for $j$=1,2. As a result, the misclassification probabilities become identical given $\boldsymbol{z}$.

### Multilevel model for cross-sectional CE data adjusting for misclassification

Using the estimates of parameters for logistic models of SE and SP from validation data, say $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, a corrected multilevel logistic model for the main observed data uses $\pi_{\mathrm{stme}}^{*} = Pr\left(Y_{\mathrm{M,stme}}^{*} = 1|\beta, \mathbf{x}_{\mathrm{stme}}, u_{\mathrm{m}},\ u_{\mathrm{tm}}, u_{\mathrm{e}}, \boldsymbol{\alpha}, \boldsymbol{\eta}, \mathbf{z}\right)$, which is the observed (corrected for misclassification) probability for CE on surface $s$ nested in tooth $t$ in mouth $m$ from the main data set given $\boldsymbol{x}_{\mathrm{stme}}$ and $\boldsymbol{z}$, a vector of covariates from the main data and validation data, respectively, and random effects $u_{\mathrm{m}}$, $u_{\mathrm{tm}}$, and $u_{\mathrm{e}}$ and estimates of SE and SP $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, respectively. This processing was done in one joint model, i.e., a model that encompasses the estimation of SE and SP and at the same time correcting for misclassification. The corrected model is given by:

$$\pi_{\mathrm{stme}}^{*} = (1 - \tau_{00}) + [\tau_{11} + \tau_{00} - 1] \times \left[g^{-1}\left(\boldsymbol{x}_{\mathrm{stme}}^{T}\boldsymbol{\beta} + u_{\mathrm{m}} + u_{\mathrm{tm}} + u_{\mathrm{e}}\right)\right]$$

where $\tau_{11}$=$\tau_{11}(\boldsymbol{z})$ and $\tau_{00}$=$\tau_{00}(\boldsymbol{z})$ are the differential SE and SP.

Bayesian estimation approach

The posterior summary measures of the parameters are obtained using a sampling approach called the Markov Chain Monte Carlo (MCMC) approach [21]. Here, non-informative or vague priors were used which express that there is no prior information on the parameters. For this purpose, JAGS 3.1.0 [22] software was used. Three MCMC chains were run, each for 100,000 iterations for each model. The convergence of these MCMC chains was checked using the CODA package (see [23]) in R. In particular, the Gelman and Rubin diagnostics measure $\widehat{R}$ was used and this value was close to 1 for all the parameters, which means there was no evidence against convergence. Finally, a sensitivity analysis on the model corrected for misclassification was performed. Specifically, a sensitivity analysis was performed by changing the prior distributions for fixed effects. This was done in order to check whether the model was robust to some perturbations.

## References


1. Pine CM, Pitts NB, Nugent Z (1997) British Association for the Study of Community Dentistry (BASCD) guidance on the statistical aspects of training and calibration of examiners for surveys of child dental health: a BASCD coordinated dental epidemiology programme quality standard. Community Dent Health 14(Suppl 1):18–29
2. Neuhaus JM (1999) Bias and efficiency loss due to misclassified responses in binary regression. Biometrika 86:843–855
3. Neuhaus JM (2002) Analysis of clustered and longitudinal binary data subject to response misclassification. Biometrics 58:675–683
4. Magder LS, Hughes JP (1997) Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol 146(2):195–203
5. Mwalili S, Lesaffre E, Declerck D (2005) A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. J R Stat Soc Ser C Appl 54:77–93
6. Lesaffre E, Mwalili S, Declerck D (2004) Analysis of caries experience taking inter-observer bias and variability into account. J Dent Res 83(12):951–955
7. Tenenbein A (1986) A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. Biometrika 73:13–22
8. Marshall RJ (1990) Validation study methods for estimating exposure proportions and odds ratios with misclassified data. J Clin Epidemiol 43:941–947
9. Küchenhoff H (2009) Misclassification and measurement error in oral health. In: Lesaffre E, Feine J, Leroux B, Declerck D (eds) Statistical and methodological aspects of oral health research. Wiley, New York, pp 279–290
10. Agbaje JO, Mutsvari T, Lesaffre E, Declerck D (2012) Examiner performance in calibration exercises compared with field conditions when scoring caries experience. Clin Oral Investig 16(2):481–488
11. Declerck D, Lesaffre E, Leroy R, Vanobbergen J (2009) Examples from oral health epidemiology: the Signal Tandmobiel and smile for life studies. In: Lesaffre E, Feine J, Leroux B, Declerck D (eds) Statistical and methodological aspects of oral health research. Wiley, New York, pp 341–357
12. Pitts NB, Evans DJ, Pine CM (1997) British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys–1996/7. Community Dent Health 14(Suppl 1):6–9
13. Klein H, Palmer CE, Knutson JW (1938) Studies on dental caries. I. Dental status and dental needs of elementary school children. Public Health Rep 53:751–765
14. Mutsvari T (2012) Misclassification in multilevel models with applications in dental caries research, PhD Dissertation, KU Leuven
15. Brenner H, Savitz DA (1990) The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case–control studies. Am J Epidemiol 132(1):181–192
16. Wacholder S, Armstrong B, Hartge P (1993) Validation studies using an alloyed gold standard. Am J Epidemiol 137:1251–1258
17. McInturf P, Johnson WO, Cowling D, Gardner IA (2004) Modelling risk when binary outcomes are subject to error. Stat Med 23:1095–1109
18. Ralph BD (1998) Propensity score methods for bias reduction in the comparison of treatment to non-randomized control group. Stat Med 17:2265–2281
19. Gustafson P (2004) Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. Chapman & Hall, London
20. Greenland S (1980) The effects of misclassification in the presence of covariates. Am J Epidemiol 112:564–569
21. Lesaffre E, Lawson B (2012) Bayesian biostatistics. Wiley, New York
22. Plummer M (2011) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: Hornik K, Leisch F, Zeileis A (eds) Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC, 2003). Technische Universitaet Wien, Vienna, Austria. http://www.ci.tuwien.ac.at/Conferences/DSC.html. Accessed 24 Nov 2011
23. Plummer M, Best N, Cowles K, Vines K (2008) CODA: output analysis and diagnostics for MCMC. R Package version 0.13-3