Children's Oral Health in Italy: Training and Clinical Calibration of Examiners for the National Pathfinder about Caries Disease

Paolo Castiglia[a], Guglielmo Campus[b], Giuliana Solinas[a], Carmelo Maida[b], Laura Strohmenger[c]

Purpose: To evaluate the dental status of 12-year-old children in 2004, an epidemiological survey called 'National pathfinder about children's oral health in Italy' was planned. The present paper describes the clinical calibration method.

Materials and Methods: Thirty children (20 caries-affected and 10 caries-free) were selected. Seven examiners received 35 hours of training and calibration in making clinical measurements by a single trainer. The trainer's diagnosis was the benchmark validity reference. Increase in sensitivity was tested using paired Student's *t*-test. Inter-examiner reliability for DMFS was evaluated using Pearson's correlation coefficients and ANOVA. Percent agreement and Kappa statistics were calculated for dichotomous judgments (tooth-by-tooth analysis and caries prevalence).

Results: The raters' sensitivity increased from calibration I to calibration II ($p < 0.05$). For DMFS a good agreement between raters was found without significant differences among them at both calibration exercises. The Pearson's correlation coefficients between all pairs of raters were significant ($p < 0.001$) after both calibration exercises. Tooth-by-tooth reliability was high: the strength of agreement was substantial or almost perfect in calibration I (Kappa statistics from 0.81 to 0.95) and almost perfect in calibration II (Kappa statistics from 0.92 to 0.97). Regarding caries prevalence, sensitivity, specificity, percent agreement and Kappa statistics results were high but without significant differences between individual pairs of observations in the two calibration sessions ($p > 0.05$).

Conclusions: The results suggest that sensitivity measurements should be used for agreement assessments in addition to the other World Health Organization-recommended reliability measurements. Kappa statistics are the eligible measurement of reliability for dichotomous judgments.

Key words: caries, dental health, epidemiology, examiner calibration

Oral Health Prev Dent 2007;3:xxx-xxx. *Submitted for publication: 04.05.06; accepted for publication: 22.09.06*

For a long time, many health services throughout the world have focused on the prevention of oral and dental diseases. To gain significant results in this field, it is crucial to have access to trustworthy data regarding oral diseases of the population (Fyffe et al, 2000).

Meticulous and reliable data collection is vital for success in all fields of research. These issues are seldom considered in dental research (Lesaffre et al, 2004). It is particularly true when multiple examiners are involved in an important epidemiological survey, since their observations must be comparable. In this context, training is necessary, and it can be defined, according to the Guidance on the Statistical Aspects of Training and Calibration of Examiners for Surveys of Child Dental Health by British Association for the Study of Community Dentistry (BASCD) (Pine et al, 1997), as teaching the agreed interpretation of the diagnostic criteria. It is essential that the full range of diagnostic situations is presented and discussed using both slides and clinical examples. Moreover, sufficient opportuni-

[a] *Institute of Hygiene and Preventive Medicine, Laboratory of Epidemiology and Biostatistics, University of Sassari, Italy.*

[b] *Dental Institute, University of Sassari, Italy.*

[c] *School of Dentistry, University of Milan, 'S.Paolo Hospital' WHO Collaborating Center of Milan for Epidemiology and Community Dentistry, Italy.*

Reprint requests: *Professor Guglielmo Campus, Dental Institute, University of Sassari, Viale San Pietro 43/c, I 07100 Sassari, Italy. Tel: +39 079 228540. Fax: +39 079 228541. Email: gcampus@uniss.it*

ties should be made for discussion of the interpretation of the criteria; these discussions should be performed in a timely manner and in a sympathetic atmosphere. Problems with the interpretation of specific conditions or misapprehensions about the use of particular codes can only be identified and solved at this time. When problems are encountered in the field, the examiners should have the confidence to give an immediate response. Theoretical discussions and clinical training are complemented by calibration, which provides a formal measure of how well the examiner can interpret the criteria in comparison to the gold standard set by the trainer.

The higher the number of cases in calibration exercises, the more successful training is; enough cases should be involved to ensure adequate material for the range of diagnostic decisions that have to be made (Pine et al, 1997). Since calibration and training become extremely important if more than one examiner is involved in the data collection, it is necessary to ensure not only that each examiner individually is consistent and reliable (intra-examiner reliability), but also that each one agrees with the others (inter-examiner reliability). Reliability is investigated by examining several patients at least twice, to discover if the diagnostic results for a rater remain reliable in both measurements. When measurements are recorded by one examiner, intra-examiner reliability can be tested; when more examiners rate the same patients, inter-examiner reliability is examined. However, testing intra-examiner reliability might be affected by the presence of some systematic biases, due to the difficulty of blinding the rater to ensure the independence of the two measurements. Thus, inter-examiner evaluation more accurately estimates reliability (McPartland and Goodridge, 1997). In this kind of study, reliability is traditionally estimated through the percent agreement, i.e. the proportion of agreement between examiners judging dichotomous variables (e.g. the presence or absence of caries disease). Nevertheless, percent agreement overestimates reliability because it contains both chance agreement, as if the rater had just guessed each rating, and beyond chance agreement. Thus, the most suitable statistic for evaluating inter-examiner reliability is Cohen's kappa coefficient, which estimates the degree of agreement corrected for chance agreement. The range of possible values of kappa is from -1 to 1, even if it usually falls between 0 (chance agreement) to 1 (perfect agreement). A negative kappa would indicate agreement worse than that expected by chance. However, the kappa statistic also has disadvantages, such as its misapplication to noncategorical data. In this case Pearson's correlation must be used. However, if a systematic bias is present (e.g. if a rater indicates a lesion on one or more teeth per patient that the other rater doesn't report), Pearson's correlation coefficient overestimates the agreement (Hunt, 1986).

For all these reasons, a reproducibility study must precede a national multi-center study in order to evaluate inter-examiner reliability.

In Italy, there are no national data on oral diseases; in particular, there are few papers on caries disease in the literature (Vogel et al, 1986; Strohmenger et al, 1997; Angelillo et al, 1998; Campus et al, 2001; Arcella et al, 2002; Campus et al, 2005). Furthermore, diagnostic criteria are different among studies and only in a few papers a calibration of the examiners was carried out (Bolin et al, 1997; Pine et al, 2004).

In 2004, an epidemiological survey called 'National pathfinder about children's oral health in Italy' was promoted by the Italian Society of Hygiene and Preventive Medicine. The aim was to evaluate dental status, related to attitudes towards dental care in 12-year-old children. The study was coordinated by the World Health Organization (WHO) collaborating center for oral health of Milan at the University Dental Institute. In the present paper, the clinical calibration method for the National pathfinder is described.

MATERIALS AND METHODS

Population and sampling

The calibration sessions were held in Milan at the University Dental Institute. For the National pathfinder, a multi-stage cluster sampling technique was chosen, considering in the first stage the Italian sections according to the National Institute of Statistics classification (North-Western, North-Eastern, Central, Southern and Insular Italy [ISTAT, 2004], in the second stage, the cities of the sections, and in the third stage the secondary schools of the cities and then each class. The final sample contained approximately 6000 children attending the first year of secondary school. Due to the high number of children and the ease of transporting the raters throughout the country in a seven-seat car, the number of examiners was set at seven. Each selected school was visited by all seven raters at the same time. This method had two advantages: 1) due to the shortness of a school visit, it reduced the school organisational disadvantages; 2) it managed to avoid the inter-cluster variability attributable to the inter-examiner variability. Nevertheless, high reproducibility between raters was necessary, thus inter-examiner relia-

Table 1 Variables registered in the Calibration exercise for the Pathfinder study (Bolin et al, 1995)

Surface level	Comment
1. Missing	4 surfaces for anterior teeth, 5 for posterior. Missing teeth in primary dentition wererecorded as extracted due to caries for molars and canine, patients were questionedregarding trauma to anterior teeth.In the permanent dentition the examiner asked the child why the tooth or the teethwere missing.
2. Filled	The surface and the kind of restorative material (amalgam, composite, glass ionomercement, crown, temporary cement) were registered. Fissure sealants were recordedseparately and not included in the F component of the DMFS index. To differentiatebetween coloured fissure sealant and composite filling, it was decided that an opaquearea closely fitting the anatomical fissure was recorded as a sealant.
3. Carious	Caries were recorded on each tooth surface only if the CPITN probe could enter thelesion, or if the lesion could be judged as clearly in the dentine, undermining theenamel, with the underlying carious dentine appearing as a brownish or whitish spot.Arrested caries of dentine with a hard, smooth, 'non-sticky surface', were alsoregistered as caries, and accordingly were included in the DMFS index.
4. DMFS	This was a calculated variable (1+2+3).

bility had to be maximised over the threshold indicated by WHO, i.e. 85% in all measurements (WHO, 1997). The calibration study was focused on 12-year-old children, following the WHO indications (WHO, 1997).

Thirty subjects were selected from schoolchildren attending a secondary school in Milan: 10 children were caries-free and 20 had caries experience at various levels (DMFS range 1–16). The benchmark rater did not see the children before the other raters, but visited them on the same occasion and under the same conditions.

Clinical examinations

The team of seven examiners received training and calibration in making clinical measurements. The calibration exercises were conducted independently from the epidemiological pathfinder study to avoid the demands of participants. Agreement was reached on the design of the main study as well as the diagnostic criteria. Every rater evaluated all 30 subjects by examining all tooth surfaces and comparing their results with the examination results of the benchmark examiner. The children were examined twice by each rater and once by the benchmark examiner; consequently, each subject was examined fifteen times. The method adopted to diagnose dental caries was the visual/tactile method and the equipment consisted of sterile mirrors and WHO CPI dental probe, without diagnostic adjuncts such as previous dental brushing and drying. This is the method recommended by the WHO (WHO, 1997) and is the best for low or moderate prevalence groups for dental caries (Assaf et al, 2004). The programme for calibration was 5 days, spanning 35 hours. Table 1 shows the diagnostic criteria.

As no true gold standards are generally available for caries diagnosis in caries prevalence surveys, a benchmark examiner may be used as a gold standard against which to assess sensitivity; in the present study, GC trainer's diagnosis was the benchmark validity reference. The benchmark examiner (GC) is a dentist who habitually uses the WHO criteria for examinations and who had previously been trained and calibrated in the diagnosis of carious lesions to ensure comparability of the clinical registration for epidemiological purposes in order to compare dental health in different countries (Bolin et al, 1995). Inter-examiner reproducibility was achieved in two different examinations of 30 individuals. No discussion was permitted between the raters and the dental advisors as to the interpretation of the criteria during the calibration sessions.

Statistical analysis

Statistical analysis was performed with Stata 9 (StataCorp, TX, USA). For DMFS, sensitivity was set as the proportion of affected (decayed plus filled) surfaces indicated by the rater on those diagnosed by the benchmark; the difference in sensitivity for each rater between the two calibration exercises was assessed using paired Student's *t*-test at a significance level of $\alpha = 0.05$. Inter-examiner reliability was evaluated through analysis of variance for fixed effect (Fleiss, 1986) and Pearson's correlation coefficients. In order

Table 2 Sensitivity (%) for DMFS measurement from calibration I to II

Rater	Calibration I	Calibration II	Difference (II–I)
1	94.8	96.5	1.7
2	80.9	90.4	9.6
3	90.4	96.5	6.1
4	91.3	95.7	4.3
5	81.7	94.8	13.0
6	80.0	93.0	13.0
7	85.2	91.3	6.1
p-value*			0.003

* Paired Student's t-test

Table 3 Inter-examiner reliability as assessed by Pearson's correlation coefficients: calibration I below diagonal, calibration II above diagonal. In diagonal, correlation coefficients between the two calibration exercises for each

	Benchmark	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
Benchmark	-	0.995	0.993	0.994	0.994	0.994	0.991	0.986
Rater 1	0.993	0.997	0.996	0.995	0.995	0.994	0.991	0.983
Rater 2	0.979	0.977	0.985	0.993	0.994	0.991	0.992	0.988
Rater 3	0.977	0.974	0.970	0.990	0.988	0.990	0.992	0.991
Rater 4	0.974	0.980	0.962	0.938	0.988	0.989	0.984	0.976
Rater 5	0.956	0.955	0.942	0.935	0.937	0.972	0.992	0.982
Rater 6	0.915	0.904	0.925	0.932	0.857	0.894	0.944	0.994
Rater 7	0.972	0.971	0.967	0.981	0.933	0.938	0.945	0.991

All correlation coefficients are significant at the 0.001 level (2-tailed).

to study inter-examiner agreement, DF values of each tooth were codified with 0 as sound and 1 as affected (decayed or filled); tooth-by-tooth percent agreement and Cohen's kappa statistics were then calculated. For caries experience, each subject was codified with 0 as sound and 1 as affected (at least one decayed or filled tooth surface). The strength of agreement associated with kappa statistics was labelled as indicated by Landis and Koch (1977) and successively modified by Jamieson and co-workers (Jamieson et al, 2004): < 0.51 slight, 0.51–0.60 fair, 0.61–0.70 acceptable, 0.71–0.80 moderate, 0.81–0.90 substantial, > 0.90 almost perfect. Kappa statistics were tested using z test at a significance level of $\alpha = 0.001$.

RESULTS

Regarding DMFS values, the agreement between each examiner against the gold standard is expressed as sensitivity in Table 2. In the first calibration exercise, the raters' sensitivity ranged from 80% to almost 95%, with a mean sensitivity value of 86.3%. As part of this project and training course, the results were discussed, and when necessary, the criteria were modified. As a result, in the second calibration exercise, the mean sensitivity increased to 94%, while the range decreased (91.3–96.5%); the average difference between individual pairs of observations in the two calibration sessions was significantly different from 0 (Student's paired *t*-test: p = 0.003). Table 3 shows the Pearson's correlation coefficients between the raters (the inter-examiner reliability) both at the first calibration exercise (values reported below diagonal in the table) and at the second clinical examination (above diagonal); comparison with benchmark is also reported. Significant correlation was observed between all pairs of raters (p < 0.001). In calibration I, rater 6 vs. raters 4 and 5 showed the lowest values (0.857 and 0.894 respectively). After calibration II, all values were more than 0.975. Values in the diagonal (bold text) refer to correlation between the two calibration exercises of each examiner; raters 6 and 5 showed the lowest values (0.944 and 0.972 respectively). The inter-examin-

Table 4 Inter-examiner reliability: analysis of variance on seven raters DMFS scores

Calibration I Source of variation	Degrees of freedom	Sum of squares	Mean square	F ratio	p value
Subjects	29	3649.35	125.84	102.78	
Raters	6	11.52	1.92	1.57	0.16
Error	174	213.05	1.22		
Total	209	3873.92			

Calibration II Source of variation	Degrees of freedom	Sum of squares	Mean square	F ratio	p value
Subjects	29	4004.84	138.10	613.58	
Raters	6	1.98	0.33	1.47	0.19
Error	174	39.16	0.23		
Total	209	4045.98			

Table 5 Tooth-by-tooth Kappa statistics and percent agreement between all pairs of examinations

Calibration I	Benchmark	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
Benchmark	-	98.7	97.0	98.1	97.9	96.6	96.7	97.5
Rater 1		0.95*	-	96.7	97.9	97.9	96.9	97.0
97.5								
Rater 2	0.87°	0.86°	-	97.1	96.6	96.1	96.5	97.5
Rater 3	0.92*	0.91*	0.87°	-	96.2	96.0	96.6	98.1
Rater 4	0.91*	0.91*	0.85°	0.84°	-	95.7	95.6	96.1
Rater 5	0.86°	0.86°	0.82°	0.82°	0.82°	-	95.9	95.9
Rater 6	0.86°	0.87°	0.83°	0.85°	0.81°	0.81°	-	97.7
Rater 7	0.89°	0.89°	0.88°	0.92*	0.83°	0.81°	0.90°	-

Calibration II	Benchmark	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
Benchmark	-	99.2	98.5	99.1	99.1	99.0	98.7	98.4
Rater 1	0.97*	-	98.7	99.1	99.1	99.0	99.0	98.4
Rater 2	0.94*	0.95*	-	98.6	98.9	98.7	98.7	98.9
Rater 3	0.96*	0.96*	0.94*	-	98.5	98.6	98.6	98.7
Rater 4	0.96*	0.96*	0.95*	0.94*	-	98.4	98.4	98.0
Rater 5	0.96*	0.96*	0.95*	0.94*	0.93*	-	98.7	98.1
Rater 6	0.95*	0.96*	0.95*	0.94*	0.93*	0.95*	-	99.1
Rater 7	0.93*	0.93*	0.95*	0.95*	0.92*	0.92*	0.96*	-

Below diagonal, kappa statistics; above diagonal, percent agreement.
Strength of agreement for Kappa: °substantial, *almost perfect (Landis and Kock, 1977), as modified by Jamieson and co-workers (Jamieson et al, 2004); z tests for all Kappa statistics were

er reliability analysed via analysis of variance is reported in Table 4. In calibration I, a good agreement between raters was found without significant differences among them (p = 0.16), although a high value of mean squares for error (EMS = 1.22) was present. After calibration II, the error value drastically decreased (EMS = 0.23).

Tooth-by-tooth Kappa statistics and percent agreement between all pairs of clinical examinations are shown in Table 5 for the two calibration exercises separately; a comparison with benchmark examination is also included (below the diagonal are the Kappa statistics; above the diagonal, percent agreement). Percent agreement ranged from 95.6% (rater 6 vs. rater

Table 6 Reliability measures for prevalence of caries

Calibration I				Calibration II				
Rater	Sensitivity	Specificity	Percent	Kappa	Sensitivity	Specificity	Percent	Kappa
	(%)	(%)	agreement	(%)	(%)	agreement		
1	90.0	100.0	93.3	0.86°	90.0	100.0	93.3	0.86°
2	95.0	90.0	93.3	0.85°	90.0	100.0	93.3	0.86°
3	90.0	100.0	93.3	0.86°	90.0	100.0	93.3	0.86°
4	95.0	90.0	93.3	0.85°	95.0	100.0	96.7	0.93*
5	80.0	100.0	86.7	0.73†	95.0	100.0	96.7	0.93*
6	85.0	100.0	90.0	0.79†	90.0	100.0	93.3	0.86°
7	90.0	100.0	93.3	0.86°	90.0	100.0	93.3	0.86°

Strength of agreement for Kappa (Jamieson et al, 2004): †moderate, °substantial, *almost perfect; z tests for all Kappa statistics were significant ($p < 0.001$).

4) to 98.7% (rater 1 vs. benchmark) in the first calibration session and from 98.0% (rater 7 vs. rater 4) to 99.2% (rater 1 vs. benchmark) in the second session. The strength of agreement, recorded as Jamieson's categories of Kappa statistics, within raters and between raters and benchmark was 'substantial' or 'almost perfect' in calibration I (Kappa statistics ranged from 0.81 to 0.95). In calibration II, Kappa statistics ranged from 0.92 to 0.97 with the strength of agreement 'almost perfect' between all pairs of measurements. In both calibrations, all Kappa statistics achieved significant values of z ($p < 0.001$).

Table 6 shows raters' agreement measures (sensitivity, specificity, percent agreement and kappa statistics) based on prevalence of caries (decayed or filled teeth: yes/no) in comparison with the benchmark for the two calibration exercises separately. Inter-raters' agreement was high at both calibration exercises without significant differences between individual pairs of observations in the two calibration sessions ($p > 0.05$). Nevertheless, at calibration II the measures became more similar and sensitivity range decreased from 80.0–95.0% to 90.0–95.0% between the two calibration exercises, specificity increased from 90.0–100.0% to 100.0%, and significant kappa values ($p < 0.001$) increased from 0.73–0.86 to 0.86–0.93. Sensitivity increased only for raters 5 and 6, but decreased for rater 2; specificity stretched to the higher value for all raters and kappa statistics reached substantial or almost perfect values.

DISCUSSION

Before an epidemiological survey is carried out, a reliability study is essential in order to assess the level of observer variability in the measurement procedures to be applied in data acquisition. These kinds of study consent to set validity of measurements concerning either the relationship between the item findings with the reference values and the inference drawn from the survey. In Italy, no national surveys have been reported on caries disease. The present study evaluated reliability after training using two calibration exercises among seven raters, both for DMFS measurement and for proportion of caries-free 12-year old children.

The results clearly show that for reliable comparison in a multi-centric epidemiological study, calibration of the raters is crucial. Although at the time of the first calibration exercise, a high level of inter-examiner reliability had already been achieved, both using analysis of variance for fixed effects and Pearson's correlation coefficients, a significant improvement of sensitivity for quantitative measurements from calibration I to calibration II was reached. The minimum acceptable level of agreement recommended by WHO is set at 85.0% (WHO, 1997); in the first calibration exercise, the sensitivity of two raters was under the threshold value, while the reliability measurements for DMFS were satisfactory. Sensitivity seemed to be a useful measurement of agreement in identifying less reliable raters both for quantitative values (DMFS) and qualitative ones (DF prevalence). No systematic bias between raters' scores was noted in our calibration exercises or for the benchmark, therefore Pearson's correlation coefficients did not overestimate inter-examiner reliability. However, this measurement was less helpful than sensitivity in identifying those raters in need of more clinical training, after the first calibration exercise. Like sensitivity, ANOVA underlined the inter-examiner reliability improvement between the two calibration exercises, even if analysis of variance cannot identify raters that highly contribute to the magnitude of variability. To reach adequate accuracy, we propose that sensitivity should be used together with the other recommended reliability measurements.

Regarding qualitative measurements, tooth-by-tooth agreement analysis showed that Kappa statistic more accurately represented the reliability of the diagnosis of caries on individual teeth, while percent agreement overestimated the true reliability of the raters. This phenomenon might be related to the high prevalence of caries-free teeth in the sample, and consequently, an agreement on the absence of disease was relatively easy. This was supported by the reliability analysis for the prevalence of caries, as specificity reached the 'ideal' value of 100% after the second calibration exercise.

In addition, the magnitude of Kappa increased after the second calibration exercise and the strength of agreement was 'almost perfect' between all pairs of measurements.

The present study has determined high levels of inter-examiner reliability for a national pathfinder study on caries disease. The results suggest that sensitivity measurements should be used for quantitative variables together with Pearson's correlation coefficients. For dichotomous judgments, such as the presence or absence of disease (tooth-by-tooth analysis or caries prevalence), we confirm that Kappa statistics are suitable measurements of reliability. The methods used for the calibration of our examiners have proved to be efficient in several ways. The training of the examiners allowed the increase of reliability at the second calibration. The main reason for the success is that all the examiners were committed to working together, and able to find compromises about examination criteria, so that all participants accepted the design of study. In order to plan an epidemiological survey that involves national centres, it is essential that the data collection is preceded by calibration of the raters' to obtain and preserve adequate agreement between the examiners.

REFERENCES

1. Angelillo IF, Anfosso R, Nobile CG, Pavia M. Prevalence of dental caries in schoolchildren in Italy. Eur J Epidemiol 1998;14:351-357.
2. Arcella D, Ottolenghi L, Polimeni A, Leclercq C. The relationship between frequency of carbohydrates intake and dental caries: a cross-sectional study in Italian teenagers. Public Health Nutrition 2002;5:553-560.
3. Assaf AV, Moneghim MC, Zanin L, Mialhe FL, Pereira AC, Ambrosano GMB. Assessment of different methods for diagnosing dental caries in epidemiological surveys. Community Dent Oral Epidemiol 2004;32:418-425.
4. Bolin AK, Bolin A, Jansson L, Calltorp J. Children's dental health in Europe. Swed Dent J 1997;21:25-40.
5. Bolin AK, Bolin A, Koch G, Alfredsson L. Children's Dental Health in Europe. Clinical calibration of dental examiners in eight EU countries. Swed Dent J 1995;19:183-193.
6. Campus G, Lumbau A, Lai S, Solinas G, Castiglia P. Socio-economic and behavioural factors related to caries in twelve-year-old Sardinian children. Caries Res 2001;35:427-434.
7. Campus G, Senna A, Cagetti MG, Maida C, Strohmenger L. Caries experience and status in school-age children residing in four north italian communities. Oral Health Prev Dent 2005;1:33-38.
8. Fleiss JL. Reliability of measurements. In: Fleiss JL. The design and analysis of clinical experiments. New York, John Wiley and Sons, 1986;1-32.
9. Fyffe HE, Deery C, Nugent ZJ, Nuttall NM, Pitts NB. Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). Community Dent Oral Epidemiol 2000;28:42-51.
10. Hunt RJ. Percent Agreement, Pearson's correlation, and Kappa as measures of inter-examiner reliability. J Dent Res 1986;65:128-130.
11. ISTAT. Codici dei Comuni, delle Province e delle Regioni. www.istat.it/strumenti/definizioni/comuni, 2004. Accessed ????????.
12. Jamieson LM, Thomson WM, McGee R. An assessment of the validity and reliability of dental self-report items used in a National Child Nutrition Survey. Community Dent Oral Epidemiol 2004;32:49-54.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
14. Lesaffre E, Mwalili SM, Declerck D. Analysis of caries experience taking inter-observer bias and variability into account. J Dent Res 2004;83:951-955.
15. McPartland JM, Goodridge JP. Counterstrain and traditional osteopathic examination of the cervical spine compared. J Bodywork Movement Ther 1997;1:173-178.
16. Pine CM, Adair PM, Nicoll AD, Burnside G, Petersen PE, Beighton D et al. International comparisons of health inequalities in childhood dental caries. Community Dent Health. 2004;21:121-130.
17. Pine CM, Pitts NB, Nugent ZJ. BASCD guidance on the statistical aspects of training and calibration of examiners for surveys of child dental health. Community Dent Health 1997;14:18-29.
18. Strohmenger L, Brambilla E, Sardi E, Sardo Infirri J. Dental caries prevalence in 12-year-old children in Milan (Italy). Riv Ital Odontoiatr Infant 1997;4:47-49.
19. Vogel G, Strohmenger L, Barozzi G, Malerba A. Dental and periodontal epidemiology of the school population in 3 Italian regions. Minerva Stomatol 1986;35:565-570.
20. WHO (World Health Organization). Oral Health Surveys: Basic Methods. Geneva: WHO, 1997.