The construction and use of diagnostic standards for radiographic caries incidence scores

OLA HAUGEJORDEN

Institute of Community Dentistry, University of Oslo, Norway

Haugejorden, O. The construction and use of diagnostic standards for radiographic caries incidence scores. *Acta Odont. Scand.* 35, 95–102, 1977

The aim was to construct and illustrate the use of diagnostic standards for recording of DFS incidence. Separate standards are presented for ranking of subjects; the mean «intermediate» DFS score per subject, which represented the majority decision of three dentists; and a certainty scale for individual tooth surfaces. The scores recorded for 40 sets of posterior bitewing radiographs of children between 13 and 15 years of age by two fourth year dental students and two dental hygienists were compared with the standards. Spearman's rank correlation coefficient reached the critical level of + 0.70 in three of eight instances. Only one mean net DFS increment score differed significantly from the mean standard score ($p < 0.05$). The level of agreement with the certainty scale for individual surfaces was low. Focusing attention on recording of DFS incidence gives a more severe test of examiners' diagnostic performance. The construction and use of control charts for monitoring the diagnostic level of examiners is suggested.

Key-words: Dental caries; roentgenology; epidemiology

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

It has been suggested that caries diagnostic variability within and between examiners may exceed the caries inhibitory effects of agents being studied and that the choice of examiner may be as important as the choice of the test substance (*Shaw & Murray*, 1975).

The need for a critical acceptance level for diagnostic performance of examiners when diagnosing caries is now recognized (*Lu*, 1975; *Haugejorden & Slack*, 1975). A diagnostic inconsistency rate of maximum 20 per cent was proposed for recording of the earliest possible radiographic lesion (*Haugejorden & Slack*, 1975) while *Kamen & Schreer* (1975), without referring to the diagnostic method, mentioned at least 30 per cent for the initial stages of caries.

In addition to the problem of reproducibility of diagnosis, it is important to maintain a constant diagnostic level throughout the duration of clinical trials. *Haugejorden* (1976) has described the construction and use of three types of standards for primary radiographic caries to guard against shifts of diagnostic level.

As the reliability coefficient (*Guilford & Fruchter*, 1973) was consistently lower for DFS increments, than for primary radiographic caries scores (*Haugejorden & Slack*, 1975), it may be useful to study

Received for publication, March 18, 1976

examiners' diagnostic performance relative to a standard when recording DFS incidence.

The purpose of this paper is to describe the construction of diagnostic standards for recording of radiographic DFS incidence during one year of a clinical trial. Separate standards will be presented for (1) the ranking of subjects according to score; (2) the mean score per subject; (3) the change or lack of change of status of individual tooth surfaces; and (4) data will be presented to illustrate the use of the standards.

MATERIAL AND METHOD

The material and methods used to obtain data for constructing the standards and for illustrating their use have been reported earlier by *Haugejorden & Slack* (1975) and *Haugejorden* (1976). Only a brief description is therefore given here.

The forty English school children who had participated in clinical trials of fluoride containing dentifrices were between 13 and 15 years of age when the pairs of posterior bitewing radiographs were taken at the third and fourth annual examinations (*Slack et al.*, 1967a, b, 1971).

The radiographs had been produced while using a standardized examination technique and exposure time and processing of the Kodak DF55 films had been in accordance with the manufacturer's recommendations.

The pairs of bitewing radiographs of each subject were mounted in coded plastic film mounts. They were assessed independently twice, by three dentists, two fourth year dental students and two dental hygienists in recording sessions 3–4 weeks apart. A Watson's viewing box with a fixed magnifying lens (x2) was used, and working conditions and procedures were identical and kept constant throughout for all examiners. The scoring codes for radiographic diagnosis were those reported by *Haugejorden & Slack* (1975). The scores were recorded simultaneously by a trained scribe and a tape recorder to permit checks for recording errors.

The terms caries increment and decrement are used as defined by *Haugejorden* (1974).

Diagnostic Standards

Since neither the true caries score nor true radiographic caries score could be established for the 40 subjects (*Haugejorden*, 1974), an alternative, practical solution for epidemiological studies had to be found for DFS increments and decrements as for primary caries scores.

Standard Ranking of Subjects

The standard ranks were determined by the sum of the six ranks available for each subject from the two assessments of the radiographs by the three dentists. This approach was acceptable as there was significant agreement in the ranking among the three dentists ($p < 0.05$) (*Moroney*, 1956).

Diagnostic Standard for the Group

The three recording levels for DFS increments are defined as follows:

(a) The «minimum DFS increment/decrement score» includes only those surfaces about which all three dentists agreed on the score transition between the third and fourth examination.

(b) The «intermediate DFS increment/decrement score» consists of the surfaces which contributed to the «minimum» score plus those surfaces for which two of the dentists had recorded an increment or a decrement, i.e. the majority decision indicated a score transition which contributed to the DFS increment or decrement.

(c) The «maximum DFS increment/decrement score» includes the surfaces mentioned under

(b) as well as surfaces which only one dentist had judged to contribute towards the DFS increment/decrement, thus accepting the minority decision.

In a limited number of instances a decision could not be reached on the basis of available scores from the second session as one dentist had recorded an increment, one a decrement, and the third no change of score which affected the DFS index. Two of the participating dentists re-examined these surfaces and their joint decision was accepted in order that a scale of scores might be established.

Diagnostic Standard for Tooth Surfaces

Degree of certainty scales for the recording of DFS increments and decrements were established. Table I shows that the same basic principles were followed as those outlined for primary radiographic caries (*Haugejorden,*

Table I. *Description of DFS increment and decrement diagnostic standards based on two independent scores by each of the three dentists; and the number of surfaces by category*

Level of agreement	Number of surfaces	Categories on diagnostic certainty scale
	DFS Increments	
++++++ +++++0	3, 12 → 15	"Positive"
++++00 +++000	13, 16 → 29	"Probable"
++0000 +00000	22, 46 → 68	"Questionable"
+0000– ++000– +000–– +00–––	3, 3, 2, 1 → 9	"Undecided"
	DFS Decrements	
–––––– –––––0	3, 1 → 4	"Positive"
––––00 –––000	4, 8 → 12	"Probable"
––0000 –00000	12, 33 → 45	"Questionable"
000000	– –	"Unlikely"

Key to signs for level of agreement:

\+ Change of status contributing to DFS increment.
– Change of status contributing to DFS decrement.
0 No change of status or a change contributing to neither DFS increment nor decrement.

1976). For example, the «positive» category of the standard contains those surfaces for which either all six diagnostic decisions, or at least five, agreed that a non-carious surface at the third, had become decayed or filled at the fourth examination when dealing with DFS increments; and the opposite score transition between these examinations for DFS decrements.

An «undecided» category has been included in Table I because a small number of surfaces had been recorded both as an increment and a decrement by the three dentists.

Comparisons of Recording by Tooth Surface

Examiner S.1. recorded a DFS increment of 52 and a DFS decrement of 40 in «Session I», i.e. a net DFS increment of 12. Thirty-one per cent of surfaces contributing to the DFS increment had been recorded in the «positive» and «probable»; and 29 per cent in the «questionable» and «undecided» categories of the standard, while 40 per cent was due to surfaces which in the opinion of the dentists did not contribute to the DFS increment. Of the 40 DFS decrement surfaces, 10 were observed in the first two categories, 7 in the «questionable», and the remaining 23 on surfaces for which the dentists had not recorded a decrement (Table V).

The same examiner's net DFS increment increased to 31 in «Session II». The increase was largely due to a reduction of the DFS decrement (Table V). The proportion of the DFS increment due to surfaces not recorded by the dentists decreased from 40 per cent in «Session I» to 23 per cent in «Session II», and from 23 (57 per cent) to 5 (23 per cent) for DFS decrements.

If the number of surfaces recorded by examiner S.I. in the «positive» and «probable», «questionable» plus «undecided», and «unlikely» categories in «Session I» and «Session II» are tabulated in a 3x2 contingency table and Chi-square calculated, then it is revealed that the distribution of DFS increments among categories of the standard did not change significantly between sessions ($X^2 = 2.86$, $p > 0.10$, 2df). However, the distribution of DFS decrements between categories did change significantly, ($X^2 = 6.81$, $p < 0.05$, 2 df).

Examiner S.2's results were similar to those described for S.1., but the distribution of the DFS increment among the categories of the standard changed significantly between recording sessions ($X^2 = 8.47$, $p < 0.025$, 2df). The changes in distribution of DFS decrements between sessions were not statistically significant ($X^2 = 4.38$, $p > 0.10$, 2df).

The results of the hygienists showed similar changes in distribution of DFS increment and decrement scores among the categories of the standard as observed between sessions for the dental students. However the intra-examiner differences between sessions were not statistically significant at the 5 per cent level. Furthermore, the hygienists had, in most instances, recorded more of their DFS increment and decrement in the «positive» and «probable» categories of the standard than did dental students (Table V).

The number of surfaces recorded by the dentists in each category of the standard which were also detected by the non-dentally qualified examiners may be deduced from Tables I and V. Thus it was observed that from 8–12 surfaces of the 15 DFS increments of the «positive» category were also recorded by these examiners. Of the 29 DFS increments in the «probable» category, the number of surfaces recorded varied between 7 and 14, while of the 68 DFS surfaces in the «questionable» category, from 10–15 surfaces had also been included by the students and hygienists.

For DFS decrements the 4 surfaces of the «positive» category had all been recorded except by examiner S.1, who included only 2, and H.2. who recorded 3 of these surfaces in «Session I». From 0–8 of the 12 surfaces in the «probable» category were also recorded by

Table II. *Spearman's rank correlation coefficient (R_S) for non-dentally qualified examiners when their ranking of subjects according to DFS increments and decrements is compared with a standard ranking*[a]

Session	R_S for examiners			
	S.1	S.2	H.1	H.2
I	0.62	0.54	0.74	0.78
II	0.68	0.60	0.85	0.57

a) Standard ranking is the resultant ranks from the DFS increments and decrements recorded by the three dentists in 'Session I' and 'Session II'.

Table III. *Mean standard DFS increments per subject and 95 per cent confidence limits based on the scoring by three dentally qualified examiners during 'Session II'*

Standard score	DFS increments		95 percent confidence limits[a]	
	Mean	SE	Lower	Upper
"Minimum"	0.28	0.19	− 0.10	0.66
"Intermediate"	0.85	0.36	0.12	1.58
"Maximum"	0.98	0.49	− 0.01	1.97

a) $t_{(0.05,39)} = 2.0227$

these four examiners, while from 6–11 of the 45 «questionable» DFS decrements had been included in their scores.

RESULTS

The results will be presented under three headings.

Table IV. *Mean DFS increment difference ($\bar{d}$) per subject between the score by four non-dentally qualified examiners in 'Session I' and 'Session II', and the "intermediate" score by three dentists in 'Session II'*

Examiner	$\bar{d}$	SE($\bar{d}$)	Student's t	p
Session I				
S.1	− 0.55	0.38	− 1.45	> 0.10
S.2	− 0.53	0.38	− 1.39	> 0.10
H.1	− 0.60	0.29	− 2.07	< 0.05
H.2	− 0.08	0.21	− 0.38	> 0.60
Session II				
S.1	− 0.08	0.27	− 0.30	> 0.70
S.2	− 0.48	0.25	− 1.85	> 0.05
H.1	− 0.13	0.21	− 0.62	> 0.50
H.2	− 0.08	0.23	− 0.35	> 0.70

Comparisons of Ranking of Subjects

Spearman's rank correlation coefficient (R_S) for the dental students varied between + 0.54 and + 0.68, while the corresponding range for the dental hygienists was + 0.57 to + 0.85 (Table II). These levels of agreement between each non-dentally qualified examiner and the standard ranking was statistically significant in every instance ($p < 0.001$).

Comparisons of Group Results

The mean standard net DFS increments per subject are given in Table III. For primary caries it was noted that the mean «intermediate» standard score was almost exactly half-way between the mean «minimum» and «maximum» standard scores per subject (*Haugejorden*, 1976), but this did not apply to the net DFS increment estimates. The mean DFS increment and decrement scores recorded by the four non-dentally qualified examiners were still compared with the mean «intermediate» standard DFS scores (Table IV).

Table V. *Distribution by standard score category of DFS increments and decrements recorded in two recording sessions by dental students and hygienists*

Categories of diagnostic certainty scale	Recording session			
	I		II	
	DFS			
	Increment %	Decrement %	Increment %	Decrement %
Examiner – S.1				
"Positive"	15.4	5.0	21.4	16.0
"Probable"	15.4	20.0	21.4	28.0
"Questionable"	25.0	17.5	26.8	36.0
"Undecided"	3.8	–	7.1	–
"Unlikely"	40.4	57.5	23.2	20.0
Total (100 %)	52	40	56	25
Examiner – S.2				
"Positive"	18.3	8.5	23.9	12.9
"Probable"	15.0	8.5	19.6	22.6
"Questionable"	16.7	23.4	34.8	29.0
"Undecided"	1.7	4.3	2.2	6.5
"Unlikely"	48.3	55.3	19.6	29.0
Total (100 %)	60	47	46	31
Examiner – H.1				
"Positive"	22.9	16.0	23.4	22.2
"Probable"	20.0	16.0	29.8	33.3
"Questionable"	31.4	32.0	31.9	33.3
"Undecided"	8.6	4.0	4.3	5.6
"Unlikely"	17.1	32.0	10.6	5.6
Total (100 %)	35	25	47	18
Examiner – H.2				
"Positive"	21.1	11.5	22.2	28.6
"Probable"	24.6	19.2	33.3	–
"Questionable"	26.3	23.1	24.4	42.9
"Undecided"	–	3.8	–	14.3
"Unlikely"	28.1	42.3	20.0	14.3
Total (100 %)	57	26	45	14

It will be seen that the four examiners' mean net DFS increment per subject without exception was lower than the mean net «intermediate» DFS score. In fact the mean score recorded by examiner H.I. in «Session I» was lower than the «minimum» net DFS increment score per subject; a result which differed significantly from the mean «intermediate» standard score ($p < 0.05$).

DISCUSSION

Whereas Spearman's rank correlation coefficient when comparing primary caries scores of the four examiners with a set of standard scores, were all above the critical level of + 0.70 quoted by *Leijon & Markén* (1968), (*Haugejorden*, 1976), only three of eight estimates in Table II were higher than + 0.70. Though there was a statistically significant level of agreement between the standard ranking of subjects and the ranking carried out by the four examiners ($p< 0.001$), and only one mean difference was significant ($p < 0.05$) (Table IV), the results presented in Table V revealed marked disagreement when considering the individual surfaces which contributed an increment or a decrement to the net DFS score.

The less favourable results of the comparisons with the standard score when focusing attention on recording of increments/decrements instead of primary caries is probably due to changes in the mouth during the interval of time between examinations; to changes in radiographic projection or factors influencing film density (*Haugejorden*, 1974); and to diagnostic errors at two examinations, instead of one, affecting incidence scores.

The standard set of radiographs for controlling the diagnostic level for recording of DFS incidence should ideally exhibit the whole range of incidence and extent of disease which may be encountered in a study of population and provide a reasonably large number of diagnostic decisions for a limited number of sets. More evidence is needed before definite recommendations can be made on the number of sets of radiographs and diagnostic decisions necessary for judging stability or otherwise at a chosen diagnostic level.

Furthermore, a decision must be taken concerning the most suitable interval of time between first and second examination. Although two- and three-year increments are often reported, it seems that sets of radiographs taken one year apart might be most useful for this purpose. This interval of time between examinations is most frequently used to determine whether or not the caries preventive effect of an agent changes with increasing period of use, and also because the growth of lesions during one year may be limited (*Hollender & Koch*, 1969; *Berman & Slack*, 1973) and therefore provide a more severe test of diagnostic performance of examiners.

The results reported here were arrived at using diagnostic criteria stipulating recording of the earliest detectable radiographic lesions without specific requirements as to depth of penetration of the radiolucency. The only limitation imposed was that «sound» should be recorded if there were doubts as to the presence of caries. The reliability coefficients reported for recording of DFS increments when employing different requirements as to depth of penetration of radiographic carious lesions were not consistently higher when counting only advanced lesions (*Haugejorden & Slack*, 1975). For this reason, and because recording only relatively advanced lesions might entail a study of caries progression rather than caries incidence, it seems advantageous to record the earliest detectable radiographic lesion, provided that acceptable reproducibility is attained.

On the basis of available evidence it seems reasonable to suggest that 3–5 trained examiners should assess the chosen sets of radiographs (*Markén*, 1962) using thoroughly tested diagnostic criteria when establishing

standard scores of the type described here. Since underscoring probably is more common than over-recording (*Quenzel, Gustafsson & Grahnén*, 1954), and to simplify the certainty scale (Table I), the surfaces classified in the «probable», «questionable» and «unlikely» categories after the first assessment, should be re-examined simultaneously by 2–3 examiners of equal professional standing (*Markén*, 1970) to arrive at a definite diagnosis.

The establishment of this type of diagnostic standards to monitor the maintenance of a stable diagnostic level may permit the construction of control charts with warning- and action limits and setting up of rules for what corrective action is necessary and when it should be implemented (*Page*, 1955; *Moroney*, 1956). Further research is required to determine practical warning- and action limits and suitable rules for intervention for various degrees of departure from an agreed diagnostic level.

A solution of these problems would facilitate control of intra-examiner variability in recording level over time, as well as, provide a method for controlling inter-examiner differences. Adoption of the methods suggested might help to reduce the uncertainties associated with comparison of results of clinical trials conducted at different times by one or by a number of different investigators.

REFERENCES

Berman, D.S. & Slack, G.L. 197 3. Caries progression and activity in approximal tooth surfaces. A longitudinal study. Brit. Dent. J. 1 3 4, 51–57

Guilford, J.P. & Fruchter, B. 197 3. *Fundamental statistics in psychology and education.* 5th ed. New York, McGraw-Hill, p 399

Haugejorden, O. 197 4. *A study of the methods of radiographic diagnosis of dental caries in epidemiological investigations.* Acta Odont. Scand. 32, Suppl. 65, pp 45, 50 and 76

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

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

Hollender, L. & Koch, G. 1969. Roentgenographic partial recording of caries in assessment of effect of prophylactic measures. Odont. Revy, 20, 267–282

Kamen, A. & Schreer, S.B. 1975. Consistency is validity – or are both possible in clinical studies of anticariogenic agents? J. Dent. Res. 5 4, 1098

Leijon, G. & Markén, K.-E. 1968. Roentgenologic diagnosis of proximal caries. Deviations between observers and comparisons between the recordings from periapical and bite-wing roentgenograms. Acta Odont. Scand. 26, 35–61

LU, K. H. 1975. Quality control of diagnostic performance in large caries studies. IADR Abstr. L 143. J.Dent.Res. 54, L 36

Markén, K.-E. 1962. *Studies of deviations between observers in clinico-odontological recording.* Trans. Royal School of Dentistry, Stockholm, Umeå, No. 8, p 186

Markén, K.-E. 1970. Observatøren i kliniskt-odontologiska undersökningar. Svensk Tandläk. Tidskr. 6 3, 787–801

Moroney, M.J. 1956. *Facts from Figures.* 3rd ed. Penguin Books Ltd., Harmondsworth, Middlesex, England, pp 1 41–172 and 3 3 4– 3 41

Page, E.S. 1955. Control charts with warning lines. Biometrika 42, 2 4 3–257

Quenzel, C.-E., Gustafsson, B.E. & Grahnén, H. 195 4. The Vipeholm dental caries study. Reliability of the method in determination of caries activity. Acta Odont. Scand. 11, 365– 388

Shaw, L. & Murray, J.J. 1975. Inter-examiner and intra-examiner reproducibility in clinical and radiographic diagnosis. Int. Dent. J. 25, 280–288

Slack, G.L., Berman, D.S., Martin, W.J. & Young, J. 1967a. Clinical testing of a stannous fluoride-insoluble metaphosphate dentifrice in Kent school girls. Brit. Dent. J. 12 3, 9–16

Slack, G.L., Berman, D.S., Martin, W.J. & Hardie, J.M. 1967b. Clinical testing of a stannous fluoride-calcium pyrophosphate dentifrice in Essex school girls. Brit. Dent. J. 12 3, 26– 3 3

Slack, G.L., Bulman, J.S. & Osborn, J.F. 1971. Clinical testing of fluoride and non-fluoride containing dentifrices in Hounslow school children. Brit. Dent. J. 1 30, 15 4–158