Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study

M. B. Saunders, K. Gulabivala, R. Holt & R. S. Kahan

Eastman Dental Institute and Hospital for Oral Health Care Sciences, University of London, London, UK

Abstract

Saunders MB, Gulabivala K, Holt R, Kahan RS. Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study. *International Endodontic Journal,* **33,** 272–278, 2000.

Aim The aim of this preliminary study was to test the reliability of radiographic evaluation of features of endodontic interest using a newly devised data collection system

Methodology Twelve endodontic MSc postgraduate students and one specialist endodontist examined sample radiographs derived from a random selection of 42 patients seen previously on an Endodontic New Patient Clinic (EDI). Each student examined a random selection of 8–9 roots on periapical radiographs of single- and multirooted teeth, with and without previous root canal therapy and 3–4 dental panoramic tomograms (DPTs). A total of 100 roots were examined. A proforma was used to record observations on 67 radiographic features using predefined criteria. Intra-observer agreement was tested by asking the students to re-examine the radiographs. The principle investigator and the specialist endodontist examined the same radiographs and devised a Gold Standard using the same criteria. This was compared with the student assessments to determine inter-observer variation. The postgraduates then attended a revision session on the use of the form. Each student subsequently examined 8–9 different roots from the pool of radiographs. A further assessment of inter-observer variation was made by comparing these observations with the Gold Standard.

Results Of the 67 radiographic features, only 25 had sufficient response to allow statistical analysis. Kappa values for intra- and inter-observer variation were estimated. These varied depending on the particular radiographic feature being assessed. Fifteen out of 25 intra-observer recordings showed 'good' or 'very good' Kappa agreement, but only three out of 25 inter-observer observations achieved 'good' or 'very good' values. Inter-observer variation was improved following the revision session with 16 out of 25 observations achieving 'good' or 'very good' Kappa agreement.

Conclusions Modification to the proforma, the criteria used, and training for radiographic assessment were considered necessary to improve the accuracy and reproducibility of the observations entered.

Keywords: observer variation, proforma, radiographic evaluation.

Received 7 December 1998; accepted 13 July 1999

Introduction

Effectiveness in the treatment of diseases is determined by using measurable parameters of disease presence and absence. Where the disease process is confined to, or arises within bony structures, such as periapical disease, a combination of clinical and radiographic criteria are used to help determine treatment outcome (Harty *et al.* 1970). Of additional interest is the influence of preoperative, intraoperative and post-operative factors on the treatment outcome (Smith *et al.* 1993). Most of our knowledge of factors

Correspondence: K. Gulabivala, Department of Conservative Dentistry, Eastman Dental Institute and Hospital for Oral Health Care Sciences, 256 Grays Inn Road, London WC1X 8LD, UK (fax: 0171 915 1028, e-mail: k.gulabivala@eastman.ucl.ac.uk).

influencing outcome of treatment of periapical disease is derived from retrospective studies (Grahnen & Hansson 1961, Bender *et al.* 1966, Kerekes & Tronstad 1979, Sjogren *et al.* 1990). Unfortunately the value of this data tends to be compromised by the nonstandardized and incomplete nature of routine clinical data recording (Chow 1993, Mokbel 1994). This problem could be addressed by prospective, systematic data collection systems. However, there does not appear to be any consensus on what constitutes the ideal. In principle, such a system should prompt the clinician to look for and record historical and clinical information in a systematic and perhaps standardized way.

Obviously, an important part of the endodontic assessment is the radiographic evaluation (pre-, intra-, and postoperatively) which is subject to considerable variation in interpretation (Goldstein *et al.* 1971). Although solutions have been suggested for improving reading of radiographs, there is little consensus on the different approaches (Grondahl 1979, Reit 1987). The reliability of interpretation of dental radiographs may be affected by the education, training and experience of the observers, the quality of radiographs, the viewing environment and examiner knowledge of the subject matter (Valachovic *et al.* 1986, Rohlin *et al.* 1991, Stheeman *et al.* 1995).

The validity of treatment outcome studies is dependent on the accuracy of the data collection, of which radiographic evaluation forms an important component. The purpose of this study was to test the hypothesis that information gathered using a newly devised radiographic data collection system (Proforma) would have poor reliability. The hypothesis was tested by using intra- and inter-observer variation amongst examiners as a measure of the reliability of the data collection and the defined criteria of observation. This was repeated after further training in the use of the form and criteria.

Materials and methods

Materials

Intra-oral periapical radiographs (PAs) and dental panoramic tomograms (DPTs) from 42 randomly selected patients examined on an Endodontic New Patient Clinic (Eastman Dental Institute) were used to test intra- and inter-observer variation.

A total of 100 roots from the PA radiographs, including single-rooted and multirooted teeth, with and without previous root canal therapy, and 42 DPTs were divided equally and given to 12 endodontic postgraduate students for assessment in the normal clinic environment using a standard fluorescent light box. Each postgraduate therefore examined eight or nine roots and three or four DPTs.

Entries were made into the radiographic proformae (Fig. 1) (Table 1) which consisted of 67 categorized radiographic features. The definitions of the radiographic categories were contained in a reference document previously given to the postgraduate students and discussed with them. The data from this assessment was consigned 'Postgraduate Observations 1'.

The same postgraduates were then asked to repeat the exercise under the same conditions 3–6 weeks later. This data was consigned 'Postgraduate Observations 2'.

All the sample radiographs were then assessed individually by the principle investigator (MS) and a specialist endodontist (RK), and the observations were compared. A consensus decision was reached by negotiation of disagreements and a 'Gold Standard' (as per Halse & Molven 1986) was established.

The postgraduate students then attended a training session on radiographic assessment and the category definitions were re-enforced with reference to early assessment of the Observation 2 Proforma. Areas of large scale disagreement were identified and discussed with the postgraduates. Immediately following this they were asked to assess a similar number of different

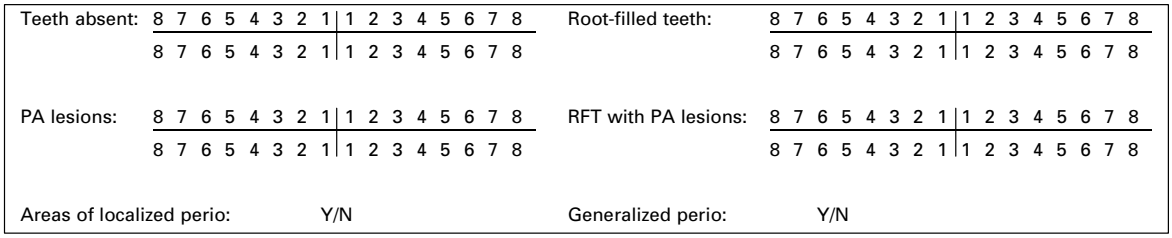

Figure 1 DPT assessment form

Table 1 Radiographic assessment form

Tooth no.						Code
(i) Crown						
Caries – Location						M/D/O/B/L/M-Rt/D-Rt
Caries – Depth						I/II/III
Restoration – Deepest						mms from pulp chamber
Restoration – Marginal Def						M/D/BL/☒
(ii) Pulp chamber						
2′ Dentine						Physiologic/Irritation
Sclerosis						☑☒
Pulp Stones						M/D/Central
Int Resorption – Symmetry						Sym/Asym
Int Resorption – Location						M/D/Central
Int Resorption – Extent						Max diam in mms
Furcation Perforation						☑☒
(iii) Root canal						
Root						M/D/B/L/P/MB/DB/etc
Curvature						<10°/10°-45°/>45°
S-shaped root						☑☒
GP in sinus – pointing						Cor/Mid/Apic/Furc
RCT – Material						GP/Ag-s/Ag-p/Pst/Am/Ind
RCT – mms from apex						Over = +/Under = −
RCT – Coronal Voids						☑☒
RCT – Middle Voids						☑☒
RCT – Apical Voids						☑☒
RCT – Coronal Adaption						Good/Fair/Poor
RCT – Middle Adaption						Good/Fair/Poor
RCT – Apical Adaption						Good/Fair/Poor
Retrograde Seal						Good/Fair/Poor
Post – Present						☑☒
Post – Decemented in past						☑☒
Post – Design						P-sm/P-ser/T-Sm/T-Ser/O
Post – Length						mms
Post – C/P Ratio						C:P
Post – Fit						Good/Fair/Poor
Sclerosis – Coronal						Moderate/Severe
Sclerosis – Middle						Moderate/Severe
Sclerosis – Apical						Moderate/Severe
Perforation – Coronal						Mesial/Distal/?BL/Furcat
Perforation – Middle						Mesial/Distal/?BL/Furcat
Resorption – Coronal						Int/Ext-inf/Ext-rep/EIP
Resorption – Middle						Int/Ext-inf/Ext-rep/EIP
Resorption – Apical						Int/Ext-inf/Ext-rep/EIP
Root Resorption – Symmetry						Sym/Assym
Suspect Ledge – Coronal						☑☒
Suspect Ledge – Middle						☑☒
Suspect Ledge – Apical						☑☒

continued—

Table 1 *—continued*

Tooth no.						Code
Horizontal # – Coronal						With Repair/No Repair
Horizontal # – Middle						With Repair/No Repair
Horizontal # – Apical						With Repair/No Repair
Vertical # – Coronal						☑ ☒
Vertical # – Middle						☑ ☒
Vertical # – Apical						☑ ☒
Apical Foramen						<0.5/0.5-1/>1 mm
Hypercementosis						Moderate/Severe
(iv) Periradicular status						
PDL Widening – Coronal						Mesial/Distal/ ☒
PDL Widening – Middle						Mesial/Distal/ ☒
PDL – Widening – Apical						Mesial/Distal/ ☒
Lesion – Furcation						☑ ☒
Lesion – Middle						Mesial/Distal/ ☒
Lesion – Apical						Mesial/Distal/ ☒
Lesion Size						Maximum Diameter – mm
Condensing Osteitis						Cor/Mid/Apic/Furc
(v) Periodontal apparatus						
Vertical Bone Loss						Cor/Mid/Apic
Horizontal bone loss						Coronal/Middle/Apical
Furcation involvement						☑ ☒

radiographs from the same sample of 100 roots and 42 DPTs. This data was consigned 'Postgraduate Observations 3'.

Data analysis

The data from these four assessments were collected from the radiographic proformae and entered into a database software program (SPSS for Windows 6.1.2, SPSS Inc., Chicago, IL, USA). The original data was entered in the abbreviated word form as stipulated in the criteria. This was reconfigured into a numerical form for entry into the software.

Statistical analysis

Observer variation using Cohen's kappa test (Cohen 1960) was calculated for three sets of comparisons.

- Intra-observer variation between Postgraduate Observations 1 & 2.
- Inter-observer variation between Postgraduate Observations 2 and the 'Gold Standard'.
- Inter-observer variation between Postgraduate Observations 3 and the 'Gold Standard'.

Results

On analysis of the collected data, it was found that only certain categories had an adequate response rate on the form to allow statistical analysis. A number of categories were either deemed not relevant by the postgraduates and neither positive nor negative responses were noted (e.g. fractures, perforations, ledges, etc.) or too few roots were attributed to an observation (e.g. post, sclerosis, hypercementosis, etc.). In total, 25 categories were selected for analysis. The numerical results are summarized in Table 2. Descriptive terms were assigned to the kappa values and the comparative results are shown in Figure 2.

The reproducibility was dependent on the category being observed. Fifteen of the 25 observations achieved 'good' or 'very good' intraobserver kappa values as defined by Landis & Koch (1977) in Table 3. Initial inter-observer agreement was considerably poorer with only three out of the 25 observations having 'good' or 'very good' kappa values. An improvement occurred following the revision session with increased kappa values for all observations, and 16 observations achieving 'good' or 'very good' kappa values.

No statistical relationship between intra- and inter-examiner observations could be elicited.

Table 2 Categories selected for comparison and kappa values for assessments

Category	Intra-observer variation	1st inter-observer variation	2nd inter-observer variation
PA radiographic observations			
1. Restoration deepest	0.52968	0.25988	0.54989
2. Marginal deficiency	0.58662	0.39825	0.65486
3. Pulp chamber sclerosis	0.84188	0.08326	0.21592
4. Curvature	0.45655	0.46403	0.65621
5. Root fill – material	0.79438	0.64070	0.88218
6. Root fill – mm from apex	0.49778	0.43654	0.61572
7. Root fill – coronal voids	0.78506	0.55754	0.58278
8. Root fill – middle voids	0.67130	0.37977	0.52890
9. Root fill – apical voids	0.76863	0.56453	0.76968
10. Root fill – coronal adaptation	0.58274	0.34865	0.59075
11. Root fill – middle adaptation	0.70019	0.42008	0.60179
12. Root fill – apical adaptation	0.75248	0.34660	0.55807
13. Periodontal ligament widening – coronal	0.56180	0.37600	0.44533
14. Periodontal ligament widening – middle	0.60511	0.41477	0.47881
15. Periodontal ligament widening – apical	0.62600	0.23104	0.37160
16. Lesion furcation	0.72930	0.38712	0.64490
17. Lesion middle	0.62600	0.33593	0.70613
18. Lesion apical	0.27103	0.24241	0.60862
19. Lesion size	0.46192	0.41026	0.62790
DPT observations			
20. Teeth missing	0.71495	0.74963	0.94512
21. Root-filled teeth	0.75339	0.83809	0.83913
22. Periapical lesions	0.73406	0.58583	0.63640
23. Root-filled teeth with periapical lesion	0.85526	0.67991	0.68056
24. Localized periodontal disease	0.73642	0.53151	0.76479
25. Generalized periodontal disease	0.51133	0.00825	0.76932

Discussion

This preliminary study was carried out in an attempt to illuminate areas of weakness in the design and implementation of a new proforma, with the expectation that further investigation using a revised form and more refined training will ensure better reproducibility. It is important to understand that the objective of this study was not to test the form itself, or specifically to gauge the diversity of radiographic interpretation. The sample size viewed by each postgraduate student, dictated by logistics, was too small to give greater statistical significance (assuming it existed), and the length of time taken for each root assessment and the number of assessments made (potentially 1719 observations), meant an increase in sample size would overload the assessors and lead to further inaccuracies.

Although various researchers (Brynolf 1971, Antrim 1983, Welander *et al.* 1983) have suggested strategies for improving the reliability of diagnosis from radiographic images, the selection of postgraduate students, radiographs and viewing conditions in this study were considered to be a realistic representation of radio-

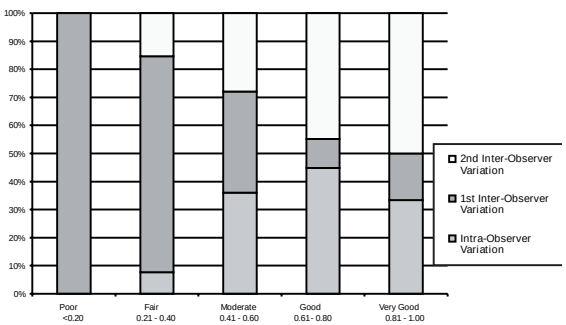

Figure 2 The percentage values of strength of agreement for each assessment

Table 3 Kappa value definitions (adapted from Landis & Koch 1977)

Kappa value	Strength of agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

graphic assessment in the prevailing clinical environment at the Eastman Dental Institute, London. Since the radiographic proforma is for general clinical use on the New Patient Assessment Clinics, the conditions needed to be replicated.

Rather than request the postgraduates to assess the same set of radiographs for all three sets of comparisons, it was decided that two different sets, randomly selected from the pool of 100 roots and 42 DPTs would be used. As there was no intention to test inter-examiner variation between the postgraduates, this method was used to ensure that no bias was introduced following conferring or discussion amongst the postgraduates.

The 'Gold Standard' was achieved through consensus opinion. Other studies use a 'Gold Standard' to indicate a true result, established through confirmation by histological examination (Cayley & Holt 1997). There were no 'true' results in the assessment of these radiographs as no true confirmation was obtained.

The kappa test determines the overall levels of agreement between observers, whilst correcting for the proportion of agreement expected by chance. This test is now the most widely accepted measure of agreement when considering data arising from nominal or ordinal scales (Brennan & Silman 1992). In practical terms, values above 0.61 (good or very good) are taken to mean a high standard of agreement, and a level at which the data can be accepted as being reproducible.

Results of intraobserver variation suggest that observer reproducibility is poor with only 15 of the 25 observations (60%) having an acceptably high standard of agreement.

Brynolf (1971) and Goldman *et al.* (1974) report intra-observer agreement levels of 72–87%. These results are incomparable to those of this study as they relate to success/failure decisions based on a variety of radiographic parameters, not specific observations as in this study.

Previous studies into inter-observer variation have found that agreement is poor, despite strategies attempting to improve reproducibility, such as observer calibration, strict criteria, and scoring indices (Eckerbom *et al.* 1986, Rohlin & Akerblom 1992). The results from the first inter-observer assessments in this study are broadly in agreement with the previous ones, with only two out of the 25 observations considered to have an acceptable standard of agreement. The attempt to improve agreement by re-enforcing the assessment criteria was successful in that 23 of the 25 observations achieved higher kappa scores, with 16 of the 25 reaching an acceptable level of agreement. This improved result following a training session is better than that achieved by Cayley & Holt (1997) using a similar experimental method in the radiographic diagnosis of interproximal caries. However, the overall result still indicates that the data entered into the radiographic proformae is not reproducible and would therefore adversely affect the reliability of any related epidemiological or treatment outcome study.

A level of intra-observer consistency suggests the observers were clear in their minds about the definition of a given category and its interpretation. However, this did not necessarily correlate with the actual definition given or to its interpretation. It was therefore essential to clarify the definition and to check its interpretation. A possible reason for the poor inter-observer agreement values may be because the postgraduate observers were unable to follow the definitions because they were poorly defined, or difficult to interpret.

It is questionable as to whether the physical layout of the form had a negative effect on performance. The postgraduates did not consider the forms to be 'user friendly' and they were difficult to use, containing too much information on the page.

Some categories did not offer appropriate negative response options. For example 'not applicable' responses should have been included in a number of categories to distinguish between a 'no entry' (empty) field and a negative observed response.

Further investigation is required to assess the relative importance of each of these possibilities and therefore identify solutions to the problems of accuracy and reproducibility.

The study was able to highlight observations with initially poor levels of inter-observer agreement that improved with training and those that did not improve with training. Of particular note were *Pulp chamber sclerosis, Periodontal Ligament Space Widening -apically*, and *Root Filling – middle voids* assessments that produced consistently high levels of inter-examiner disagreement. Intra-examiner agreement values for these observations were either good or very good. By identifying sources of inaccuracies such as these, a greater emphasis can be placed on these definitions whilst training the postgraduate observers.

As a result of this study and further discussion with the postgraduate students involved, modifications to the layout of the form have been undertaken. The definitions have been reframed and training sessions scheduled at regular intervals. A further study to assess the effect of these changes is planned. The

refinement of this system of data collection should facilitate prospective analysis of outcome of endodontic treatment and factors that influence it.

References

Antrim DD (1983) Reading the radiograph: a comparison of viewing techniques. *Journal of Endodontics* **9**, 502–5.

Bender IB, Seltzer S, Soltanoff W (1966) Endodontic success a reappraisal of criteria. II. *Oral Surgery, Oral Medicine and Oral Pathology* **22**, 790–802.

Brennan P, Silman A (1992) Statistical methods for assessing observer variability in clinical measures. *British Medical Journal* **304**, 1491–4.

Brynolf I (1971) Improved viewing facilities for better roentgenodiagnosis. *Oral Surgery, Oral Medicine and Oral Pathology* **32**, 808–11.

Cayley AS, Holt RD (1997) The influence of audit on the diagnosis of occlusal caries. *Caries Research* **31**, 97–102.

Chow NM (1993) A retrospective study of patient attendance. (MSc Project) Dept. of Periodontology, Eastman Dental Institute, University of London.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Eckerbom M, Andersson J-E, Magnusson T (1986) Inter-observer variation in radiographic examination of endodontic variables. *Endodontics and Dental Traumatology* **2**, 243–6.

Goldman M, Pearson AH, Darzenta N (1974) Reliability of radiographic interpretations. *Oral Surgery, Oral Medicine and Oral Pathology* **38**, 287–93.

Goldstein IL, Mobley WH, Chellemi SJ (1971) The observer process in the visual interpretation of radiographs. *Journal of Dental Education* **35**, 485–91.

Grahnen H, Hansson L (1961) The prognosis of pulp and root canal therapy. A clinical and radiological follow-up examination. *Odontologisk Revy* **12**, 146–65.

Grondahl H-G (1979) The influence of observer performance in radiographic caries diagnosis. *Swedish Dental Journal* **3**, 101–7.

Halse A, Molven O (1986) A strategy for the diagnosis of periapical pathosis. *Journal of Endodontics* **12**, 534–8.

Harty FJ, Parkins BJ, Wengraf AM (1970) Success rate in root canal therapy. *British Dental Journal* **128**, 65–70.

Kerekes K, Tronstad L (1979) Long term results of endodontic treatment performed with a standardised technique. *Journal of Endodontics* **5**, 83–90.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–74.

Mokbel NM (1994) Post surgical periodontal maintenance. (MSc Project). Dept. of Periodontology, Eastman Dental Institute, University of London.

Reit C (1987) The influence of observer calibration on radiographic periapical diagnosis. *International Endodontic Journal* **20**, 75–81.

Rohlin M, Akerblom A (1992) Individualised periapical radiography determined by clinical and panoramic examination. *Dento-Maxillo-Facial Radiology* **21**, 135–41.

Rohlin M, Kullendorff B, Ahlqwist M, Stenstrom B (1991) Observer performance in the assessment of periapical pathology: a comparison of panoramic with periapical radiography. *Dento-Maxillo-Facial Radiology* **20**, 127–31.

Sjogren V, Hagglund B, Sundqvist G, Wing K (1990) Factors affecting the long-term results of endodontic treatment. *Journal of Endodontics* **16**, 498–504.

Smith CS, Setchell DJ, Harty FJ (1993) Factors influencing the success of conventional root canal therapy a five year retrospective study. *International Endodontic Journal* **26**, 321–33.

Stheeman SE, Mileman PA, Van't Hof MA, Van der Stelt PF (1995) Diagnostic confidence and the accuracy of treatment decisions for radiopaque periapical lesions. *International Endodontic Journal* **28**, 121–8.

Valachovic RW, Douglas CW, Berkley CS, McNeil BJ, Chauncey HH (1986) Examiner reliability in dental radiography. *Journal of Dental Research* **65**, 432–6.

Welander U, McDavid WD, Higgins NM, Morris CR (1983) The effect of viewing conditions on the perceptibility of radiographic details. *Oral Surgery, Oral Medicine and Oral Pathology* **56**, 651–4.