

# The construction and use of diagnostic standards for primary radiographic caries

OLA HAUGEJORDEN

Dental Division, Directorate of Health, Oslo, Norway

Haugejorden, O. The construction and use of diagnostic standards for primary radiographic caries. *Acta Odont. Scand.* 34, 69—77, 1976.

The aim was to establish diagnostic standards for primary approximal radiographic caries diagnosis. These were described with reference to ranking of subjects, mean »intermediate» score per subject, which was based on the majority diagnosis recorded by three examiners, and by category of a certainty scale for individual tooth surface scores. The results recorded by two fourth year dental students and two hygienists, when reading 40 pairs of posterior bitewing radiographs of children between 13 and 15 years of age, showed an acceptable level of agreement for ranking of subjects. Agreement was less satisfactory, when comparing the mean scores of these four examiners with the »intermediate» score per subject and with a standard for individual surfaces. The proposed standards do not represent true scores, but serve as guidelines, as they depend on the diagnostic level of the examiners on whose scores the standards are based, and on the diagnostic criteria employed.

*Key-words:* Dental caries; diagnosis; oral; radiography; dental

*Ola Haugejorden, Institute of Community Dentistry, University of Oslo, Blindern, Oslo 3, Norway*

Only short references have been made in the literature to measures taken to ensure maintenance of a constant caries diagnostic level in periodic cross-sectional and longitudinal studies (*Backer-Dirks et al.*, 1961, 1967; *Horowitz et al.*, 1973; *Haugejorden*, 1974). Thorough pre-survey training of examiners to achieve uniform interpretation and application of diagnostic criteria, combined with continuous or periodic diagnostic reproducibility checks throughout the survey period have been the methods most commonly reported. When it has been necessary to replace examiners during the course of a survey, the new examiner has usually been calibrated to some experienced person to

whom the retiring examiner had also been calibrated. This approach is satisfactory provided evidence can be presented to show that once thoroughly trained to record caries at one diagnostic level, an examiner will continue to do so for long periods of time, e.g. 10 years or more as implied in some investigations (*Arnold et al.*, 1962).

Since visual acuity decreases with age and knowledge of a disease process may change with time, it would appear that reliance on the experience of an examiner is inadequate, as his diagnostic reproducibility may remain constant within relatively narrow limits even though the recording level employing the same

diagnostic criteria may have changed. Standards for controlling that the diagnostic level is maintained are needed. Reference standards are described in the Scandinavian Moulage System (*Westin & Wold, 1943*). The use of extracted teeth as examples of diagnostic standards and the use of standard radiographs were recommended by *Backer Dirks et al. (1967)* and *Horowitz et al. (1973)*. Limited use seems to have been made of such aids to ensure maintenance of a constant clinical caries diagnostic level. The use of carious lesions in extracted teeth or photographs of them as reference standards for clinical caries diagnosis is possible, but not completely satisfactory since the tactile component of clinical diagnosis is not taken into account.

Radiographic caries diagnosis is a visual process. The properly produced and stored radiograph constitutes a permanent record (*Pfahler, 1908*). It may be re-examined at intervals over extended periods of time and may, therefore, serve as a reference standard for checking on maintenance of a constant radiographic caries diagnostic level.

Sets of radiographs have been reassessed at intervals during the examination period to guard against change of diagnostic level in a study by *Backer Dirks et al. (1961)*. The authors did, however, not report the magnitude of deviation from pre-survey training results which were tolerated before recorded data were rejected and the examiners required to submit for training to re-establish the previously agreed diagnostic level.

No other reports on the use of such safeguards against shifts of recording level were discovered.

The scope for comparisons between results of epidemiological studies is limited by the use of different caries diagnostic criteria, intra- and inter-examiner va-

riation in recording level. Description of principal requirements for controlled clinical trials of caries preventive agents and procedures (*Horowitz et al., 1973*) and standardization of basic oral health survey methods (*World Health Organization, 1971*) will facilitate more reliable comparisons of results of epidemiological investigations. However, it seems possible to extend this work further in relation to clinical trials by making full use of the advantages of radiographs (*Haugejorden, 1974*).

The purpose of this paper was to describe the construction of standards for primary approximal radiographic caries diagnosis. Separate standards will be presented for, (1) the ranking of subjects according to caries score; (2) the mean score per subject; (3) the status of individual tooth surfaces; and (4) results will be reported to illustrate the use of these standards.

#### MATERIAL AND METHODS

The material and methods used to obtain data for illustrating the construction and use of diagnostic standards for radiographic approximal caries has partly been reported by *Haugejorden & Slack (1975)*. Only a summary and emphasis of important details are therefore included here.

The 40 subjects were English school children who participated in clinical trials of fluoride containing dentifrices. They were between 13 and 15 years of age at the time when these radiographs were taken.

The 40 pairs of posterior bitewing radiographs were taken at the third and fourth examinations of three-year clinical trials. The films used were Kodak DF55 which had been exposed while using a modification of the film holder described by

*Backer Dirks, Van Amerongen & Winkler* (1951) (*Slack et al.*, 1967). The darkroom procedure had been standardised and was according to the manufacturer's recommendations.

The radiographs of each subject were mounted in transparent film mounts which were given code numbers and arranged in random order, in two cardboard boxes, to ensure identical order of reading by each examiner.

A Watson's radiographic viewing box with a fixed ( $\times 2$ ) magnifying lens was used throughout by all examiners. A cardboard overlay ensured that each pair of radiographs was assessed independently.

The radiographs were read twice by all examiners in separate recording sessions subsequently described as "Session I" and "Session II". Three to four weeks elapsed between recording sessions to eliminate memory regarding diagnostic decisions. The scores recorded in 'Session I' were never accessible to the examiner in 'Session II'.

The longitudinal method of reading radiographs was used, i.e. the third and fourth sets of radiographs of each subject were assessed in chronological order in quick succession (*Haugejorden*, 1974).

The scoring codes used for radiographic diagnosis were those described by *Haugejorden & Slack* (1975).

The score codes were recorded on scoring sheets by a trained scribe and simultaneously by a tape recorder to allow checks for recording errors. The tape recordings of scores revealed no disagreements of importance to the primary caries scores when checked against the completed scoring sheets.

Seven examiners participated in the study; (a) three dentists engaged in post-graduate study (examiners: D.1, D.2, and D.3); (b) two fourth year dental students

(examiners: S.1, and S.2); and (c) two dental hygienists working at a dental school (examiners: H.1, and H.2). All examiners received written diagnostic criteria and before work began, the recording procedure was carefully explained and demonstrated. In addition, the dental hygienists who had no previous education or experience in radiographic caries diagnosis, were given three hours of intensive training. Disregarding the calibrating effect which common training by examiner D.1 gave the dental hygienists, the results of the study will be those recorded by uncalibrated examiners working under identical conditions and employing the same diagnostic criteria.

#### *Diagnostic Standards*

Three diagnostic standards were established.

1. *Standard Ranking of Subjects.* It was decided to use the sum of ranks resulting from the duplicate assessments of the radiographs by the three dentists as the basis for the standard ranking of the subjects. This approach was permissible as there was a significant measure of agreement between the three dentists ( $p < 0.001$ ) (*Moroney*, 1956).

2. *Diagnostic Standard for the Group.* Three recording levels were defined by employing the scores recorded by the three dentists in 'Session I'. The three recording levels have been labeled »minimum score«, »intermediate score«, and »maximum score«, and were defined as follows.

- (a) The »minimum score« included only those approximal surfaces which were unanimously judged to be carious.
- (b) The »intermediate score« consisted of the »minimum score« plus those surfaces which were recorded as carious by two dentists, but not by the

Table I. Description of the caries diagnostic standard and the number of surfaces for which the three dentists recorded caries in each category

Caries diagnostic certainty scale	Two independent scores each by three dentists				
	Combinations of score codes <sup>1)</sup>	Number of surfaces			
		Third examination	Fourth examination		
»Positive»	c c c c c c	53	69	62	85
	c c c c c x	16		23	
»Probable»	c c c c x x	17	38	26	38
	c c c x x x	21		12	
»Questionable»	c c x x x x	17	60	23	66
	c x x x x x	43		43	
»Unlikely»	x x x x x x	—	—	—	—

<sup>1)</sup> Key: c = carious surface; x = other score code.

third, i.e. the majority decision for these surfaces was 'carious'.

- (c) The »maximum score» included all surfaces counted as 'carious' under (b) above, plus those surfaces which one dentist had judged to be carious, i.e. it includes a number of surfaces for which the minority decision indicated 'carious'.

3. *Diagnostic Standard for Individual Tooth Surfaces.* It was decided to design a diagnostic certainty scale based on the two independent scores recorded by each dentist during 'Session I' and 'Session II'. The use of six diagnostic decisions to establish a standard meant six categories describing levels of examiner agreement; from a unanimous decision that a surface was 'carious' to a situation in which only one dentist recorded 'carious' on one occasion. For convenience of tabulation and because of a relatively small number of surfaces judged to be carious in this material, the categories were grouped as shown in Table I.

The category labeled »unlikely» was included in Table I to take into account surfaces for which other examiners re-

corded 'carious' whereas the standard score was 'non-carious'.

*Statistical analysis.* Student's t test for paired observations and Chi-square analysis were used to determine the statistical significance of observed differences.

Spearman's rank correlation coefficient is used to present the results of the ranking of subjects according to caries score and Student's t test to determine the significance of the level of concordance of results recorded by dentists, dental students and dental hygienists (Moroney, 1956).

## RESULTS

The results of the comparisons between the diagnostic standards and the results obtained by dental students and dental hygienists, will be presented under three headings.

### 1. Comparisons of Ranking of Subjects

Spearman's rank correlation coefficient ( $R_s$ ) was calculated for each of the four non-dentally qualified examiners. The results for ranking according to primary caries scores are given in Table II.

Table II. Spearman's rank correlation coefficient ( $R_s$ ) for non-dentally qualified examiners when their ranking of subjects according to primary caries scores on third and fourth examination radiographs is compared with a standard ranking\*)

Session	$R_s$ for examiners			
	S.1	S.2	H.1	H.2
	Third examination			
I	0.85	0.81	0.86	0.92
II	0.83	0.89	0.93	0.86
	Fourth examination			
I	0.88	0.86	0.93	0.93
II	0.90	0.91	0.94	0.93

\*) The standard ranking is based on the ranking of subjects by three dentists in two independent recording sessions.

$R_s$  for the two dental students varied between +0.81 and +0.91. The coefficients for fourth examination results were consistently slightly higher than those for the third examination radiographs in each session. The range for the hygienists was from +0.86 to +0.94.

When Student's  $t$  was calculated for the coefficients shown in Table II, it gave  $p < 0.001$ , 38 degrees of freedom, i.e. the results of the non-dentally qualified examiners correlated highly significantly with the dentists' standard ranking.

## 2. Comparisons of Group Results

The mean primary caries score per subject for third examination radiographs are given in Table III. When 95 per cent confidence limits were calculated, it became apparent that the mean »minimum score» per subject corresponded approximately to the lower limit of the 95 per cent confidence interval about the mean »intermediate score», while the upper confidence limit nearly coincided

Table III. Mean standard primary caries score per subject and 95 per cent confidence limits based on the scores recorded by three dentally qualified examiners during 'Session II'

Standard score	Primary caries		95 per cent	Confidence limits*)
	$\bar{x}$	S.E.M.	Lower	Upper
	Third examination			
'Minimum'	1.40	0.34	0.71	2.09
'Inter-mediate'	2.15	0.43	1.28	3.02
'Maximum'	3.10	0.53	2.03	4.17
	Fourth examination			
'Minimum'	1.75	0.39	0.96	2.54
'Inter-mediate'	2.80	0.51	1.77	3.83
'Maximum'	3.90	0.58	2.73	5.07

\*)  $t_{(0.05,39)} = 2.0227$

with the mean »maximum score». In view of these findings it was decided to compare the primary caries score recorded by the dental students and hygienists with the »intermediate score».

The mean difference ( $\bar{d}$ ) between the caries scores of the four non-dentally qualified examiners and the standard »intermediate scores», its standard error, and Student's  $t$  for paired observations were calculated. The results are presented in Table IV.

The mean primary caries scores per subject recorded by the dental students were higher than the mean »intermediate score» for third and fourth examination radiographs in 'Session I' ( $p$  always  $< 0.01$ ). In fact three of their four mean scores exceeded the mean »maximum score». In 'Session II' only examiner S.1's mean score per subject was significantly different from the mean »intermediate score» ( $p < 0.05$ ).

The dental hygienists' results did not differ significantly from the mean »inter-

Table IV. Mean primary caries score difference ( $\bar{d}$ ) per subject between the score recorded by four non-dentally qualified examiners in 'Session I' and 'Session II' and the 'intermediate' standard score recorded by three dentists in 'Session II'

Examiner	$\bar{d}$	S.E.M.	Student's t	p
Session I				
Third examination				
S.1	1.40	0.31	4.52	<0.001
S.2	1.63	0.32	5.09	<0.001
H.1	-0.10	0.22	-0.45	>0.60
H.2	0.28	0.19	1.47	>0.10
Fourth examination				
S.1	0.90	0.29	3.10	<0.01
S.2	1.28	0.26	4.92	<0.001
H.1	-0.68	0.19	-3.58	<0.001
H.2	0.38	0.17	2.23	<0.05
Session II				
Third examination				
S.1	0.38	0.18	2.11	<0.05
S.2	0.38	0.22	1.50	>0.10
H.1	0.05	0.15	0.33	>0.70
H.2	-0.30	0.17	-1.76	>0.05
Fourth examination				
S.1	0.35	0.23	1.52	>0.10
S.2	0.03	0.22	0.14	>0.80
H.1	-0.10	0.17	-0.59	>0.50
H.2	-0.30	0.21	-1.43	>0.10

Table V. Distribution of primary caries scores recorded by non-dentally qualified examiners according to score categories of the standard

Ex- ami- ner	Total score (100%)	'Caries diagnostic certainty scale'			
		'Posi- tive' %	'Prob- able' %	'Question- able' %	'Un- likely' %
Third examination — session I					
S.1	142	45.8	19.7	15.5	19.0
S.2	151	45.0	19.9	13.9	21.2
H.1	82	73.2	14.6	8.5	3.7
H.2	97	63.9	21.6	12.4	2.1
(3 dentists × 2*)		69	38	60	—
Fourth examination — session I					
S.1	148	53.4	17.6	14.9	14.2
S.2	163	50.9	12.9	14.7	21.5
H.1	85	78.8	15.3	5.9	—
H.2	127	63.0	17.3	13.4	6.3
(3 dentists × 2*)		85	38	66	—
Third examination — session II					
S.1	101	64.4	18.8	9.9	6.9
S.2	101	60.4	24.8	8.9	5.9
H.1	88	76.1	18.2	4.5	1.1
H.2	74	75.7	14.9	6.8	2.7
Fourth examination — session II					
S.1	126	62.7	15.1	13.5	8.7
S.2	113	69.0	15.9	10.6	4.4
H.1	108	74.1	17.6	7.4	0.9
H.2	100	73.0	18.0	8.0	1.0

\*) Number of surfaces according to category from Table II.

mediate score» for third examination radiographs ( $p > 0.05$ ). For fourth examination radiographs the differences were statistically significant for both hygienists in 'Session I' ( $p < 0.05$ ) but not in 'Session II' (Table IV).

Whereas the dental students' mean primary caries scores were always higher than the mean »intermediate score» per subject, examiner H.1 had only one mean score higher, and examiner H.2, two mean scores higher than the mean »intermediate score».

### 3. Comparisons of Recording by Tooth Surface

The distribution of the primary caries scores recorded by the four non-dentally qualified examiners according to standard score category are given in Table V.

For third examination radiographs assessed in 'Session I' the percentage of the total score recorded by the dental students in each category of the standard, never differed by more than 2 per cent, and never by more than 8 per cent for fourth examination radiographs. Since the dental

students returned higher scores than everybody else in 'Session I', it was not unexpected to find a relatively high proportion of their total score in the »questionable» and »unlikely» categories of the standard, but 14—22 per cent of the score in the »unlikely» category appears high in view of the fact that about 30 per cent of the surfaces of the three first categories of the standard had not been recorded.

The hygienists had recorded only 0—6 per cent of their primary caries score in the »unlikely», and 80—93 per cent in the »positive» and »probable» categories of the standard in 'Session I'.

The differences between results obtained by students and hygienists in the distribution of their total scores among the categories of the standard, were highly significant for third ( $X^2 = 35.37$ ,  $p < 0.001$ , 9 degrees of freedom) and fourth examination results ( $X^2 = 35.83$ ,  $p < 0.001$ , 9 degrees of freedom).

There was a marked reduction of primary caries scores for the students from the first to the second recording session, but they still scored higher than the hygienists. This change reduced the number of surfaces in the »questionable» and »unlikely» categories of the standard, from 49 to 17 surfaces for examiner S.1 and from 53 to 15 surfaces for S.2. The differences in distribution of scores among the categories of the standard were no longer statistically significant at the 5 per cent level (Table V).

Another way of looking at these results is to compare the scores in each category with the corresponding number of surfaces recorded by the dentists.

Table V shows that from 1—13 surfaces of the »positive» category were missed by the four non-dentally qualified examiners in the two recording sessions when

considering third examination radiographs. Dental students missed fewer of the surfaces included in this standard category than did hygienists. The only exception was due to examiner H.1 who came closest to the standard score in 'Session II'. The corresponding range for the fourth examination was from 2—18 surfaces. The greatest deviation from the standard count was returned by examiner H.1 in 'Session I'.

Of the 38 surfaces which the three dentists had recorded in the »probably» carious category, from 8—27 per cent had not been recorded by the non-dentally qualified examiners when assessing third examination radiographs. Similarly the students and hygienists did not record 9 to 25 surfaces of the »probable» category for fourth examination radiographs.

Of the 60 surfaces in the »questionable» category only 7—31 per cent were also recorded for third, while 8—36 per cent were diagnosed by the non-dentally qualified examiners from fourth examination radiographs.

#### DISCUSSION

The diagnostic standard with which examiners' recorded diagnoses might be compared would ideally be the true caries score for each tooth surface of the subjects. A true caries score cannot be established *in vivo* (Haugejorden, 1974), nor can a true radiographic caries status, as no infallible diagnostic method exists. It was for these reasons that diagnostic standards were constructed on the basis of routinely recorded scores.

The reasoning behind the use of standard score categories was that surfaces contributing to the »minimum»/»positive» score categories almost certainly were carious. The surfaces which made up the

»intermediate»/»probable» score categories were probably carious, but the uncertainty is greater than for the surfaces of the first mentioned categories. The difference between »maximum» and »intermediate» scores, though it probably includes some surfaces exhibiting true radiographic caries, will predominantly consist of surfaces with borderline lesions, and surfaces for which a false positive diagnosis was recorded. The »questionable» category includes surfaces scored falsely positive for caries and some borderline lesions.

The use of scores from two recording sessions for the surface standard reduced the probability of false negative diagnoses, but this approach would tend to increase the number of false positive diagnoses in the »probable» and »questionable» categories of the standard.

There will always be an element of uncertainty inherent in the use of this type of standard since it depends on the recording level of the examiner's whose scores were used to construct it and the diagnostic criteria employed. Thus the comparison of an examiner's scores with this type of standard score does not provide a measure of accuracy.

Though there is highly significant agreement between the standard ranking of subjects and that recorded by the four non-dentally qualified examiners, some non-systematic deviation of ranking has occurred. *Leijon & Markén* (1968) suggested  $R_s = + 0.70$  as an acceptable ranking. The coefficients reported in Table II are all well above this critical level. The results presented here, which were recorded by inexperienced examiners, suggest that  $R_s = + 0.85$  to  $+ 0.90$  might be attained with comparatively little training and calibration of examiners.

It is also encouraging to find that these

examiners achieved the reported level of agreement with the »intermediate» score per subject without preliminary training and only through using the same written diagnostic criteria as the dentists (Table IV). On this basis it seems reasonable to suggest that agreement with a standard mean score might be relatively easy to achieve for radiographic caries. On the other hand, it may indicate that this standard is insufficiently sensitive to reveal the magnitude of the problem of caries diagnostic variability of examiners.

The disagreement between scores of the non-dentally qualified examiners and the standard scores became more obvious when analysis focused on individual tooth surfaces (Table V). It seems therefore that a standard based on scores for each tooth surface will give more scope for monitoring changes of caries diagnostic levels.

The radiographs constituting the standard set, should ideally exhibit the widest possible range of stages of development of radiographic caries. The set should also provide for a reasonable number of diagnostic decisions for a relatively limited number of sets of bitewing radiographs. Furthermore, diagnostic criteria must be clearly formulated to facilitate uniform interpretation and application.

Without being able to identify them, examiners should reassess the radiographs of the standard set at intervals during the examination period. When using the standard set in practice it is not only necessary to establish a standard score, as shown here, but also to define control limits on either side of it to indicate the deviations which may be tolerated before corrective action is taken.

As investigations to determine the relative effectiveness and efficiency of preventive agents become more numerous

the need for a diagnostic standard to facilitate more reliable comparison between results is likely to increase.

*Acknowledgements.* For the use of radiographs, equipment and facilities the author would like to thank Professor G. L. Slack and the London Hospital Medical College Dental School.

#### REFERENCES

- Arnold, F. A. Jr., Likins, R. C., Russel, A. L. & Scott, D. B.* 1962. Fifteenth year of the Grand Rapids fluoridation study. *J. Amer. Dent. Ass.* 65, 780—785
- Backer Dirks, O., Baume, L. J., Davies, G. N. & Slack, G. L.* 1967. Principal requirements for controlled clinical trials. *Int. Dent. J.* 17, 93—103
- Backer Dirks, O., Houwink, B. & Kwant, G. W.* 1961. The results of 6½ years of an artificial fluoridation of drinking water in the Netherlands. The Tiel-Culemborg experiment. *Archs Oral Biol.* 5, 284—300
- Backer Dirks, O., van Amerongen, A. J. & Winkler, K. C.* 1951. A reproducible method for caries evaluation. *J. Dent. Res.* 30, 346—359
- Haugejorden, O.* 1974. *A study of the methods of radiographic diagnosis of dental caries in epidemiological investigations.* *Acta Odont. Scand.* 32, Suppl. 65, 269 p
- Haugejorden, O. & Slack, G. L.* 1975. A study of intra-examiner error associated with recording of radiographic caries at different diagnostic levels. *Acta Odont. Scand.* 33, 169—181
- Horowitz, H. S., Baume, L. J., Backer Dirks, O., Davies, G. N. & Slack, G. L.* 1973. Principal requirements for controlled clinical trials of caries preventive agents and procedures. *Int. Dent. J.* 23, 506—516
- Leijon, G. & Markén, K.-E.* 1968. Roentgenologic diagnosis of approximal caries. Deviations between observers and comparisons between the recordings from periapical and bite-wing roentgenograms. *Acta Odont. Scand.* 26, 35—61
- Moroney, M. J.* 1956. *Facts from Figures.* 3rd ed., Penguin Books Ltd., Harmondsworth, Middlesex, England. pp. 334—341
- Pfahler, G. E.* 1908. The use of the roentgen rays in dentistry. *Dental Cosmos* 50, 916—923
- Slack, G. L., Berman, D. S., Martin, W. J. & Young, J.* 1967. Clinical testing of a stannous fluoride-insoluble metaphosphate dentifrice in Kent school girls. *Brit. Dent. J.* 123, 9—16
- Westin, G. & Wold, H.* 1943. 1942 års tandmönstring av inskrivningskyldiga. *Odont. Tidskr.* 51, 487—615