Clin Oral Invest (2012) 16:481–488
DOI 10.1007/s00784-011-0523-1

ORIGINAL ARTICLE

Examiner performance in calibration exercises compared with field conditions when scoring caries experience

Jimoh Olubanwo Agbaje · Timothy Mutsvari · Emmanuel Lesaffre · Dominique Declerck

Received: 27 July 2010 / Accepted: 4 February 2011 / Published online: 23 February 2011

Abstract The objective of this study was to verify how valid misclassification measurements obtained from a 'pre-survey' calibration exercise are by comparing them to validation scores obtained in 'field' conditions. Validation data were collected from the '*Smile for Life*' project, an oral health intervention study in Flemish children. A calibration exercise was organized under 'pre-survey' conditions (32 age-matched children examined by eight examiners and the benchmark scorer). In addition, using a pre-determined sampling scheme blinded to the examiners, the benchmark scorer re-examined between six and 11 children screened by each of the dentists during the survey. Factors influencing sensitivity and specificity for scoring caries experience (CE) were investigated, including examiner, tooth type, surface type, tooth position (upper/lower jaw, right/left side) and validation setting (pre-survey versus field). In order to account for the clustering effect in the data, a generalized estimating equations approach was applied. Sensitivity scores were influenced not only by the calibration setting (lower sensitivity in field conditions, $p<0.01$), but also by examiner, tooth type (lower sensitivity in molar teeth, $p<0.01$) and tooth position (lower sensitivity in the lower jaw, $p<0.01$). Factors influencing specificity were examiner, tooth type (lower specificity in molar teeth, $p<0.01$) and surface type (the occlusal surface with a lower specificity than other surfaces) but not the validation setting. Misclassification measurements for scoring CE are influenced by several factors. In this study, the validation setting influenced sensitivity, with lower scores obtained when measuring data validity in 'field' conditions. Results obtained in a pre-survey calibration setting need to be interpreted with caution and do not (always) reflect the actual performance of examiners during the field work.

Keywords Caries experience · Calibration setting · Misclassification · Sensitivity · Specificity

J. O. Agbaje · D. Declerck (✉)
School of Dentistry, Oral Pathology and Maxillofacial Surgery,
Catholic University Leuven,
Kapucijnenvoer 7 blok a bus 7001,
3000 Leuven, Belgium
e-mail: dominique.declerck@med.kuleuven.be

T. Mutsvari · E. Lesaffre
L-Biostat, Catholic University Leuven,
U.Z. St. Rafael Kapucijnenvoer 35,
3000 Leuven, Belgium

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

Introduction

In oral health surveys, calibration exercises are organized in order to guarantee the reliability of the data obtained by different examiners. For this purpose, most guidelines for recording caries experience (CE) include instructions on the organization and interpretation of calibration sessions [1–3]. Based on data obtained from these exercises, the level of agreement between scores obtained by the examiners and a benchmark examiner (inter-examiner agreement) and/or between repeated examinations of individuals by the same examiner (intra-examiner agreement) can be assessed [2–5].

However, calibration exercises for the assessment of inter-examiner agreement are often organized in circumstances widely different from the conditions experienced by the dental examiner during the field work. In most cases,

these exercises consist of the examination of a rather small group of subjects by the different examiners and a benchmark scorer (if applicable). Often these examinations are organized in a setting different from the one in which the field work will be performed (e.g. dental institute instead of school setting), and 'preselected' subjects are used (e.g. cooperative patients with enough pathology present, interesting cases…). In addition, the examiner is aware of the purpose of the calibration exercise and will try to perform as good as possible. Therefore, the agreement measures obtained from these calibration sessions possibly present a distorted picture of the actual scoring behaviour of the examiner during the survey. Even when the examiners are instructed to re-examine some of the children during the field work (repeat examinations, allowing the calculation of intra-examiner agreement), possible bias can be introduced by the fact that the examiner is aware of the purpose of this exercise. As a consequence, the assessment of the reliability and validity of the CE scores can be questioned. When this information would be used to correct for misclassification in the main survey [5, 6], this may lead to biased estimates.

However, several other factors might also influence the quality of scoring of CE. It is plausible that the type of tooth (e.g. incisor versus molar), its position within the mouth (e.g. upper versus lower jaw) and the type of surface considered (e.g. distal versus occlusal surface) all impact on the accuracy of scoring. In addition, variations between examiners will exist.

The objective of this study, therefore, was to verify how valid misclassification measurements obtained from a 'pre-survey' calibration exercise are by comparing them to the scores obtained in 'field' conditions. The impact of the calibration setting on the quality of scoring of CE was evaluated considering other possible influencing conditions.

Materials and methods

The comparison of examiner performance was undertaken within the scope of an ongoing epidemiological survey.

Survey

The 'Smile for Life' (Tandje de Voorste) project is an oral health promotion intervention study in very young children (and their parents) launched in 2003 in Flanders (Belgium). Before starting the intervention, baseline data were collected in 3- and 5-year-old children. The results of this survey have been described in detail elsewhere [7]. Examiners participating in the oral health screenings were trained according to the guidelines published by the British Association for the Study of Community Dentistry (BASCD) [2, 8].

Recording of caries experience

Caries experience was recorded using the criteria proposed by the BASCD [2]. No radiographs were taken. The recording of CE on individual tooth surfaces took place at d_1-level (initial lesion), but allowing reporting of results at d_3-level ('cavitation' stage).

Calibration of examiners before start of the survey

The calibration exercise, involving eight dentist-examiners and the benchmark examiner (DD), was undertaken in a group of 32 (5-year-old) children. For practical and organizational reasons, they were selected from a school in the surroundings of the dental institute. A group of children showing good cooperation and presenting with a variety of pathologies, including untreated disease and recurrent caries and fillings, was selected by the benchmark scorer. Some caries-free children were also included. Informed consent was obtained from the parents of all the children.

The oral cavity of the children was examined in a classroom with the child seated on an ordinary chair, using a mouth mirror with a built-in light source (Mirrorlite™ by Defend® from Medident, Belgium) and WHO-CPITN type E probe (Prima Instruments, Gloucester, UK).

The examinations took place in sessions of each eight children (four sessions in total) and were organized in such a way that the children remained seated during the whole procedure, and the examiners circulated following a pre-determined scheme until every child was examined by every examiner. Each dentist-examiner was assisted by a recorder who was responsible for recording the clinical data into a prepared form.

Assessment of data reliability under field conditions

In order to assess the quality of the scoring of the different examiners under field conditions, the benchmark paid an unannounced visit to each of the examiners at the location they were working on that specific day. The benchmark scorer arrived just after (within a few hours) the 5-year-old children participating in the *Smile for Life* project had been examined by the dental examiner and re-examined the children. This re-examination by the benchmark scorer is different from what is undertaken in the case of intra-examiner monitoring of diagnostic consistency in which an examiner re-examines some of the survey children (usually 10%). Between six and 11 children per examiner were re-examined by the benchmark scorer, which amounts to 70 children in total.

Caries experience assessment was performed using identical methods and materials as described above. Data

were entered on site in an electronic database by a logistic assistant using Dental Survey Plus 2 (version 1.1) software (School of Computing, Dundee University).

Data management and analysis

Data from the calibration exercise (recorded on paper forms) were entered into an electronic database by a logistic assistant using Dental Survey Plus 2 (version 1.1) (School of Computing, Dundee University). All electronic data files were transferred to SAS files for further analysis.

Measurement of agreement (sensitivity and specificity)

The level of agreement between the scores obtained by the examiners and the benchmark was assessed by calculating sensitivity and specificity scores, at d_1-level and at two levels, i.e. at tooth and at surface levels. When estimating sensitivity and specificity, the obtained data were split into two groups according to the score attributed by the benchmark: (1) surface with CE and (2) surface without CE. Therefore, the proportion of surfaces in group 1 scored as presenting CE according to the examiner is the sensitivity, and the proportion of surfaces in group 2 without CE according to the examiner is the specificity. Since kappa scores are widely used in the dental literature, they were also calculated between the dental examiner and the benchmark scorer and are shown in the tables. All agreement measures were calculated both for the data obtained in the calibration exercise and under field conditions.

Agreement measures can be influenced by several factors, not only the condition under which the data were collected. Possible factors include: tooth type (molar tooth versus non-molar tooth), surface type (mesial, distal, occlusal, lingual or buccal surface), jaw (mandible or maxilla), quadrant (right or left side) and examiner. In order to assess the influence of the different factors on sensitivity and specificity, a logistic regression model was applied. All the above-mentioned factors were included in the model as categorical variables. For more details on the model, we refer to the Appendix.

The simple logistic regression model assumes that the data are independent. However, when data are clustered, as is the case in this example (surfaces nested within teeth and teeth within subjects), the logistic model needs to be extended [9]. If not accounted for clustering or correlation, the variance of the parameters—and hence, the *P* value associated with a parameter—will be wrongly estimated. A common approach to deal with correlated data is the generalized estimating equations (GEE) approach [10]. The GEE approach is based on a model where the parameters have a population average (classical logistic regression) interpretation while the variances of these parameters are corrected for the clustering. The level of significance was set at 0.05.

Results

In a first step, the quality of scoring of the examiners from the *Smile for Life* project was assessed by calculating sensitivity, specificity and kappa scores for each examiner versus the benchmark scorer. This was undertaken both for the data obtained in the calibration exercise and the data collected under field conditions (Table 1). Scores were calculated at surface level. Sensitivity scores ranged from 53% to 87% (with a mean of 75%) during calibration and from 42% to 71% (with a mean of 52%) under field conditions. For all examiners except one examiner (ex 8),

Table 1 Sensitivity, specificity and kappa scores obtained by the different examiners (n=8) involved in the *Smile for Life* project during the calibration exercise and under field conditions (data analysed at d_1-level and at surface level)

Examiner	Calibration exercise			Field conditions		
	Sensitivity (%)	Specificity (%)	Kappa value	Sensitivity (%)	Specificity (%)	Kappa value
Ex 1	70.00	99.67	0.75	50.00	99.72	0.40
Ex 2	68.33	99.71	0.75	–	100.00	–
Ex 3	85.00	99.00	0.73	41.67	99.64	0.49
Ex 4	72.73	99.49	0.73	42.86	99.47	0.51
Ex 5	86.67	98.94	0.73	60.00	99.02	0.42
Ex 6	76.67	99.12	0.70	53.85	97.86	0.44
Ex 7	83.33	99.23	0.76	41.94	98.46	0.47
Ex 8	53.33	99.60	0.61	70.59	99.16	0.64
Average	74.52	99.34	0.72	51.56	99.17	0.48

– no sensitivity and kappa values calculated since both examiner and benchmark did not record any CE

sensitivity scores were higher for data obtained during the calibration session. Specificity scores were high (97.8% and above) under both conditions and seem not to be influenced by the setting. Kappa scores ranged from 0.61 to 0.76 (with a mean of 0.72) during calibration and from 0.40 to 0.64 (with a mean of 0.48) under field conditions. Kappa values were lower under field conditions (except for examiner 8) and showed limited variability among examiners when measured during calibration.

In the GEE model, other possible influencing conditions were incorporated. Results for sensitivity and specificity are depicted (separately) in Tables 2 and 3, respectively. Positive estimates reflect a higher sensitivity/specificity of scoring CE for that category compared to the reference category; negative estimates reflect lower scores. Odds ratios provide information on the effect sizes.

The sensitivity of scoring CE was significantly influenced by tooth type (lower in molar teeth compared to non-molar teeth), tooth surface (lower for lingual surfaces compared to occlusal surfaces), jaw (lower for mandible versus maxilla), setting (lower under field conditions compared to calibration exercise) and examiner (higher for examiners 3, 5, 6 and 7 compared to examiner 8) (Table 2).

Specificity was influenced by tooth type (higher for non-molar teeth compared with molar teeth), surface (lower for occlusal surface compared to all other surfaces) and examiner (lower for examiners 3, 5, 6 and 7 compared to examiner 8)(Table 3).

The GEE approach can be used to estimate SE and SP of scoring CE for a specific combination of factors. As an example, Table 4 shows estimates for the SE of scoring CE on the buccal surface of teeth in the right quadrant, separately for mandible and maxilla and also for calibration exercise and field conditions. The sensitivity of scoring CE on a buccal surface on the right side of the mouth is lower in the mandible than in the maxilla, this when measured in a pre-survey calibration exercise as well as under field conditions. This trend is seen for all examiners but with considerable variation among them regarding the extent of the differences. The impact is more pronounced in molars.

Discussion

Caries experience surveys usually involve several examiners with different scoring behaviours [11, 12]. Accurate and reliable assessment of disease by these examiners is an important contributory aspect to the overall quality of an epidemiological survey. To this end, the validity of scoring by the examiners involved in the survey needs to be

Table 2 GEE parameter estimates (standard deviation) and odds ratio for the sensitivity of scoring CE at surface level

Factor	Parameter	Estimate (SD)	*p* Value	Odds ratio estimate
Tooth type	Intercept	−1.15 (0.41)	0.01	–
	Non-molar	1.45 (0.43)	<0.01	4.26
	Molar	–	–	–
Surface	Buccal	0.18 (0.29)	0.54	1.20
	Distal	0.17 (0.24)	0.47	1.19
	Mesial	0.11 (0.31)	0.71	1.12
	Lingual	−0.59 (0.30)	0.05	0.55
	Occlusal	–	–	–
Jaw	Maxilla	1.78 (0.31)	<0.01	5.93
	Mandible	–	–	–
Quadrant	Right	0.03 (0.25)	0.90	1.03
	Left	–	–	–
Setting	Field	−1.28 (0.29)	<0.01	0.28
	Calibration	–	–	–
Examiner	1	0.65 (0.46)	0.16	1.92
	2	0.27 (0.46)	0.56	1.31
	3	1.23 (0.47)	0.01	3.42
	4	0.57 (0.43)	0.17	1.77
	5	1.69 (0.51)	<0.01	5.42
	6	0.97 (0.44)	0.03	2.64
	7	0.90 (0.43)	0.04	2.46
	8	–	–	–

– indicates the reference category

Table 3 GEE parameter estimates (standard deviation) and odds ratio for the specificity of scoring CE at surface level

Factor	Parameter	Estimate (SD)	*p* Value	Odds ratio estimate
Tooth type	Intercept	4.28 (0.30)	<0.01	–
	Non-molar	1.73 (0.24)	<0.01	5.64
	Molar	–	–	–
Surface	Buccal	0.44 (0.16)	0.01	1.55
	Distal	0.89 (0.20)	<0.01	2.44
	Mesial	1.43 (0.26)	<0.01	4.18
	Lingual	0.66 (0.19)	<0.01	1.93
	Occlusal	–	–	–
Jaw	Maxilla	−0.01 (0.18)	0.95	0.99
	Mandible	–	–	–
Quadrant	Right	−0.25 (0.18)	0.16	0.78
	Left	–	–	–
Setting	Field	−0.13 (0.23)	0.57	0.88
	Calibration	–	–	–
Examiner	1	0.38 (0.45)	0.40	1.46
	2	0.73 (0.46)	0.12	2.08
	3	−0.63 (0.33)	0.05	0.53
	4	−0.18 (0.36)	0.62	0.84
	5	−0.84 (0.32)	0.01	0.43
	6	−0.86 (0.34)	0.01	0.42
	7	−0.74 (0.34)	0.03	0.48
	8	–	–	–

– indicates the reference category

considered. In order to guarantee this aspect, examiners receive training before (and sometimes repeated during) the survey. The outcome of the training is assessed in calibration exercises. Data obtained from these calibration exercises are used to assess the agreement between examiners and between examiners and the benchmark examiner. These agreement measures are reported in publications, and they inform the reader about the reliability of the data obtained. In addition, they can be used to correct for misclassification in the main data.

Table 4 Sensitivity estimated by the GEE approach for a specific combination of categories of factors (buccal, right)

Examiner	Tooth type	Pre-survey calibration		Field conditions	
		Mandible (%)	Maxilla (%)	Mandible (%)	Maxilla (%)
Ex 1	Non-molar	76.13	94.98	47.00	84.02
	Molar	42.80	81.61	17.22	55.23
Ex 2	Non-molar	68.57	92.82	37.75	78.24
	Molar	33.85	75.21	12.46	45.76
Ex 3	Non-molar	85.07	97.13	61.30	90.38
	Molar	57.20	88.80	27.09	68.78
Ex 4	Non-molar	74.65	94.58	45.02	82.92
	Molar	40.85	80.38	16.11	53.25
Ex 5	Non-molar	90.02	98.17	71.50	93.70
	Molar	67.92	92.62	37.05	77.73
Ex 6	Non-molar	81.46	96.30	54.98	87.87
	Molar	50.75	85.94	22.27	62.95
Ex 7	Non-molar	80.38	96.05	53.25	87.10
	Molar	49.00	85.07	21.08	61.30
Ex 8	Non-molar	62.48	90.80	31.65	73.30
	Molar	28.09	69.85	09.80	39.17

However, calibration exercises are often organized in circumstances different from the conditions during field work, and this could impact the quality of the validation data obtained [13, 14]. Therefore, in this study, a comparison between the performances of examiners under both conditions was made (Table 1). To our knowledge, this is the first study that compares examiner performance under both conditions.

For assessing agreement among raters, the kappa statistic is often used for binary or ordinal (in general, categorical) measurements. However, when a benchmark examiner is available, the use of sensitivity and specificity is recommended [8]. The kappa statistic is a function not only of sensitivity and specificity, but also of prevalence and therefore is more difficult to interpret [15]. For example, different kappa values can be obtained in different studies (differing in disease prevalence) even when the examiners score the same way [16]. Also, the same kappa value can correspond to different kinds of agreement [16]. Finally, here, we model sensitivity and specificity as a function of covariates while taking the multilevel structure of the data into account. A similar modelling exercise with kappa software is lacking for the situation considering more than two levels [17].

As shown in Table 1, sensitivity seems to be influenced by the validation setting with higher scores obtained during the calibration session than under field conditions (difference of about 20% to 40%). This difference in sensitivity points towards an underestimation of CE in field circumstances. The differences in scoring behaviour could be due to the fact that, for the calibration session, 'preselected' subjects were used [13]. These subjects were likely to have a disease level, cooperation pattern and variation of pathologies different from what was observed in the field. In addition, work load and time constraints during the survey are likely to contribute to the differences in sensitivity and specificity observed in the two settings. Also important is the fact that the examiners were aware of the purpose of the calibration exercise, and they will have tried to perform as good as possible. When active in field conditions, without expecting any 'checking' of their performance, the accuracy of their scoring might be lower. On the other hand, learning experience during the survey might improve their performance (see, e.g. examiner 8).

In Table 1, it is also observed that some examiners had low sensitivity values (e.g. examiner 8). To allow this examiner to participate in the actual survey can have an impact on the quality of the obtained results. One way to solve this problem is to omit the examiner from the survey. However, as can be derived from Table 1, this same examiner performed very well under field conditions (highest sensitivity score). A possible way of handling this situation is to correct for misclassification for all examiners involved. This will take into account that some are good scorers, and others are bad scorers [6].

For reasons of comparison, kappa values were also presented here (Table 1). Values obtained during the calibration session ranged between 0.61 and 0.76, showing little variation among examiners, and dropped to much lower values under field conditions. It is clear that sensitivity and specificity scores provide more information on both the nature and the extent of the disagreement. They indicate over- and/or underestimation of the scoring process (which cannot be derived from the kappa score) and show more variability among the examiners.

To validate the findings presented in Table 1 statistically, the validation setting—among other factors—was included in a logistic regression model using a GEE approach in order to account for the clustering of data. It was observed that there was a significant effect of validation setting on the sensitivity of scoring CE, with sensitivity being higher when measured in the pre-survey calibration exercise than in field conditions. This is reflected by the statistically significant negative GEE estimate (−1.28) obtained for field compared to pre-survey conditions. From the analysis, it became clear that sensitivity is also influenced by tooth type, as indicated by the statistically significant positive estimate (1.45) for non-molar teeth compared to molar teeth. Higher sensitivity scores were obtained for detecting CE on incisors and canines compared to molars. This is possibly due to the more complex anatomical structure of the molars and also to the fact that these teeth are more posteriorly positioned in the mouth which could impair their visualisation by the examiners. Apart from tooth type, the jaw examined also influenced the sensitivity of CE scoring with higher sensitivity observed for surfaces in the maxilla than in the mandible. It is unclear how this finding can be explained. The position of the tongue and presence of saliva might hamper CE detection more in the mandible than in the maxilla.

Furthermore, there was also an 'examiner' effect on sensitivity of scoring CE, with all the examiners performing better than examiner 8 as shown by the positive estimates. The observed differences in the scoring behaviour of examiners could relate, among other things, to their experience in CE scoring in an epidemiological setting and number of years in practice [18, 19]. There was no influence observed of surface type and quadrant on sensitivity of scoring.

The scoring of incisors and canines yielded higher specificity scores than in molars. Surfaces of incisors and canines are located more anteriorly in the oral cavity compared to molars; they have less complex anatomical structures, and direct visualisation is usually easier. This can explain the differences in specificity scores obtained. Furthermore, it is possible that discolorations and food

accumulation, frequently occurring on the occlusal surfaces of molars, are responsible for the lower specificity observed when scoring these surfaces compared to other surfaces [20, 21].

The logistic regression model proposed in this paper yields parameter estimates that allow the calculation of estimates for a combination of categories. In this way, differences between conditions and/or examiners can be considered, allowing more informative feedback. It is not only important to have an idea of overall examiner performance; the possibility of estimating sensitivity and specificity scores for specific situations (e.g. occlusal surfaces versus other surfaces, etc.) yields information that can be used to provide more specific and detailed feedback to the examiners involved in the survey.

Conclusions

Knowledge about factors influencing the reliability of CE scoring is important not only when setting up agreement measurement and constructing a validation data set, but also when interpreting the outcome of this assessment. In order to guarantee the quality of the validation of examiner performance, the characteristics of selected individuals should be as close as possible to those of the target population of the planned survey in age, disease distribution, etc. Also, examiner calibration should be organised in a setting closely resembling the actual circumstances in which the survey will take place. However, this is often difficult to realise because of practical and organisational issues. It is clear that the assessment of data reliability is an important issue. Results obtained in a pre-survey calibration setting need to be interpreted with caution. In the present paper, this was shown for CE scoring, but the same conclusion is likely to be valid also for the scoring of other (oral) conditions. More work is needed to optimize the whole process: the collection of validation data, interpretation of results and feedback to examiners.

Acknowledgements This investigation was supported by Research Grant OT/05/60, Catholic University Leuven. The following partners collaborated in the "Smile for Life Project": Dominique Declerck (Project Coordinator) and Roos Leroy (both from the Department of Dentistry, Catholic University Leuven), Karel Hoppenbrouwers (Youth Health Care at the Catholic University Leuven, and the Flemish Society for Youth Health Care), Emmanuel Lesaffre (Centre for Biostatistics, Catholic University Leuven), Stephan Vanden Broucke (Research Group for Stress, Health and Well-being at the Catholic University Leuven), Luc Martens (Dental School, Ghent University), Erwin Van Kerschaver and Martine Debyser (Child and Family). The study was supported financially by GABA Benelux and GABA International.

Conflict of interests The authors declare that they have no conflict of interests.

Appendix

Logistic regression model for sensitivity and specificity

Suppose that the binary score for CE is denoted as Y for the benchmark and Y^* is the score attributed by the examiner. Thus, $Y=1$ corresponds to CE truly present, and $Y=0$ means no CE present. Using this notation, sensitivity is equal to $\pi_{se} = \Pr(Y^* = 1|Y = 1)$ and specificity is equal to $\pi_{sp} = \Pr(Y^* = 0|Y = 0)$.

A logistic regression model relating π_{se} and π_{sp} to p factors x_1, x_2,…x_p is given by:

$$\text{logit}(\pi_{se}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \quad (1)$$

$$\text{logit}(\pi_{sp}) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + ... + \gamma_p x_p, \quad (2)$$

respectively, where logit $(\pi) = \log(\pi/[1-\pi])$. The coefficients β_0, β_1,…, β_p and γ_0, γ_1,…, γ_p are called regression coefficients and are estimated according to the method of maximum likelihood. The coefficients β_0 and γ_0 are called intercepts. The other coefficients measure the strength of the relationship between the regressors and sensitivity and specificity, respectively.

Models 1 and 2 assume that the data are independent. However, here, the data are clustered, since surfaces are nested within teeth, and teeth are nested within mouths. Therefore, the models were extended further to account for this clustering using the GEE approach. This approach is based on supposing at the start a correlation structure for the outcomes, called working correlation matrix. The term 'working' refers to the fact that it is good enough that the correlation matrix roughly represents the true correlation structure. Here, an exchangeable working correlation was assumed, which means that the correlation of CE among surfaces on the same tooth and on teeth in the same mouth are all equal.

References

1. International Caries Detection and Assessment System Coordinating Committee (2009) Criteria manual—international caries detection and assessment system (ICDAS II)
2. Pitts NB, Evans DJ, Pine CM (1997) British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. Community Dent Health 14 (Suppl 1):6–9
3. World Health Organization (1997) Oral health surveys. Basic methods. World Health Organization, Geneva
4. Assaf AV, Tagliaferro EP, Meneghim MC, Tengan C, Pereira AC, Ambrosano GM, Mialhe FL (2007) A new approach for interexaminer reliability data analysis on dental caries calibration. J Appl Oral Sci 15:480–485

5. Mwalili SM, Lesaffre E, Declerck D (2008) The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Stat Meth Med Res 17:123–139
6. Mutsvari T, Lesaffre E, Garcia-Zattera MJ, Diya L, Declerck D (2010) Factors that influence data quality in caries experience detection: a multilevel modeling approach. Caries Res 44:438–444
7. Declerck D, Leroy R, Martens L, Lesaffre E, Garcia-Zattera MJ, Vanden BS, Debyser M, Hoppenbrouwers K (2008) Factors associated with prevalence and severity of caries experience in preschool children. Community Dent Oral Epidemiol 36:168–178
8. Pine CM, Pitts NB, Nugent ZJ (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 2. Community Dent Health 14(Suppl 1):18–29
9. William FM, Nan G (2006) Estimation of sensitivity and specificity of clustered binary data. Statistics and data analysis, SUGI 31 proceedings, SAS Proceedings
10. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
11. Gorelick MH, Atabaki SM, Hoyle J, Dayan PS, Holmes JF, Holubkov R, Monroe D, Callahan JM, Kuppermann N (2008) Interobserver agreement in assessment of clinical variables in children with blunt head trauma. Acad Emerg Med 15:812–818
12. Helderman WH, Mulder J, van'T Hof MA, Truin GJ (2001) Validation of a Swiss method of caries prediction in Dutch children. Community Dent Oral Epidemiol 29:341–345
13. Castiglia P, Campus G, Solinas G, Maida C, Strohmenger L (2007) Children's oral health in Italy: training and clinical calibration of examiners for the National Pathfinder about caries disease. Oral Health Prev Dent 5:255–261
14. Cleaton-Jones P, Hargreaves JA, Fatti LP, Chandler HD, Grossman ES (1989) Dental caries diagnosis calibration for clinical field surveys. Caries Res 23:195–199
15. Agbaje JO, Mutsvari T, Lesaffre E, Declerck D (2010) Measurement, analysis and interpretation of examiner reliability in caries experience surveys: some methodological thoughts. Clin Oral Investig (in press)
16. Lesaffre E, Mwalili SM, Declerck D (2004) Analysis of caries experience taking inter-observer bias and variability into account. J Dent Res 83:951–955
17. Williamson JM, Lipsitz SR, Manatunga AK (2000) Modeling kappa for measuring dependent categorical agreement data. Biostatistics 1:191–202
18. Heifetz SB, Brunelle JA, Horowitz HS, Leske GS (1985) Examiner consistency and group balance at baseline of a caries clinical trial. Community Dent Oral Epidemiol 13:82–85
19. Poorterman JH, Verheij JG, Kieft JA, Eijkman MA (1997) Variations among dentists in the diagnosis of caries and assessment of dental restorations. Ned Tijdschr Tandheelkd 104:214–218
20. Cortes DF, Ellwood RP, Ekstrand KR (2003) An in vitro comparison of a combined FOTI/visual examination of occlusal caries with other caries diagnostic methods and the effect of stain on their diagnostic performance. Caries Res 37:8–16
21. Mojon P, Favre P, Chung JP, Budtz-Jorgensen E (1995) Examiner agreement on caries detection and plaque accumulation during dental surveys of elders. Gerodontology 12:49–55