Community Dental Health (2007) **24,** 21-25
Received 17 January 2005; Accepted 30 August 2005

Agreement amongst examiners assessing dental fluorosis from digital photographs using the TF index.

J. Tavener, R.M. Davies and R.P. Ellwood

Dental Health Unit, Skelton House, Manchester Science Park, Lloyd St North, Manchester, M15 6SH, England.

Objective To compare the scoring of dental fluorosis by experienced examiners from digital photographs using the TF index. ***Basic Research design*** 120 images were selected from 703 photographs obtained during a clinical trial (Tavener *et al*., 2004). The selection process was stratified so that the full range of defects seen in the main study was included. The children, aged 8-10 years, were from deprived areas of Manchester, England with fluoride levels in the drinking water of less the 0.1 ppm F. The photographs of the upper and lower anterior sextants were taken after cleaning and drying the teeth. The examiners were identified by searching Medline for individuals who had previously used the TF index or had experience of scoring dental fluorosis. Of the 12 examiners identified, 10 agreed to take part. Each examiner was provided with identical CDs containing a PowerPoint presentation of the images. Twelve images were duplicated and interspersed amongst the 120 images to assess intra examiner agreement. Each examiner was also supplied with a table listing the criteria and illustrations for each of the TF index scores (Fejerskov *et al*., 1988). ***Results*** The prevalence of fluorosis (TF>0) amongst the 10 examiners ranged from 43% to 70% and from 2% to 13% for the more severe scores (TF 3 or 4). Paired agreements amongst subject scores for the 10 examiners, measured using a weighted Kappa score, ranged from 0.40 to 0.71. ***Conclusion*** It is concluded that although the criteria for the TF index are well defined, it is possible that examiners may interpret the criteria in different ways and conditions in which images are viewed may need to be standardised. This study may explain some of the differences in the prevalence and severity of fluorosis reported in different studies. There is a need to standardise the methods used to score dental fluorosis.

Key words: dental fluorosis, digital images, reproducibility, TF index

Introduction

The efficacy of fluoride in preventing and treating dental caries has contributed significantly to the decline in dental caries experienced in many parts of the world. However, inappropriate use of fluoride increases the risk of dental fluorosis. Mild fluorosis, characterised by thin white lines or diffuse frosting (hypomineralisation) of the enamel surface may only be recognised by experienced examiners after thorough drying of the teeth. At high levels of fluoride exposure the formation of the enamel may be so disrupted that the entire surface is lost but such quantitative defects are very rare in most populations.

The methods used to record dental fluorosis have been the subject of great debate with many epidemiologists unable to agree whether dental fluorosis can be differentiated from other forms of developmental enamel effects. This has led to the introduction of a wide variety of indices of which the three most widely used are the DDE Index (FDI, 1992), Dean's Index (Dean, 1942) and the TF index (Thylstrup and Fejerskov, 1978).

The DDE Index records all enamel defects but does not attempt to ascribe a cause. Thus when conducting studies focused specifically on the effects of fluoride in a population it may not be the most appropriate method to apply. The fluorosis specific Dean's Index has been used for more than 60 years and there is a wealth of data enabling comparisons between studies in a wide range of populations. However, some concerns with the criteria employed by Dean's Index prompted the development of the TF Index. This index produces an ordinal score that has been histologically validated against the degree of hypomineralisation and is particularly useful in populations with low levels of fluoride exposure as the teeth are dried and the criteria allow differentiation of the mildest forms of fluorosis. The index is well documented with clear descriptive criteria for each of the scores supplemented by line drawings and photographs.

Various factors, other than the type of index used, may influence the prevalence of fluorosis measured in a particular population. For example, drying the teeth improves the contrast between normally mineralised and fluorotic enamel and the angle of viewing and the lighting conditions are also important.

It might be expected that, when using the same index and methods, populations with broadly similar levels of fluoride intake would have similar levels of fluorosis but this is not always the case. For example, Hamdan and Rock (1991) reported a prevalence of fluorosis of 8% in a population with drinking water containing 0.1 ppm F and 26% in a population with 1 ppm F in the drinking water. In contrast, Ellwood and O'Mullane (1994) reported a prevalence of 36% in a non-fluoridated and 54% in a fluoridated population. Clearly, some of the difference between studies may be due to differences in other fluoride exposure such as toothpaste and fluoride

Correspondence to: Dr. Roger P Ellwood , Dental Health Unit, Skelton House, Manchester Science Park, Lloyd St North, Manchester, M15 6SH, England. E-mail: roger.ellwood@manchester.ac.uk

tablets but it is also likely that differences may be due to the way fluorosis is recorded by examiners.

Several studies have used photographs to capture images of teeth that can then be scored for fluorosis (Cochran *et al*., 2004; Holt *et al*., 1994; Ellwood and O'Mullane 1995; Levine *et al*., 1989; Sabieha and Rock, 1998). This has a number of advantages which include; standardisation of the viewing conditions, the ability to score the images using a variety of methods and examiners, the ability to mix photographs taken at different times so that images from different populations might be scored randomly, the ability to easily blind examiners to the area of residence of subjects, and the opportunity to archive photographs so that images from different studies, or those taken at different times, can be compared.

In a recent study (Tavener *et al*., 2004) digital photographs of 703 children were recorded providing a range of images encompassing the range of fluorosis generally seen in populations with a low fluoride exposure from the drinking water. During the course of scoring these photographs it became apparent that different examiners interpreted the criteria of the TF index in quite different ways. It is possible therefore that variation in the application of criteria rather than true differences between populations might explain some of the differences between studies performed in populations with apparently similar levels of fluoride exposure.

The aim of this study was to compare the agreement amongst experienced examiners scoring dental fluorosis from digital photographs using the TF index.

Methods

The 120 images were selected from 703 photographs taken in a previously reported clinical trial (Tavener *et al*., 2004). The selection process was stratified so that the full range of defects seen in the main study was included. The children were aged 8-10 years when photographed and from deprived areas of Manchester, England with fluoride levels in the drinking water of less the 0.1 ppm F.

Standardised digital photographs of the upper and lower anterior sextants were taken by an experienced examiner (JT). In order to minimise specula reflection, the photographs were taken from approximately 15 degrees above the perpendicular using a Fuji Finepix S1 Pro with a Micro Nikkor 105mm lens and Nikon SB 21 ring flash using the top element for illumination. The reproduction ratio was set for 1:1 or life-size. The upper incisor teeth were wiped with a cotton wool roll and allowed to air dry for one minute.

Examiners were identified using a Medline search for individuals who had previously used the TF index or had experience of scoring dental fluorosis. Of the 12 examiners identified, 10 agreed to participate. Each examiner was provided with identical CD's containing a PowerPoint presentation showing the images. Twelve of the 132 images were duplicates of the original images and interspersed within the presentation to assess intra examiner agreement. Each examiner was provided with the criteria and illustrations for each of the TF index scores (Fejerskov *et al*., 1988). Printed Microsoft Access scoring sheets were included for examiners to record their scores and return by post. Instructions requested that each image be scored dichotomously for the presence or absence of fluorosis. Ssubsequently, if scored positively, examiners were requested to ascribe a severity score using the Thylstrup and Fejerskov ordinal scale for each central incisor (Thylstrup and Fejerskov, 1978). The criteria for this index may be summarised as: TF1: thin white lines running across the tooth surface, TF2: pronounced white lines, TF3: merging of white lines and cloudy areas of opacity and TF4 the entire surface is chalky white, TF scores of five or more are associated with pitting and enamel loss but none were found in the population examined.

Statistical analysis

For each subject the highest TF score on either of the upper central incisors was used in analysis. In addition to the examiner based assessments four summary variables were calculated from the subject level scores. These were minimum and maximum scores recorded by any of the ten examiners and the modal and mean scores for the ten examiners. In the case of the mean score the number was rounded to the nearest whole number. The frequency distribution of TF scores for each examiner and the summary variables were tabulated.

The subject level scores for each of the examiners were compared using a weighted Kappa statistic (Fleiss, 1981). A score of less than 0.2 is considered poor agreement, 0.2 to 0.4 fair agreement, 0.4 to 0.6 moderate agreement, 0.6 to 0.8 good agreement and 0.8 to 1 very good agreement. (Landis and Koch, 1977)

Ethical approval for the study was obtained from the Local Research Ethics Committee. Passive consent for participation in the study was sought using pre-paid post cards at the start of the study and again for participants to have their upper anterior teeth photographed in school.

Results

A total of 120 photographs were scored by each of the 10 examiners. Repeat examinations of 12 sets of photographs for each examiner yielded intra examiner weighted kappa scores ranging from 0.25 to 0.85 (Table 1). The standard errors for the assessments were large (0.19 to 0.25) so that differences amongst examiners did not attain statistical significance.

The prevalence of fluorosis (TF score>0) ranged from 43% to 70% (Table 1). A total of 69% of images were scored as having no fluorosis by one or more examiners, but only 17% were scored as fluorosis free by all examiners. Using the modal score for each subject for the ten examiners 57% of subjects had fluorosis and using the mean score 59% had fluorosis. The prevalence of TF scores 3 or 4 ranged from 2% to 13%. No subjects were scored as having either TF score 3 or 4 by all examiners but 24% were scored as TF 3 or 4 by one or more of the examiners. There was no correlation between the prevalence of fluorosis (TF>0) and the prevalence of the more severe scores (TF 3 or 4) for the ten examiners (r^2=0.01, p>0.05).

Table 1. Frequency distribution TF scores for individual examiners and the minimum, maximum, mode and mean score for the ten examiners.

		TF Score					*Intra examiner*
Examiner		*0*	*1*	*2*	*3*	*4*	*Kappa (SE)*
1	n	43	31	35	9	2	0.81
	%	*36*	*26*	*29*	*8*	*2*	*(0.23)*
2	n	60	26	24	10	0	0.53
	%	*50*	*22*	*20*	*8*	*0*	*(0.21)*
3	n	55	25	27	10	3	0.85
	%	*46*	*21*	*23*	*8*	*3*	*(0.21)*
4	n	53	39	18	7	3	0.5
	%	*44*	*33*	*15*	*6*	*3*	*(0.19)*
5	n	68	29	17	6	0	0.59
	%	*57*	*24*	*14*	*5*	*0*	*(0.24)*
6	n	49	22	34	14	1	0.49
	%	*41*	*18*	*28*	*12*	*1*	*(0.22)*
7	n	50	25	28	16	1	0.25
	%	*42*	*21*	*23*	*13*	*1*	*(0.21)*
8	n	55	45	18	2	0	0.58
	%	*46*	*38*	*15*	*2*	*0*	*(0.21)*
9	n	36	51	29	4	0	0.70
	%	*30*	*43*	*24*	*3*	*0*	*(0.22)*
10	n	40	42	33	5	0	0.69
	%	*33*	*35*	*28*	*4*	*0*	*(0.21)*
Lowest score by 1 or more examiners	n	83	32	5	0	0	
	%	*69*	*27*	*4*	*0*	*0*	
Highestscore by 1 or more examiners	n	20	35	36	24	5	Not applicable
	%	*17*	*29*	*30*	*20*	*4*	
Mode score	n	52	33	30	5	0	
	%	*43*	*28*	*25*	*4*	*0*	
Mean score	n	49	40	25	6	0	
	%	*41*	*33*	*21*	*5*	*0*	

Table 2. Percentage agreement and weighted kappa agreement (standard error) for paired comparisons amongst the ten examiners.

Examiner	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1		62%	54%	69%	60%	62%	56%	65%	65%	63%
2	0.59 (0.06)		77%	70%	65%	63%	71%	71%	65%	55%
3	0.52 (0.06)	0.77 (0.07)		59%	58%	59%	64%	64%	55%	55%
4	0.64 (0.06)	0.67 (0.07)	0.56 (0.06)		63%	61%	63%	72%	71%	58%
5	0.46 (0.06)	0.54 (0.07)	0.47 (0.06)	0.52 (0.06)		58%	62%	68%	55%	53%
6	0.57 (0.07)	0.60 (0.07)	0.58 (0.07)	0.57 (0.06)	0.46 (0.06)		68%	59%	59%	66%
7	0.51 (0.07)	0.63 (0.07)	0.60 (0.07)	0.58 (0.06)	0.51 (0.06)	0.65 (0.07)		58%	60%	59%
8	0.56 (0.06)	0.64 (0.07)	0.56 (0.06)	0.64 (0.06)	0.56 (0.07)	0.51 (0.06)	0.49 (0.06)		71%	58%
9	0.59 (0.06)	0.63 (0.06)	0.53 (0.06)	0.66 (0.06)	0.45 (0.06)	0.54 (0.06)	0.54 (0.06)	0.64 (0.06)		61%
10	0.54 (0.06)	0.46 (0.07)	0.45 (0.06)	0.47 (0.06)	0.40 (0.06)	0.63 (0.06)	0.50 (0.06)	0.46 (0.06)	0.49 (0.07)	

PERCENTAGE AGREEMENT

WEIGHTED KAPPA (STANDARD ERROR)

The comparison of the scores by the 10 examiners is shown in Table 2. The top right half of the table shows percentage agreement and the bottom left half the weighted kappa scores (standard error) for all possible pairings of examiners. The highest percentage agreement was 77% between examiners 2 and 3. The lowest agreement was 53% between examiners 5 and 10. Weighted kappa scores ranged from 0.40 between examiners 5 and 10 to 0.77 between examiners 2 and 3.

Discussion

The results of this study suggest that even using a well described index, such as the TF index, interpretation of criteria is highly subjective. For the ten examiners the prevalence of dental fluorosis ranged from 43 to 70% with the prevalence of the higher scores (TF 3 or 4) varying from 2-13%. These differences are of a similar magnitude to those seen between populations with and without fluoridated drinking water. Fluorosis (TF >0) was recorded by one or more examiners in 83% of subjects but 69% of subjects were also scored as having no fluorosis by one or more of the examiners. Similar inconsistencies were seen for the higher TF scores. This suggests that increasing the threshold to exclude minor defects, a method commonly used in epidemiology to improve consistency amongst examiners, would not improve diagnostic consistency. Thus examiners had difficulties agreeing on the presence or absence of fluorosis and its severity. It might be expected that examiners scoring a high prevalence of defects might also score a high severity but this was not the case. There was no correlation between the prevalence of fluorosis and its severity for the examiners.

It is interesting to note that the examiners scoring the highest and lowest prevalence of fluorosis achieved a weighted kappa score of 0.40, which is described as moderate agreement. Despite this, prevalences of fluorosis were 70% and 43% for the two examiners. In this case, however, the two examiners scored similar prevalences of scores TF 3 or 4 with 5% and 4% respectively. At TF score 3 or 4 the highest prevalence scored was 13% and the lowest 2% but the overall agreement for the two examiners was 0.49, which again would be described as moderate agreement. It is possible that classification of subjects based on the highest score of the two maxillary central incisors might introduce some bias and reduce the agreement amongst examiners. However, subject level outcome measures need to be used in statistical analysis and in 89% of cases the scores of left and right teeth were the same.

Overall the examiners did not demonstrate good intra-examiner reliability with six of the ten examiners having intra-examiner kappa scores of less than 0.6. Clearly when the examiners themselves are inconsistent in their scoring it is difficult to compare meaningfully between examiners.

Although care was taken to ensure that examiners scored the images in as similar conditions as practical, it is possible that at least part of the difference in scores between the examiners might be explained by differences in the way the images were viewed. For example the contrast and brightness of different monitors might be expected to modify the ability to detect lesions. The magnitude of the effect that differences in the way images were viewed might affect the prevalence and severity of fluorosis was not explored in this study and this problem needs further research. It is possible prints of images might be used for example to reduce the effect of monitor and viewing conditions. In cross-sectional studies the ability to be able to compare groups might not be affected by these differences but comparisons between studies must be treated with caution.

The TF index was developed as a clinical index in areas of high concentrations of fluoride in the water, and it is noted by the developers that difficulties exist when applying the index to cases when fluoride exposure has not been continuous throughout amelogenesis. (Fejerskov *et al*., 1988). The authors suggested that a tentative diagnosis can be made in such cases, which may be confirmed by careful history taking. This presents challenges for researchers attempting to use this index in blind studies and is impossible when scoring from images. The digital images used in this study present a 'key hole' view of the dentition. Differential diagnosis of fluorosis is aided by being able to view the complete dentition and, if this had been possible, may have improved the agreement between the examiners. Repeating the study using a whole mouth clinical scoring methodology would be of benefit, but was prohibited by logistics. It might also be argued that scoring the anterior teeth provides the most valuable information, as these teeth are most important aesthetically.

Taken overall these results are worrying and suggest caution is required when reviewing trends in fluorosis prevalence and differences between populations, particularly when studies using different methods and examiners are compared. Consistency of scoring between examiners might be improved by the introduction of a training set of images showing a wide range of presentations of each lesion type. It would be enlightening to repeat this study using other indices in particular Dean's index to see if similar problems apply.

Acknowledgment

The authors would like to thank, Ole Fejerskov, John Clarkson, Keith Milsom, Peter Rock, Elizabeth Treasure, Andrew Rugg-Gunn, Roger Ellwood, Jacqui Tavener, Deirdre Browne and Rose Kingston for scoring the images and helpful advice in conducting this study.

References

Cochran, J.A., Ketley, C.E., Sanches, L., Mamai-Homata, E., Arnadottir, I. B., Loveren, C., Whelton, H.P., O'Mullane, D.M. (2004): A standardised photographic method for evaluating enamel opacities including fluorosis. *Community Dentistry and Oral Epidemiology* **32** (Suppl 1), 19-27.

Dean, H.T. (1942): The investigation of physiological effects by the epidemiological method. In: Moulton FR, editor. Fluorine and Dental Health. Washington (DC) *American Association for the Advancement of Science*. 23-31.

Ellwood, R.P. and O'Mullane, D.M. (1994): Association between enamel opacities and dental caries in a North Wales population. *Caries Research* **28**, 383-387.

Ellwood, R.P. and O'Mullane, D.M. (1995): Dental enamel opacities in three groups with varying levels of fluoride in their drinking water. *Caries Research* **29**, 137-142.

Federation Dentaire Internationale. (1992): A review of the developmental defects of enamel index (DDE index). *International Dental Journal* **42**, 411-426.

Fejerskov, O., Manji, F., Baelum, V., Moller, I.J. (1988): Dental fluorosis ; a Handbook for Health Workers. Copenhagen. Munksgaard.

Fleiss JL. Statistical Methods for Rates and Proportions (2nd edition). New York: Wiley 1981.

Hamdan, M. and Rock W.P. (1991): The prevalence of enamel mottling on incisor teeth in optimal fluoride and low fluoride communities in England. *Community Dental Health* **8**, 111-119.

Holt, R.D., Morris, C.E., Winter, G.B., Downer, M.C.(1994): Enamel opacities and dental caries in children who used a low fluoride toothpaste between 2 and 5 years of age. *International Dental Journal* **44,** 331-341.

Landis JR, Koch G. (1977):The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

Levine, R.S., Beal, J.F., Fleming, C.M. (1989): A photographically recorded assessment of enamel hypoplasia in fluoridated and non-fluoridated areas in England. *British Dental Journal* **166**, 249-252.

Sabieha, A. M. and Rock, W.P.(1998): A comparison of clinical and photographic scoring using the TF and modified DDE Indices. *Community Dental Health* **15**, 82-87.

Tavener, J.A., Davies, G.M., Davies, R.M., Elwood, R.P. (2004): The prevalence and severity of fluorosis and other developmental defects of enamel in children who received free fluoride toothpaste containing either 440 or 1450 ppm F from the age of 12 months. *Community Dental Health* **21**,217-223.

Thylstrup, A. and Fejerskov, O. (1978): Clinical appearance of dental fluorosis in permanent teeth in relation to histological changes. *Community Dentistry and Oral Epidemiology* **6**, 315-328.