Original Paper

Caries Research

Caries Res 2010;44:438–444
DOI: 10.1159/000319539

Received: December 13, 2009
Accepted after revision: July 21, 2010
Published online: September 13, 2010

Factors That Influence Data Quality in Caries Experience Detection: A Multilevel Modeling Approach

T. Mutsvari[a] E. Lesaffre[a, d] M.J. García-Zattera[a, c] L. Diya[a] D. Declerck[b]

[a]L-BioStat and [b]School of Dentistry, Catholic University of Leuven, Leuven, Belgium; [c]Department of Statistics, Pontifical Catholic University of Chile, Santiago, Chile; [d]Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands

Key Words
Calibration · Multilevel · Sensitivity · Specificity

Abstract
Caries experience detection is prone to misclassification. For this reason, calibration exercises which aim at assessing and improving the scoring behavior of dental raters are organized. During a calibration exercise, a sample of children is examined by the benchmark scorer and the dental examiners. This produces a 2 × 2 contingency table with the true and possibly misclassified responses. The entries in this misclassification table allow to estimate the sensitivity and the specificity of the raters. However, in many dental studies, the uncertainty with which sensitivity and specificity are estimated is not expressed. Further, caries experience data have a hierarchical structure since the data are recorded for the surfaces nested in the teeth within the mouth. Therefore, it is important to report the uncertainty using confidence intervals and to take the clustering into account. Here we apply a Bayesian logistic multilevel model for estimating the sensitivity and specificity. The main goal of this research is to find the factors that influence the true scoring of caries experience accounting for the hierarchical structure in the data. In our analysis, we show that the dentition type and tooth or surface type affect the quality of caries experience detection.

Dental caries is one of the most prevalent chronic diseases that have spread worldwide affecting persons in all age groups. Many epidemiological surveys and clinical studies are carried out to obtain a further understanding of this disease entity. However, the process of detecting the presence of caries experience (CE) is not an obvious issue, affecting the quality of the obtained data. In order to standardize data collection techniques in epidemiological surveys, but also in clinical trials, CE assessment guidelines were developed by the World Health Organization [WHO, 1997] and the British Association for the Study of Community Dentistry (BASCD) [Pine et al., 1997] and more recently the International Caries Detection and Assessment System [ICDAS, 2005] was proposed. All of these underline the need for training the examiners and measuring the reliability of the obtained scores. However, the process of CE detection remains prone to misclassification error even after adhering to these standardized criteria, for the following reasons. Firstly, the circumstances or conditions in which the examinations are carried out may have an impact on the outcome of scoring CE, e.g. light conditions [Assaf et al., 2004; Kassawara et al., 2007], type and quality of the tools used [WHO, 1997; Pitts, 2001], the physical position of the patient during examination or the time available to do the examination. Secondly, factors that are related to the examiner may also influence the disease

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel
0008–6568/10/0445–0438$26.00/0

Accessible online at:
www.karger.com/cre

Emmanuel Lesaffre
L-BioStat
Kapucijnenvoer 35, Block D, Bus 7001
BE–3000 Leuven (Belgium)
Tel. +32 16 336896, Fax +32 16 337015, E-Mail Emmanuel.Lesaffre@med.kuleuven.be

detection, e.g. the examiner's experience [Heifetz et al., 1985; Poorterman et al., 1997], whether the examiner is right or left handed and the examiner's visual capacities. Lastly, characteristics of the patient under examination can also have an impact, e.g. the presence of plaque which may hide signs of CE [Assaf et al., 2004], the level of cooperation of the patient, tooth-colored fillings may hamper a correct detection but also tooth position, and dentition type and surface type may affect the quality of the data.

Common measures that are used to assess the reliability of the data are sensitivity (SE) and specificity (SP) of the raters vis-à-vis a gold standard (or a benchmark scorer in the absence of a gold standard). A gold standard is an instrument (or an examiner) who is 100% error free, while a benchmark is an experienced examiner or a tested measuring instrument which is assumed to be error free or nearly so. 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 people with dental CE who are identified as having the condition). SP measures the proportion of negatives which are correctly identified (e.g. the proportion of people without dental CE who are identified as not having the condition). Estimates of SE and SP are usually obtained from calibration exercises where a sample of children is examined by the benchmark scorer and the raters producing a 2 × 2 contingency table with the recordings from both the benchmark and the raters. Confidence intervals (CI) of SE and SP express the uncertainty with which these parameters are estimated from a sample.

To evaluate the factors that determine SE and SP, simple logistic regression models can be fitted. However, when the data are clustered, then the logistic model needs to be extended further, the reason being that the variance of the parameters is inflated due to the correlation between observations in the same cluster since they share similar characteristics [see e.g. William and Nan, 2006]. Therefore, approaches that account for this correlation should be applied. A common approach to deal with correlated data is the generalized estimating equations (GEE) approach [see e.g. Liang and Zeger, 1986]. The GEE approach is based on a marginal model where the parameters have a population average interpretation while accounting for the clustering effect in the data. In the GEE approach, an assumed working correlation is specified, but this correlation is treated as nuisance in the estimation. The parameters obtained then are consistent even if this working correlation is wrongly specified [see e.g. Liang and Zeger, 1986]. The GEE approach for estimating SE and SP with clustered data was first applied by Smith and Hadgu [1992]. If also the correlation in the data is of interest, then a multilevel model is recommended [see e.g. Leyland and Goldstein, 2001]. In contrast to the simple logistic regression and GEE, the parameters in a multilevel model have a subject-specific interpretation. For an application of the multilevel approach in the field of oral health, we refer to Burnside et al. [2007].

The main goal of the present contribution is to investigate the factors that have an impact on the true scoring of CE (measured by SE and SP) while taking the data clustering into account. In order to do this, we apply a multilevel logistic model. This model was applied to the data that were obtained from the calibration exercises during the Signal Tandmobiel® study [Vanobbergen et al., 2000].

Materials and Methods

Epidemiological Dataset

The Signal Tandmobiel project is a longitudinal (1996–2001) oral health project in Flanders (North of Belgium). At the first examination, the average age of the children was 7.1 years (standard deviation = 0.4) and varied from 6.12 to 8.09 years. For this project, 16 trained dentists (examiners) conducted annual examinations of 4,468 children (2,315 boys and 2,153 girls) from 179 primary schools, after parental consent had been obtained. Data on oral hygiene and dietary habits were obtained through structured questionnaires, completed by the parents. The children received a clinical examination using the standardized and widely accepted criteria as recommended by the WHO [1987] and based on the diagnostic criteria for caries prevalence surveys published by the British Association for the Study of Community Dentistry [Pine et al., 1997]. The clinical examinations took place in a mobile dental clinic, with a standard dental chair and dental artificial light. Detection was performed by visual-tactile method, using a disposable mouth mirror, visual-tactile and a WHO/CPITN type E probe. No radiographs were taken. For a more detailed description of the Signal Tandmobiel® study we refer to Vanobbergen et al. [2000].

Calibration Data

Training sessions for scoring CE were organized and the scoring behavior of each of the 16 dental examiners was compared to that of the benchmark scorer (last author). The benchmark scorer was trained by a BASCD trainer in 1990. It is recommended that once every 2 years, the benchmark scorers or trainers from different districts should meet to undertake a national training and calibration amongst themselves, see the BASCD [Pine et al., 1997]. However, this was not possible in our case due to logistical constraints. During the study period (1996–2001), 3 calibration exercises for scoring CE (1996, 1998, 2000), involving 92, 32 and 24 children, respectively, were organized. A large number of children was involved in 1996 compared to the other years. Four sessions were organized in 1996 with each session comprising approximately 25 children. For practical reasons, a more efficient organization was needed with only a single session for the years 1998 and 2000. Note that the age of the children for the calibration exercises of 1998 and 2000 was not recorded in the database. However, the ages of the children examined in 1996, 1998 and 2000 were age-matched with the school children in the 1st, 3rd and 5th class, respectively. During the calibration exercises, the children were not sampled at random from the main study. Rather, a school was selected where a relatively high prevalence of CE could be expected. At the end of each of the 3 calibration exercises, the SE and SP of each dental examiner were determined. In the present work, data of the 3 calibration exercises were combined. A multilevel logistic model specified in the next section was fitted to this calibration data set.

Uncertainty on Estimation of SE and SP

The traditional way of presenting the uncertainty of an estimate is through its CI. A CI gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. The width of the CI gives us an idea about how uncertain we are about the unknown parameter. Consider a 2 × 2 table below where G (0 = no caries and 1 = caries) represents the gold standard or benchmark score, E (0 = no caries and 1 = caries) is the score given by the examiner and n_{ij} is the number of subjects with score i according to the examiner and j according to the benchmark scorer:

$$\begin{array}{cc} & \begin{array}{cc} G=0 & G=1 \end{array} \\ \begin{array}{c} E=0 \\ E=1 \end{array} & \begin{pmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{pmatrix} \end{array}$$

The naive estimates of SE and SP of the examiner are

$$\hat{\tau}_{11} = \frac{n_{11}}{n_{11}+n_{01}} \text{ and } \hat{\tau}_{00} = \frac{n_{00}}{n_{00}+n_{10}},$$

respectively, with corresponding variances

$$\widehat{var}(\hat{\tau}_{11}) = \frac{\hat{\tau}_{11}(1-\hat{\tau}_{11})}{(n_{11}+n_{01})} \text{ and } \widehat{var}(\hat{\tau}_{00}) = \frac{\hat{\tau}_{00}(1-\hat{\tau}_{00})}{(n_{00}+n_{10})}.$$

Note that naive refers to the estimates that assume that the observations are independent. In calculating the CI of an estimate, we assume that the estimate follows a normal distribution for large samples. Hence the 100(1 – α)% CI is estimated by:

$$\left[\hat{\gamma} - z_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\gamma})}; \hat{\gamma} + z_{1-\alpha/2}\sqrt{\widehat{var}(\hat{\gamma})}\right], \tag{1}$$

where $\hat{\gamma}$ is the estimate of SE or SP and $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

The methodology for estimating variance and CI explained above assumes that the data are independent. However, this naive binomial approach to estimate variance tends to underestimate the variance if there is a positive correlation between responses and this results in too narrow CI, hence a possibility of inaccurate conclusions [see e.g. William and Nan, 2006]. Due to this shortfall, approaches such as GEE are opted for in order to obtain estimates with corrected variances and CI. Under the GEE approach, the most commonly used working correlations are (i) exchangeable, which assumes that any 2 observations in the same cluster have the same correlation; (ii) autoregressive, which indicates that 2 observations taken close in time within an individual tend to be more correlated than 2 observations from the same individual taken far apart in time, and (iii) independence, which is similar to the simple logistic model. The GEE is fitted using the SAS® procedure GENMOD (SAS version 9.2). The multilevel model is another way to deal with clustering. In this approach we are further interested in making inferences on the random effects which explain the variability due to the clustering. In multilevel models, the total variance in the outcome or response variable is decomposed into variances of random effects associated with each level. For example in the analysis of CE, we decompose the total variance into mouth and tooth level variability. The intracluster correlation coefficient is used to measure the proportion of variance attributed to each level of the data.

Statistical Modeling

The SE and SP of scoring CE can be affected by many factors. In order to investigate the impact of these factors one can use a simple logistic model. This model assumes independence of the data. However, the CE for surfaces of the same tooth are correlated since they are exposed to similar characteristics, similarly for the teeth that belong to one mouth resulting in a hierarchical structure. In the first part of this section, we will introduce the multilevel model. We will then describe the estimation of the parameters for this model in the second part of this section. Note that in this section we illustrate the multilevel model for SE only. The model for SP has a similar structure.

Multilevel Logistic Model Specification

Our methodology 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 binary, a popular model to apply is logistic regression. In this section we describe the multilevel logistic model for SE, which is an extension of the simple logistic model. The model uses π_{stme}, which is the SE of scoring CE on surface *s* within tooth *t* and mouth *m*, by examiner *e*. More details of the model are given in the appendix. The multilevel logistic model relating *p* regressors to the SE is given by:

$$logit(\pi_{stme}) = \\ \beta_0 + \beta_1 x_{1,stme} + \beta_2 x_{2,stme} + ... + \beta_p x_{p,stme} + u_m + u_{tm} + u_e \tag{2}$$

where $x_{1,stme}$, $x_{2,stme}$, ..., $x_{\mathrm{p},stme}$ are the covariates with the associated regression coefficients $\beta = (\beta_0, \beta_1, ..., \beta_p)$. The quantities u_m, u_{tm} and u_e are random intercepts at mouth, tooth (nested in mouth) and examiner level, respectively.

The binary covariates included in our study were: dentition type (deciduous = 0, permanent = 1), position in the mouth (2 variables were used), i.e. jaw (lower = 0, upper = 1) and quadrant (left = 0, right = 1). The nominal covariates included were: tooth type (incisor = 0, canine = 1, premolar = 2, molar = 3), surface type (buccal = 0, distal = 1, mesial = 2, lingual = 3, occlusal = 4) and year (1996 = 0, 1998 = 1, 2000 = 2). The nominal covariates were recoded into binary covariates when used in the logistic regression models.

Table 1. Estimates (percent) of SE and SP with the 95% CI using naive, GEE (independence, autoregressive and exchangeable) and 95% CI for multilevel approaches

Level	Naive	GEE			Multi-level
		Ind	AR(1)	Exch	
Sensitivity					
Surface	79.2 (77.0–81.2)	79.2 (75.4–82.5)	78.5 (74.7–81.8)	75.0 (71.2–78.5)	85.8 (77.2–91.7)
Tooth	82.8 (79.8–85.4)	82.8 (78.1–86.7)	82.7 (78.1–86.5)	81.3 (76.4–85.4)	87.0 (80.3–92.7)
Mouth	99.6 (97.0–99.9)	99.6 (97.0–99.9)	99.6 (97.0–99.9)	99.6 (97.0–99.9)	99.6 (97.0–99.9)
Specificity					
Surface	99.0 (98.9–99.1)	99.0 (98.8–99.2)	99.0 (98.7–99.1)	98.9 (98.7–99.1)	99.96 (99.92–99.98)
Tooth	97.7 (97.2–98.2)	97.7 (97.1–98.2)	97.7 (97.1–98.2)	97.7 (97.2–98.2)	98.43 (97.86–98.93)
Mouth	84.6 (54.9–96.1)	84.6 (54.9–96.1)	84.6 (54.9–96.1)	84.6 (54.9–96.1)	84.6 (54.9–96.1)

Table 2. Parameter estimates of the multilevel logistic model for SE and SP (using WinBUGS program 1.4.3)

Parameter		Sensitivity			Specificity		
		mean (SD)	2.5%	97.5%	mean (SD)	2.5%	97.5%
Fixed effects	intercept	–2.001 ± 1.410	–4.741	0.758	10.864 ± 0.932	9.183	12.850
Dentition type	permanent	–1.113 ± 0.512	–2.155	–0.139	0.768 ± 0.342	0.138	1.443
	deciduous	–	–	–	–	–	–
Tooth type	canine	–0.035 ± 1.533	–2.952	2.991	–1.741 ± 0.736	–3.302	–0.338
	molar	3.101 ± 1.299	0.676	5.675	–1.291 ± 0.797	–2.987	0.157
	pre-molar	3.419 ± 1.532	0.608	6.536	–1.291 ± 0.797	–2.987	0.157
	incisor	–	–	–	–	–	–
Surface	distal	1.563 ± 0.390	0.847	2.340	0.183 ± 0.249	–0.311	0.688
	mesial	0.734 ± 0.346	0.091	1.417	–0.023 ± 0.232	–0.476	0.420
	lingual	0.103 ± 0.377	–0.639	0.839	–0.101 ± 0.228	–0.543	0.343
	occlusal	0.875 ± 0.337	0.265	1.562	–1.486 ± 0.226	–1.920	–1.037
	buccal						
Year	1998	1.006 ± 0.703	–0.463	2.359	–0.994 ± 0.543	–2.078	0.004
	2000	0.698 ± 0.880	–0.922	2.374	–1.847 ± 0.622	–3.160	–0.595
	1996						
Jaw	upper	–0.424 ± 0.353	–1.186	0.240	0.434 ± 0.281	–0.069	1.025
	lower	–	–	–	–	–	–
Quadrant	right	0.358 ± 0.446	–0.495	1.258	0.269 ± 0.368	–0.452	1.015
	left	–	–	–	–	–	–
Random effects	σ^2_{mouth}	2.678 ± 1.045	1.096	5.070	2.104 ± 0.680	1.006	3.645
	σ^2_{tooth}	3.810 ± 1.105	2.105	6.386	6.428 ± 1.173	4.626	9.279
	$\sigma^2_{examiner}$	0.664 ± 0.437	0.159	1.788	0.304 ± 0.212	0.038	0.862
ICC	mouth level	0.274			0.180		
	tooth level	0.390			0.544		

Mean = Posterior mean; SD = posterior SD; 2.5%, 97.5% = 95% credible intervals; ICC = intracluster correlation coefficient.

Random effects regression models are popular when analyzing clustered data. An important feature of this model is that inference can be made on the variability at each level in the data hierarchy. Further, in multilevel models, the intracluster correlation coefficient can be calculated to measure the proportion of variance which is attributable to each level in the data. For example, at mouth level for SE the intracluster correlation coefficient can be calculated as $\sigma_m^2/(\sigma_m^2 + \sigma_t^2 + \pi^2/3)$. The value $\pi^2/3$ is the approximate variance of the logistic regression model assuming that the underlying continuous variable follows a logistic distribution.

Estimation of Parameters

To estimate the parameters in model (2), a Bayesian approach will be used. In a Bayesian approach the prior knowledge about the parameters is combined with the observed data (likelihood) to yield the posterior distribution. The posterior summary measures of the parameters are obtained using a sampling approach called the Markov-Chain Monte Carlo (MCMC) approach [see e.g. Spiegelhalter et al., 1996]. Here we used noninformative or vague priors which express that we do not have prior information on the parameters. A popular software to perform the MCMC calculations is WinBUGS 1.4.3 [Lunn et al., 2000]. This software was also used here. Three MCMC chains were run, each for 100,000 iterations for all the models. The convergence of these MCMC was checked using the CODA package [see Plummer et al., 2008] in R. In particular, we used the Gelman and Rubin diagnostics measure $\hat{R}$ and this value was close to 1 for all the parameters, which means there was no evidence against convergence. Technical details of parameter estimation are given in the appendix.

Results

A total of 4,797 deciduous tooth surfaces was involved in the analysis with 424 (8.8%) showing CE according to the benchmark scorer. Out of the 4,944 permanent tooth surfaces, 95 (1.9%) showed CE according to the benchmark scorer. In our analysis, we first considered the estimates of SE and SP using different methods without including any covariates. The naive, GEE and multilevel model estimates (across examiners) of SE and SP with the 95% CI are shown in table 1.

We then included covariates in our analysis starting with a simple logistic model for estimating the SE and SP using a Bayesian approach (results not shown). These estimates were used as starting values for the estimation of parameters in the multilevel model. The results of the multilevel model for the SE of scoring CE are shown in table 2. In this table, a positive estimate reflects a higher SE of scoring CE compared to the reference level for categorical variables, e.g. for the variable tooth type, the scoring of CE on incisor teeth is taken as the reference. For a continuous variable a positive estimate reflects an increase in the SE with a unit increase in that variable. Negative estimates reflect the opposite, i.e. a lower SE. The results for SP in table 2 have a similar interpretation. The last column shows the 95% credible interval for the regression coefficients. Note that an estimate is considered to be statistically significant in a Bayesian sense if its 95% credible interval does not contain 0.

There was no significant effect of the SE of scoring CE on the variables year, jaw and quadrant. However, there was a significantly lower SE of scoring CE on permanent teeth as compared to the deciduous teeth. Pertaining to the tooth type, there is a significantly higher SE of scoring CE on the molars and premolars as compared to the incisor teeth. Also, the distal, mesial and occlusal surfaces demonstrate a higher SE of scoring CE than the buccal surface. The variance estimates of the random effects show that there was more variability at tooth level than mouth level, which resulted in a higher intracluster correlation coefficient value. The estimate of the intracluster correlation coefficient shows that 39% of the variance was attributable to the variation between teeth within individuals, and 27% was attributable to variation at mouth level. This means that the correlation of any pair of surfaces belonging to one single tooth was higher compared to the correlation between any pair of teeth that belongs to one single mouth.

The results of the multilevel model for SP are shown in table 2. A nonsignificant effect on the SP of scoring CE of the variables jaw and quadrant was observed. The process of scoring CE on permanent teeth shows a significantly higher SP than for the deciduous teeth. The scoring of CE on canines and molars shows a lower SP when compared to the incisors. One can also see the significantly lower SP of scoring on occlusal surfaces compared to the buccal surfaces. In contrast to the SE, there is a significantly lower SP for scoring CE in the calibration exercise held in the year 2000 than in 1996. More variability was observed at tooth level compared to the mouth level for SP.

Discussion

To illustrate the effect of data dependency, we estimated the SE and SP (without covariates) using the naive, GEE and multilevel approaches. The GEE and multilevel models are the most commonly used approaches for discrete repeated or clustered measurements and these models can be viewed as direct extensions of the general linear models for independent observations. Although there are similarities between the marginal and multilevel specifications, the 2 approaches often produce different results [see Molenberghs and Verbeke, 2005]. The parameters in both models have completely different interpretations. When one fits a marginal model, then the parameter estimates obtained characterize the population average. In our case, the GEE and naive estimates are interpreted as the population average SE or SP. However, when one fits a multilevel model, then the parameters are specific to a subject, for example in a multilevel logistic model for SE and SP where the estimates are conditioned upon the level of the subject-specific effect or random effects. This phenomenon is shown in the results where there are discrepancies between the marginal and multilevel approaches. This is to be expected since the multilevel estimates depend on the variability at each level. In case the random intercepts variability is large, the parameters obtained from fitting the marginal and multilevel models are very different, while equal parameter values result if the variance of the random effects is 0. Further, the cluster size, e.g. the number of teeth or surfaces with CE within a subject, can be related to the outcome of interest (e.g. SE or SP), and this is referred to as nonignorable or informative cluster size and can influence the parameter estimates in models for clustered data [see Williamson et al., 2003]. This aspect is the subject of further research.

In our analysis, we observed that the 95% CI for GEE and multilevel are wider as compared to the naive approach since they account for the correlation in the data. One can see an increase in the SP from mouth to surface level. This is explained by the fact that if there is no CE at mouth level, there is no CE at tooth and surface level. On the other hand, we observed an increase in SE from surface to mouth level. Indeed, if there is CE at surface level, CE at tooth and mouth level is implied. While this looks plausible, there is no strict order of the results for SE. We refer to Lesaffre et al. [2009] for a detailed explanation. Note that the estimates of SE and SP in table 1 at mouth level remained constant regardless of any method applied since there is no clustering at this level.

At the time of examination for this study, most of the children had a mixed dentition type. In line with this, we discovered using the multilevel model (with covariates) that scoring CE on the permanent teeth had a significantly lower SE than on the deciduous teeth. This is possibly due to the more complex anatomy of the permanent teeth. The results also show that it is easier to detect CE on molars and premolars when it is truly present as compared to the incisors. This is possibly due to the higher CE prevalence encountered in molars and premolars than is the case in incisors. It is also interesting to see that the examiners' performance did not significantly change between 1996 and 2000, based on SE. The results for SP show that it is easier to detect non-CE on permanent teeth when it is truly non-CE compared to the deciduous teeth. Further, molars and canines are more difficult to score as non-CE when they are truly non-CE compared to the incisors. There was relatively low variability among the examiners. This is not unexpected since the examiners used a standardized procedure to do the examination, thereby introducing some homogeneity in the data.

Many epidemiological and clinical surveys involve a large number of subjects to be examined. For this reason, often several examiners are involved in the collection of data. Estimates of examiners' specific SE and SP are often used to monitor the quality of the data. Traditionally, however, researchers assume that the data are independent when estimating these parameters even when the data have a more complex structure. In this paper we have investigated the effects of several factors on scoring CE using a logistic multilevel model. The multilevel model corrects for the bias in the variance estimates when there is clustering in the data. The higher the observations are correlated in the cluster, the more likely that ignoring the clustering will result in biasedly estimated variances. We used a Bayesian approach since this can incorporate prior knowledge in the analysis if available. Also, the Bayesian software provides a more flexible way to fit more complex models. For example, we can include random effects for more levels in our model without too much programming effort. In summary, we applied multilevel modeling to CE scoring data and we hope that our results might be useful for the improvement in quality of the caries experience data.

Acknowledgements

This investigation was supported by research grant OT/05/60, Catholic University Leuven; the data collection was supported by Unilever, Belgium. The Signal Tandmobiel project has the following partners: D. Declerck (Dental School, Catholic University of Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Dental School, University of Ghent), P. Bottenberg (Dental School, University of Brussels), E. Lesaffre (Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands, and L-Biostat, Catholic University of Leuven, Leuven, Belgium) and K. Hoppenbrouwers (Youth Health Department, Catholic University of Leuven; Flemish Association for Youth Health Care). The third author is supported by the National Scholarship for Doctoral Studies 2009, Conicyt (Chile). We acknowledge Jimoh Olubanwo Agbaje for the medical explanations of the data and results.

Appendix

Multilevel Logistic Model

Let Y^{*}_{stme} be a binary outcome for surface s, ($s = 1, ..., n_t$) nested in tooth t ($t = 1, ..., n_m$), which is nested in mouth m ($m = 1, ..., N$) according to examiner e, ($e = 1, ..., n_e$). The corresponding binary outcome according to the benchmark scorer is Y_{stm}.

Therefore,

$$\pi_{stme} = \pi_{stme}\,(\mathbf{x}_{stme}, \mathbf{u}_{tme}) = Pr(Y^{*}_{stme} = 1 | Y_{stm} = 1, \mathbf{x}_{stme}, \mathbf{u}_{tme})$$

is the SE. Mouth, tooth and examiner random effects for SE have variances σ^2_m, σ^2_t and σ^2_e, respectively, and we refer to this set of variances as **D**.

Estimation of Parameters

Let $\theta = (\beta, \mathbf{D})$ be the parameters of interest pertaining to model (2). The likelihood contribution (conditioned on the random effects) of tooth t in mouth m according to examiner e for 1 subject is given by:

$$\begin{aligned} L_{tme}\left(\beta, \mathbf{D} | \mathbf{y}^{*}_{tme}, \mathbf{y}_{tm}, \mathbf{u}_{tme}\right) &= p\left(\mathbf{y}^{*}_{tme} | \beta, \mathbf{D}, \mathbf{y}_{tme}, \mathbf{u}_{tme}\right) \\ &= \prod_{t=1}^{n_m}\prod_{e=1}^{n_e}\prod_{s=1}^{n_t} \pi\left(\mathbf{x}_{stme}, \mathbf{u}_{stme}\right)^{y^{*}_{stme}} \left(1 - \pi\left(\mathbf{x}_{stme}, \mathbf{u}_{stme}\right)\right)^{1 - y^{*}_{stm}} \end{aligned} \qquad (3)$$

Note that n_t represents the total number of surfaces for a particular tooth, n_e represents the total number of examiners who screened that particular mouth and n_m represents the total number of teeth in the mouth.

The total conditional likelihood is given by:

$$L_c\left(\beta, \mathbf{D} | \mathbf{y}^{*}_{tme}, \mathbf{y}_{tm}\right) = \prod_{m=1}^{N} L_{tme}\left(\beta, \mathbf{D} | \mathbf{y}^{*}_{tme}, \mathbf{y}_{tm}\right). \qquad (4)$$

A marginal likelihood $L(\beta, \mathbf{D} \mid \mathbf{y}^{*}, \mathbf{y})$ is therefore obtained by integrating out the random effects in the conditional likelihood above.

Priors for Model (2)

The marginal likelihood $L(\beta, \mathbf{D} \mid \mathbf{y}^{*}, \mathbf{y})$ is combined with the prior knowledge (prior probability) on the parameters β and **D** to update the information about them (posterior).

(1) For the regression coefficients, $\beta_0, ..., \beta_p$, vague independent priors were assumed to follow a normal distribution with mean 0 and large variance, i.e. $\beta_i \sim N(0, 10^6)$, $i = 0, ..., p$.

(2) The prior distribution for each of the standard deviations (σ_m, σ_t, σ_e) of the random effects was taken as uniform, i.e. U[0,100]. The motivation for using a uniform distribution for the prior distribution on the standard deviation is discussed in e.g. Gelman and Hill [2007].

References

Assaf AV, Meneghim MC, Zanin L, Mialhe FL, Pereira AC, Ambrosano GM: Assessment of different methods for diagnosing dental caries in epidemiological surveys. Community Dent Oral Epidemiol 2004;32:418–425.

Burnside G, Pine CM, Williamson PR: The application of multilevel modelling to dental caries data. Stat Med 2007;26:4139–4149.

Gelman A, Hill J: Data Analysis Using Regression and Multilevel/Hierachical Models. Cambridge, Cambridge University Press, 2007.

Heifetz SB, Brunelle JA, Horowitz HS, Leske GS: Examiner consistency and group balance at baseline of a caries clinical trial. Community Dent Oral Epidemiol 1985;13:82–85.

International Caries Detection and Assessment System Coordinating Committee 2005 Criteria Manual: International Caries Detection and Assessment System (ICDAS II), 2005.

Kassawara AB, Assaf AV, Meneghim MC, Pereira AC, Topping G, Levin K, Ambrosano GM: Comparison of epidemiological evaluations under different caries diagnosis thresholds. Oral Health Prev Dent 2007;5:137–144.

Lesaffre E, Küchenhoff H, Mwalili S, Declerck D: On the estimation of the misclassification table for finite count data with an application in caries research. Statist Modelling 2009;9: 99–118.

Leyland AH, Goldstein H: Multilevel Modelling of Health Statistics. Hoboken, Wiley & Sons, 2001.

Liang KY, Zeger SL: Longitudinal data analysis using generalized estimating equation models. Biometrika 1986;73:13–22.

Lunn D, Thomas A, Best N, Spiegelhalter D: WinBUGS – A Bayesian modelling framework: concepts, structure, and extensibility. Statist Computing 2000;10:325–337.

Molenberghs G, Verbeke G: Models for Discrete Longitudinal Data. Berlin, Springer, 2005.

Pine C, Pitts N, Nugent Z: 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 1997; 14(suppl 1):18–29.

Pitts NB: Clinical diagnosis of dental caries: a European perspective. J Dent Educ 2001;65: 972–978.

Plummer M, Best N, Cowles K, Vines K: Coda – output analysis and diagnostics for MCMC, R Package Version 0.13-3, 2008.

Poorterman JH, Verheij JG, Kieft JA, Eijkman MA: Variations among dentists in the diagnosis of caries and assessment of dental restorations. Ned Tijdschr Tandheelkd 1997; 104:214–218.

Smith P, Hadgu A: Sensitivity and specificity for correlated observations. Stat Med 1992;11: 1503–1509.

Spiegelhalter D, Thomas A, Best N, Gilks W: Bayesian Inference Using Gibbs Sampling Manual. Cambridge, 1996.

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. Eur J Paediatr Dent 2000;2:87–96.

WHO: Oral Health Surveys: Basic Methods, ed 3 (public health report). Geneva, World Health Organization, 1987.

WHO: Oral Health Surveys: Basic Methods, ed 4 (public health report). Geneva, World Health Organization, 1997.

William FM, Nan G: Estimation of sensitivity and specificity of clustered binary data, statistics and data analysis, SUGI 31 proceedings. SAS Proceedings, 2006.

Williamson JM, Datta S, Satten GA: Marginal analyses of clustered data when cluster size is informative. Biometrics 2003;59:36–42.