BRIEF COMMUNICATION

Dean's Fluorosis Index: an Assessment of Examiner Reliability

Jayanth V. Kumar, DDS, MPH; Philip A. Swango, DDS, MPH; Pandy N. Opima, DDS, MPH; Elmer L. Green, DDS, MPH

Abstract

Objectives: *This report describes the interexaminer reliability achieved using Dean's Index in a study of dental fluorosis, and shows the effect on kappa values of assigning different weights to the various components of Dean's Index.* ***Methods:*** *Three dentists conducted replicate fluorosis examinations on 202 children in Newburgh and Kingston, NY. Examiner reliability was assessed by computing percent agreement and weighted and unweighted kappa statistics.* ***Results:*** *Agreement on the presence or absence of fluorosis using Dean's definition of fluorosis ranged from 92 to 97 percent and the respective kappa values ranged from 0.75 to 0.94. A comparison of subject-level severity scores for Dean's Index resulted in percent agreement ranging from 79.6 percent to 86.8 percent and kappa values ranging from 0.67 to 0.76. Weighting the kappa statistics improved agreement and reduced the differences.* ***Conclusions:*** *Examiners showed good to excellent agreement beyond chance in the use of the index. Subject level kappa scores were higher than tooth-level scores. [J Public Health Dent 2000;60(1):57-59]*

Key Words: reliability, dental fluorosis, Dean's Index, kappa statistics.

Dean's Index of dental fluorosis has been in continuous use for more than 50 years (1). Since the time of Dean's original work, several other indices of dental fluorosis have been developed to address some of the shortcomings of Dean's Index when used in analytic epidemiology (2-5). However, Dean's Index has remained popular for prevalence studies because of its simplicity and its ability to make comparisons with numerous earlier studies.

Rozier has suggested that due to the subjective nature of classification, assessments of fluorosis may be subject to more variation than those of many other oral conditions (5). However, studies of examiner reliability of Dean's Index have not been widely reported. In 26 studies using Dean's Index conducted between 1980 and 1992, Rozier found that only 11 included estimates of examiner reliability (5), and most reported only percent agreement scores rather than the generally more meaningful kappa statistic. Mabelya et al. (6) compared Dean's Index to the Thylstrup-Fejerskov Index and found the overall kappa scores to be 0.69 and 0.70, respectively. Evans reported a kappa score of .61, and Songpaisan and Davies reported a kappa score of .35 (5). The latter score falls below the range of kappa values generally regarded as representing acceptable agreement among examiners. Consequently, it is difficult to generalize from the existing literature about the reliability of Dean's Index or other indices of fluorosis.

This report presents findings for interexaminer reliability in a study of dental fluorosis conducted in Newburgh and Kingston, NY. The objectives of this report are to present the levels of interexaminer agreement achieved during the study, and to determine the effect of reducing the seriousness of disagreement by assigning different weights to the various components of Dean's Index.

Methods

Data for this communication were obtained in a study of 3,326 children, grades 1–8, who attended schools in fluoridated Newburgh or in nonfluoridated Kingston. Data from 202 replicate examinations conducted by two experienced dental epidemiologists and a third dentist who had no previous experience in epidemiologic studies were used in the assessment of reliability of Dean's Index of fluorosis. These examiners were trained in the fluorosis evaluations by a consultant with extensive experience in the use of Dean's Index. Training consisted of a didactic presentation of Dean's classification criteria using color transparencies and discussions of fluorosis cases in the field at a study site.

Field examinations were conducted using portable fiber optic lights; teeth were not air-dried prior to the fluorosis assessments. Subjects were first given an overall score for Dean's Index based on the two teeth displaying the most severe signs of fluorosis, with the overall subject score being determined by the score assigned to the least severely affected of the two index teeth. Following the overall score, each fully erupted tooth was scored individually. A distinction was made between dental fluorosis and other opacities that appeared to be nonfluoride in origin, as described by Russell (7). The prevalence of questionable and very mild to severe dental fluorosis in these children was 15 percent and 21 percent, respectively. At the tooth level, the prevalence of

Send correspondences and reprint requests to Dr. Kumar, Bureau of Dental Health, New York State Department of Health, Empire State Plaza, Albany, NY 12237-0619. E-mail: jvk01@health.state.ny.us. Dr. Green is with the New York State Department of Health, Albany, and the School of Public Health, University at Albany. Dr. Swango is a consultant in Albuquerque, NM. Dr. Opima was a resident in the New York State Dental Public Health Residency Program. Web site: www.health.state.ny.us. This study was supported by a grant from the National Institute for Dental Research (R01 DE 1088801). This paper was presented at the 74th general session of the International Association for Dental Research. Manuscript received: 11/6/98; accepted for publication: 12/9/98.

questionable and very mild to severe dental fluorosis was 9 percent each.

Using an SAS computer program (8), percent agreement on Dean's classification and the respective kappa scores were computed. Examiner agreement was determined for each pair of examiners. Analyses were carried out first at the level of the individual subject, next for all permanent teeth combined, and then for subsets of permanent teeth composed of the incisors and first molars. A final analysis was carried out to obtain percent agreement on the simple presence or absence of fluorosis (prevalence), without regard to severity. Procedures for computing the unweighted kappa statistics are those described by Fleiss (9). Because disagreements between the various categories of Dean's Index may not be of equal importance (depending on the study objectives), a weighting scheme was devised that would assign less emphasis when examiners disagreed by only one severity level of Dean's Index, and greater emphasis when they disagreed by more than one severity level (Table 1). The procedure described by Cohen (10) was used for deriving weighted agreement and kappa values.

To provide a qualitative meaning to the numeric kappa values, the guidelines suggested by Landis and Koch (11) were used. Kappa values of less than 0.4 indicate unacceptable agreement beyond chance between examiners, and values of 0.75 or greater denote excellent agreement beyond chance. Values between 0.4 and 0.75 represent fair to good agreement beyond chance.

Results

Table 2 shows the percent agreement and the respective kappa scores among the three examiners. At the subject level of comparison of Dean's Index, unweighted percent agreement and the unweighted kappa values ranged from 79.6 to 86.8 percent and 0.67 to 0.76, respectively. The weighted percent agreement and the corresponding kappa values exceeded 88.9 percent and 0.76, respectively.

When all teeth, rather than subjects, were considered as the units of observation, percent agreement scores were relatively unaffected, but kappa values were reduced. Unweighted kappa and weighted kappa values ranged from 0.52 to 0.65 and 0.59 to 0.75, respectively. To evaluate reliability by

TABLE 1
Weighting Matrix Used for Computing Weighted Kappa Statistics for Dean's Index

	Normal	Questionable	Very Mild	Mild	Moderate	Severe	Exclude
Normal	1	.5	0	0	0	0	0
Questionable	.5	1	.5	0	0	0	0
Very mild	0	.5	1	.5	0	0	0
Mild	0	0	.5	1	.5	0	0
Moderate	0	0	0	.5	1	.5	0
Severe	0	0	0	0	.5	1	0
Exclude	0	0	0	0	0	0	1

TABLE 2
Interexaminer Reliability Using Dean's Index, Unweighted and Weighted Percent Agreement and Kappa Statistics, by Examiner and Type of Analysis

		Unweighted					Weighted		
	n	Examiner A vs B	*n*	Examiner A vs C	*n*	Examiner B vs C	Examiner A vs B	Examiner A vs C	Examiner B vs C
Subjects	68		80		54				
Agreement (%)		86.8		82.5		79.6	92.7	91.3	88.9
Kappa		0.76		0.67		0.68	0.83	0.78	0.76
All teeth	1,018		1,726		978				
Agreement (%)		87.4		88.2		84.7	92.8	93.3	91.1
Kappa		0.65		0.53		0.52	0.75	0.63	0.59
Incisors	272		320		216				
Agreement (%)		84.9		85.3		80.1	90.3	91.6	88.2
Kappa		0.68		0.60		0.59	0.77	0.70	0.71
First molars	272		320		216				
Agreement (%)		88.2		82.8		80.1	93.4	89.1	88
Kappa		0.76		0.52		0.58	0.84	0.62	0.64
Prevalence	68		80		54				
Agreement (%)		97.0		92.0		96.3	—	—	—
Kappa		0.94		0.75		0.9	—	—	—

Note: Prevalence of fluorosis is based on Dean's classification of very mild to severe categories.

tooth type, analyses were carried out separately for permanent maxillary incisors and permanent first molars. These results did not show any clear-cut overall conclusions with respect to reliability by tooth type. Rather, the results appeared to vary by examiners. For incisors, Examiner A versus Examiner B displayed an unweighted kappa of 0.68, while the value for Examiner B versus Examiner C was 0.59. The weighted analysis improved these scores to 0.77, 0.70, and 0.71 for the three pairs of examiners. On first molars, both the unweighted and weighted percent agreement kappa scores were slightly higher for Examiner A versus Examiner B when compared to the other pairs. Overall, Examiner A showed slightly better agreement and higher kappa scores on every comparison at the subject level and on tooth level scores. The improved percent agreement and kappa values in the weighted analyses illustrate that the examiners' disagreement tended to be less than one severity level while using Dean's classification.

Examiner agreement on the presence or absence of dental fluorosis was assessed using the criterion proposed by Dean, where presence of fluorosis is defined as a subject having very mild or greater severity of fluorosis. Agreement on the presence or absence of fluorosis ranged from 92 to 97 percent and the respective kappa values ranged from 0.75 to 0.94.

Discussion

The assessment of the reliability of Dean's Index showed excellent agreement beyond chance when subjects were classified with respect to presence or absence of fluorosis. This type of analysis would be appropriate in studies to determine the prevalence of fluorosis, without regard to degrees of severity, or in studies of fluorosis using logistic regression, in which the dependent variable must be dichotomized. Dean's subject level scores also showed excellent levels of agreement. Even our inexperienced examiner achieved good agreement with other examiners after proper training.

However, differences among the examiners were noted, even though they were not great in most instances. Agreement was better between examiners when subjects were the unit of observation rather than teeth. This probably was due to several factors. For one, fewer diagnostic decisions were required. For another, some classifications applicable to individual teeth, such as unerupted and nonfluoride opacity, are easier to record at the person level. Finally, the characteristic clinical features of dental fluorosis such as its bilateral manifestation and generalized presence of opacities are easier to observe at the person level. In short, fewer chances for examiners were available to disagree on subject scores as compared to tooth scores.

We expected to find a higher level of agreement on incisors than molars; however, differences were not consistent among the examiners. It should be noted that Dean's criteria for some milder forms of fluorosis make specific reference to tooth type, e.g., "snow-capping" on cusp tips of molars and premolars. It may be that the examiners applied Dean's criteria somewhat differently for molars than for incisors.

The use of weighted kappas resulted in improved agreement over the unweighted values and reduced the differences among the examiners. The question arises as to whether the unweighted or the weighted analysis gives a more appropriate measure of reliability. In prevalence studies, where the chief parameter of interest is the proportion of persons affected by fluorosis without regard to severity, a weighted analysis might be considered more meaningful. However, in studies where all disagreements about severity are considered equally important, unweighted analyses would be more appropriate.

A second question is how to choose the component weights. In describing the weighted kappa analysis, Cohen proposed that weights be assigned on rational or clinical grounds, independent of the data actually collected (10). Thus, the relative seriousness of each possible disagreement can be quantified, where exact agreement is given a weight of 1 and all disagreements are given a weight of less than 1. In our weighting scheme, a weight of 0.5 was assigned when examiners disagreed by only one severity level of Dean's Index, and a weight of 0 was assigned to disagreements of more than one severity level. This scheme is somewhat arbitrary and it simply reflects our best estimate of the relative importance of the disagreements in the context of the study's objectives.

This study showed good to excellent interexaminer agreement beyond chance for Dean's Index. The results also compare favorably with the previous reports of reliability of fluorosis examinations (5). However when more categories were used for classifying fluorosis, it invariably resulted in lower agreement levels. Although the mere fact that Dean's Index has fewer categories compared to other indices of fluorosis should be advantageous in attaining excellent agreement levels, Mabelya et al. (6) found the Thylstrup and Fejerskov Index to be slightly better. They attributed difficulties encountered in the application of Dean's Index to not drying the teeth before clinical examination and imprecise definitions of categories (6).

References

1. Dean HT. The investigation of physiological effects by the epidemiological method. In: Moulton FR, ed. Fluorine and dental health. Washington, DC: American Association for the Advancement of Science, 1942;pub no 19:23-31.
2. Thylstrup A, Fejerskov O. Clinical appearance and surface distribution of dental fluorosis in permanent teeth in relation to histological changes. Community Dent Oral Epidemiol 1978;6:315-19.
3. Horowitz HS, Heifetz SB, Driscoll WS, Kingman A, Myers RJ. A new method for assessing the prevalence of dental fluorosis: the Tooth Surface Index of Fluorosis. J Am Dent Assoc 1984;109:37-41.
4. Pendrys DG. The Fluorosis Risk Index: a method for investigating risk factors. J Public Health Dent 1990;50:291-8.
5. Rozier RG. Epidemiologic indices for measuring the clinical manifestations of dental fluorosis: overview and critique. Adv Dent Res 1994;8:39-55.
6. Mabelya L, van't Hof MA, Konig KG, van Palenstein Helderman WH. Comparison of two indices of dental fluorosis in low, moderate, and high fluorosis Tanzanian populations. Community Dent Oral Epidemiol 1994;22:415-20.
7. Russell AL. Differential diagnosis of fluoride and nonfluoride enamel opacities. Public Health Dent 1961;21:143-6.
8. SAS Institute Inc. SAS users' guide statistics. Version 6. 4th ed. Cary, NC: SAS Institute, 1990.
9. Fleiss J. The measurement of interrater agreement. In: Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons, 1981.
10. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.