ACTA ODONTOLOGICA SCANDINAVICA
2021, VOL. 79, NO. 7, 554–561
https://doi.org/10.1080/00016357.2021.1910728

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Calibration improves observer reliability in detecting periapical pathology on panoramic radiographs

Dan Sebring[a], Thomas Kvist[a], Kåre Buhlin[b], Peter Jonasson[a], EndoReCo* and Henrik Lund[c]

[a]Department of Endodontology, Institute of Odontology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [b]Department of Dental Medicine, Division of Periodontology, Karolinska Institutet, Huddinge, Sweden; [c]Department of Oral Maxillofacial Radiology, Institute of Odontology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

ABSTRACT

Objective: To determine whether calibration improves observer reliability when assessing DMFT-score, root-filled teeth and periapical lesions on panoramic radiographs.

Material and methods: A sample of 100 panoramic radiographs was selected from a cohort of myocardial infarction patients ($n = 797$) and matched controls ($n = 796$). The following variables were assessed: DMFT-score, remaining teeth, root-filled teeth and periapical lesions. Two specialists, an endodontist and a radiologist, served as reference examiners and undertook two separate assessments. Disagreement cases were jointly assessed and the final results were used as the reference standard. Three observers undertook three separate assessments, the first without prior training, the second after calibration against the reference standard and the third with the sample concealed in the complete material. Statistical analysis was made with Wilcoxon Signed rank test and Sign test. Agreement was calculated as Intraclass Correlation Coefficient (ICC) (95% CI) and Weighted Kappa (κ) (95% CI).

Results: Periapical lesions disclosed high inter-observer variability for the reference examiners and diverged significantly between the observers and the reference standard. For the reference examiners, inter-observer agreement was $\kappa = 0.53$. The observers, in their first assessments had κ values ranging from 0.22 to 0.60 in relation to the reference standard. Following calibration, the κ values increased, ranging from 0.59 to 0.80. For the third assessment, the κ values ranged from 0.54 to 0.75. DMFT-score, remaining teeth and root-filled teeth disclosed high reliability throughout all assessments (ICC = 0.88-0.98 and $\kappa = 0.98$–0.99).

Conclusions: DMFT-score, remaining teeth and root-filled teeth can be reliably assessed on panoramic radiographs. Calibration against a reference standard improves observer reliability in the detection of periapical lesions.

ARTICLE HISTORY
Received 20 January 2021
Revised 23 March 2021
Accepted 23 March 2021

KEYWORDS
Apical periodontitis; calibration; observer variation; panoramic radiograph

Introduction

Radiography is an essential tool in many areas of dentistry, not the least in the field of endodontics. Endodontic inflammatory conditions, such as pulpitis and apical periodontitis, do not always exhibit any clinical signs or symptoms. Detection of carious and periapical lesions are both highly dependent on radiographic examination. Intra-oral radiography is the most commonly used method for caries and periapical diagnosis. Cone-beam computed tomography (CBCT) has recently become a useful adjunct in certain cases. Compared with intra-oral radiography, panoramic radiography (PR) is of limited value for endodontic diagnosis [1,2] or for detection of approximal caries [3]. Compared with CBCT, an even greater discrepancy is reported [4,5]. However, in the absence of an independent gold standard, the results should be interpreted with caution.

PR offers a rapid, simple means of providing a broad overview of both jaws and teeth and is likely to be well accepted by patients and/or study participants. It is therefore commonly used in epidemiological studies. Variables such as the number of remaining teeth, restorative therapy and endodontic treatments are accurately disclosed. Despite low sensitivity in detecting caries and periapical pathology, for larger epidemiological studies PR is considered to provide an adequate overview [1,2].

Any study reliant on interpretation of a radiograph, regardless of method, is subject to intra- and inter-observer variation. This issue was clearly demonstrated in the classic study by Goldman et al. [6]. Six observers were instructed to determine independently the periapical status of endodontically treated teeth on intra-oral radiographs: there was agreement in fewer than 50% of the cases. Subsequent studies have also reported on the issue of reliability in radiographic

CONTACT Dan Sebring dan.sebring@gu.se Department of Endodontology, Institute of Odontology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
*Collaborators: Lars Bjørndal, Victoria S. Dawson, Helena Fransson, Fredrik Frisk, Peter Jonasson, Thomas Kvist, Merete Markvart, and Maria Pigg

assessment of periapical pathology, both for conventional [7,8] and digital [9] intra-oral radiographs, and more recently also for CBCT [10]. Epidemiological and large-scale studies often involve multiple observers and variation in assessment between two or more observers can greatly influence the results. Uncertainties in the registration of independent variables will have implications for the reliability of any conclusions drawn about the material.

Strategies have been proposed for improving reliability in radiographic assessment. The positive effects of observer calibration have been reported in a number of studies in which periapical status was assessed on intra-oral radiographs [7,11,12]. However, to our knowledge, there are no such studies with respect to periapical pathology detected exclusively on PR. As PR is currently widely used in epidemiological studies, and this is likely to continue, e.g. for studies investigating associations between endodontic inflammatory disease and other conditions, the issue of reliability must be addressed. The aim of this study was to investigate whether calibration against a reference standard improves observer reliability when using PR to assess variables indicative of endodontic inflammatory disease, including calculation of DMFT-score and detection of root-filled teeth and periapical lesions.

Materials and methods

In accordance with the instructions for calibrating the periapical index (PAI) [13], the sample selected for this study comprised 100 PRs from a cohort previously described by Ryden et al. [14]. In short, the cohort comprised clinical data and PRs from 797 patients who had suffered an initial myocardial infarction and 796 controls, matched for age, sex and postal code area. Dental examinations, as well as radiographic examinations with either film or digital PRs, were undertaken during 2010-2014 at dental clinics at, or near, 17 different Swedish hospitals.

Training and calibration procedures

The study layout is illustrated in Figure 1.

Initially, two experienced specialists, one endodontist (TK) and one radiologist (HL), served as reference examiners. Prior to the calibration process instructions were formulated for the radiographic assessment of the following variables; remaining teeth, root-filled teeth, teeth with periapical lesions and DMFT-score, calculated from assessment of decayed, missing and filled teeth.

The instructions were as follows: 'Remaining teeth' comprise, besides apparent ones, also root remnants which are at least partially connected to the surrounding bone, impacted teeth covered by bone (mucosa and follicle) by no more than 2/3 of the crown and also any deciduous teeth. 'Root-filled teeth' are any teeth in which the pulp cavity and/or root canals contain radiopaque material. 'Teeth with periapical lesions' are any teeth that present distinct periapical bone destruction, noticeably larger than a widened periodontal space, around at least one root. For 'DMFT-score'

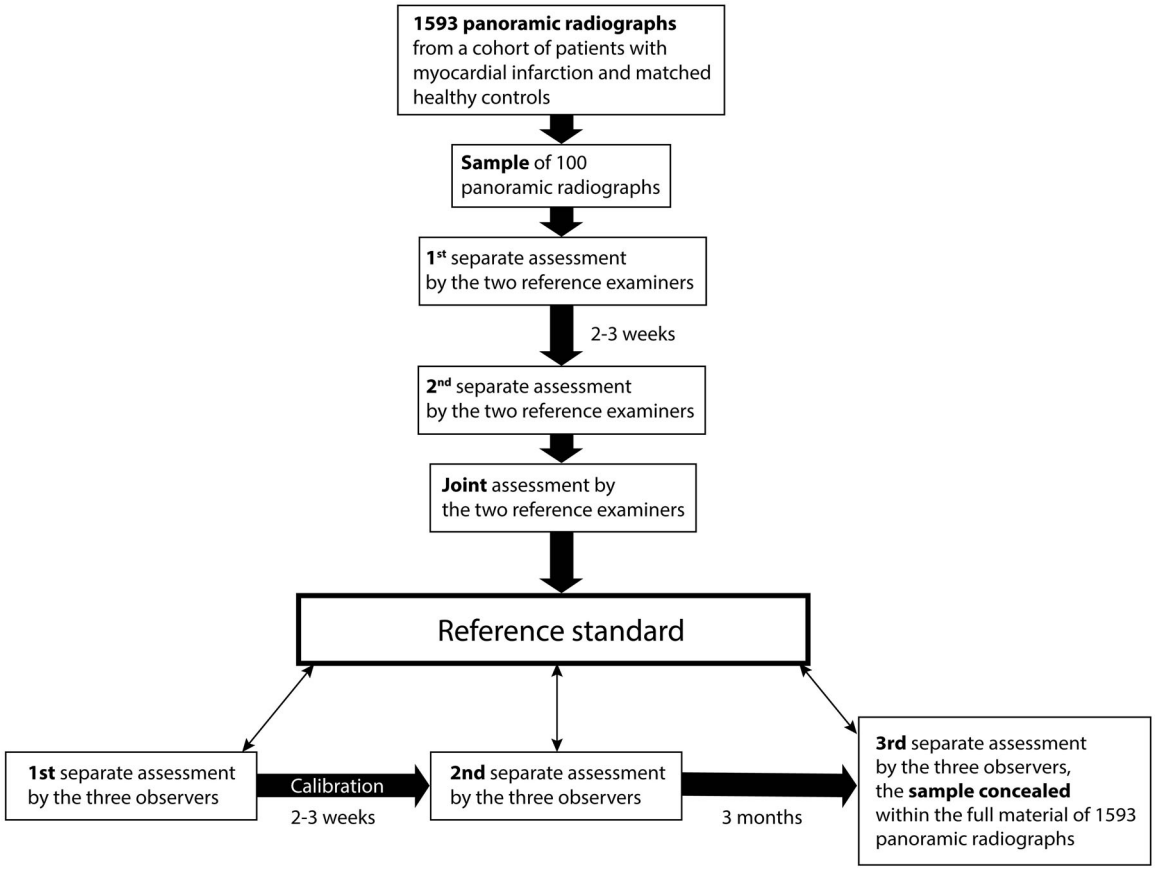

Figure 1. Flow chart of the study.

root remnants lacking restoration are registered as decayed teeth and dental implants are registered as missing teeth.

All PR assessments were made on the same high-resolution screen, in a dimly lit room. All images were assessed using the image-processing program ImageJ (U.S. National Institutes of Health, Bethesda, Maryland, USA, https://imagej.nih.gov/ij/), allowing adjustment of contrast and use of magnification tools. The two specialists assessed the sample independently and re-examined the material 3-4 weeks later, in order to determine intra-observer variation. The results were analysed and all cases with either intra- or inter-observer variations were subsequently subjected to a third, joint assessment, in which contradictory findings were discussed until consensus was reached. The results from the joint assessment then served as the reference standard for all future assessments of this sample.

Three observers, one general dentist (observer 1) and two fifth-year dental students (observers 2 and 3), were presented with written instructions for the radiographic assessment. Each observer made an initial, independent assessment of the sample. Thereafter the results were compared with the reference standard and the agreement was calculated. For calibration, each observer was individually presented with the results and given instructions on how to achieve closer agreement with the reference standard. Specifically, two observers (2 and 3) were directed to apply the criteria for registering periapical lesions more stringently, while one observer (1) was directed to lower slightly the threshold for reporting periapical lesions, to match the reference standard more closely. A second assessment of the sample followed and the results were again compared with the reference standard.

In order to evaluate the long-term effect of the calibration process, a third assessment was undertaken approximately three months later. This time the sample was randomly concealed in the complete material of 1593 PRs. When the entire material had been reassessed, agreement between the sample and the reference standard was again calculated.

Statistical analyses

Intra-observer reliability (relevant for the reference examiners) was analysed for DMFT-score and remaining teeth with the distribution of the difference between the measurements, given as mean, SD, median, minimum and maximum, intra-individual SD, repeatability and Intraclass Correlation Coefficient (ICC) with 95% CI.

Inter-observer reliability (relevant for the reference examiners) and observer-reference reliability (relevant for the observers) were analysed for DMFT-score and remaining teeth with the distribution of the difference between measurements, expressed as the mean, SD, median, minimum and maximum, Limits of Agreement and Intraclass Correlation Coefficient (ICC) with 95% CI.

Inter- and intra-observer reliability and observer-reference reliability were analysed for root-filled teeth and teeth with periapical lesions with the distribution of the difference between measurements expressed as the mean, SD, median, minimum and maximum, Weighted Kappa with 95% CI and percent agreement. Interpretation of numeric Kappa-values is described as $\kappa \leq 0.20$ = poor, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = good, and 0.81–1.00 = very good [15].

To calculate observer-reference reliability registrations from the radiographic assessment of each observer and session were compared to the reference standard findings. To determine whether there was significant improvement between the initial and second assessments by observers 1-3, the absolute difference between the observer and the reference standard was calculated for each image for both the first and the second assessment. The difference between these two assessments was analysed with the Wilcoxon Signed rank test and Sign test. Bland-Altman difference plots and scatter plots illustrate inter-observer reliability for DMFT-score.

Ethical approval

The PAROKRANK study was approved by the Regional Ethics Committee in Stockholm (Dnr: 2008/152-31/2) and all participants provided written informed consent prior to the study procedures. The study was conducted according to the principles outlined in the Helsinki Declaration.

Results

The results of the initial two independent assessments by each reference examiner are presented in Table 1. For the variables DMFT-score, remaining teeth and root-filled teeth the ICC and Weighted Kappa values were high overall (ICC = 0.97–0.98 and 0.98–1.00, κ = 0.95–0.98 respectively), i.e. 'very good' intra-observer agreement. However, assessment of teeth with periapical lesions revealed an intra-observer agreement of 0.69 and 0.72, i.e. 'good' agreement.

The results from the inter-observer agreement analysis are presented in Table 2. The observations of the two reference examiners differed with respect to DMFT-score ($p = 0.013$) and teeth with periapical lesions ($p < .0001$). For DMFT-score, remaining teeth and root-filled teeth ICC and Weighted Kappa were overall high (ICC = 0.95 and 0.97, $\kappa = 0.96$ respectively), i.e. 'very good' inter-observer agreement. However, with respect to teeth with periapical lesions, inter-observer agreement was 0.53, i.e. 'moderate'.

Observer-reference agreement for observers 1, 2 and 3 in relation to the reference standard for DMFT-score and remaining teeth is presented in Table 3. Throughout all three assessments, both DMFT-score and remaining teeth exhibited 'very good' agreement (ICC = 0.89–98 and 0.88–0.99 respectively) among all three observers and the reference standard. Figure 2 illustrates Bland-Altman difference plots and scatter plots for the variable DMFT-score for assessment 3 for all observers.

Observer-reference agreement for observers 1, 2 and 3 in relation to the reference standard for the variables root-filled teeth and teeth with periapical lesions is presented in Table 4. As with the assessments by the reference examiners, teeth with periapical lesions exhibited the greatest variability,

Table 1. Intra-observer reliability for the reference examiners.

	Assessment 1 Mean (SD) Median (min; max)	Assessment 2 Mean (SD) Median (min; max)	Change from assessment 1 to 2 Mean (SD) Median (min; max)	Intra-individual standard deviation (IISD)	Repeatability 2.77*IISD	ICC (95% CI)	Weighted Kappa (95% CI)	Percent agreement
Endodontist								
DMFT-score	21.4 (5.6) 21.5 (0.0; 32.0)	21.4 (5.6) 22.0 (0.0; 32.0)	−0.03 (1.3) 0.00 (−7.0; 4.0)	0.91	2.52	0.97 (0.96; 0.98)	–	–
Remaining teeth	25.9 (4.9) 27.0 (5.0; 32.0)	26.0 (4.9) 27.0 (5.0; 32.0)	0.07 (1.03) 0.0 (−4.0; 8.0)	0.73	2.02	0.98 (0.97; 0.99)	–	–
Root–filled teeth	2.2 (2.3) 2.0 (0.0; 11.0)	2.2 (2.1) 2.0 (0.0; 11.0)	−0.07 (0.5) 0.0 (−4.0; 1.0)	–	–	–	0.95 (0.91; 0.99)	94.0%
Teeth with periapical lesions	1.1 (1.3) 1.0 (0.0; 6.0)	1.0 (1.1) 1.0 (0.0; 5.0)	−0.1 (0.8) 0.0 (−5.0; 1.0)	–	–	–	0.69 (0.58; 0.79)	68.0%
Radiologist								
DMFT-score	21.8 (5.5) 22.5 (2.0; 32.0)	21.9 (5.6) 23.0 (1.0; 32.0)	0.1 (1.1) 0.00 (−4.0; 6.0)	0.80	2.22	0.98 (0.97; 0.99)	–	–
Remaining teeth	25.9 (4.8) 27.0 (5.0; 32.0)	25.9 (4.8) 27.0 (5.0; 32.0)	−0.08 (0.52) 0.0 (−2.0; 1.0)	0.30	0.83	1.00 (0.99; 1.00)	–	–
Root-filled teeth	2.2 (2.2) 2.0 (0.0; 11.0)	2.2 (2.1) 2.0 (0.0; 11.0)	0.03 (0.2) 0.0 (−1.0; 1.0)	–	–	–	0.98 (0.96; 1.00)	95.0%
Teeth with periapical lesions	0.7 (1.1) 0.0 (0.0; 7.0)	0.6 (1.0) 0.0 (0.0; 5.0)	−0.1 (0.6) 0.0 (−1.0; 1.0)	–	–	–	0.72 (0.61; 0.83)	78.0%

For comparison between assessments, the Wilcoxon signed rank test was used for the variables DMFT-score and remaining teeth and sign test was used for the variables root-filled teeth and teeth with periapical lesions.

Table 2. Inter-observer reliability for the reference examiners.

	Endodontist Mean (SD) Median (min; max)	Radiologist Mean (SD) Median (min; max)	Difference between radiologist and endodontist Mean (SD) Median (min; max) *p*-value	Limits of agreement	ICC (95% CI)	Weighted Kappa (95% CI)	Percent agreement
DMFT-score	21.4 (5.6) 21.5 (0.0; 32.0)	21.8 (5.5) 22.5 (2.0; 32.0)	0.41 (1.7) 0.00 (−5.0; 7.0) 0.013	−3.0–3.8	0.95 (0.93; 0.97)	–	–
Remaining teeth	25.9 (4.9) 27.0 (5.0; 32.0)	25.9 (4.8) 27.0 (5.0; 32.0)	0.04 (1.16) 0.0 (-4.0; 8.0) 1.00	−2.2–2.3	0.97 (0.96; 0.98)	–	–
Root-filled teeth	2.2 (2.3) 2.0 (0.0; 11.0)	2.2 (2.2) 2.0 (0.0; 11.0)	−0.04 (0.4) 0.0 (−3.0; 1.0) 0.52	–	–	0.96 (0.92; 0.99)	93.0
Teeth with periapical lesions	1.1 (1.3) 1.0 (0.0; 6.0)	0.7 (1.1) 0.0 (0.0; 7.0)	−0.5 (0.9) 0.0 (−5.0; 2.0) <.0001	–	–	0.53 (0.43; 0.64)	60.0

For comparison between assessments, the Wilcoxon signed rank test was used for the variables dmft-score and remaining teeth and sign test was used for the variables roo-filled teeth and teeth with periapical lesions.

with 'fair' to 'moderate' agreements ($\kappa = 0.22$, 0.30 and 0.60) in assessment 1. For root-filled teeth agreement was 'very good' ($\kappa = 0.98$). The difference in mean (SD) for the variable teeth with periapical lesions revealed that observer 1 registered fewer and observers 2 and 3 registered more such lesions than the reference standard. In assessment 2, following calibration, all observers achieved a significant improvement in agreement on periapical lesions. In assessment 3, all observers showed improved agreement compared to pre-calibration, but less improvement than in assessment 2. The results for the variable teeth with periapical lesions for all observers and assessments are illustrated in Figure 3.

Discussion

The present study explores a method for calibrating observers with respect to radiographic assessment of endodontic variables in a large material of PRs. The statistical analyses revealed high inter- and intra-observer reliability for the reference examiners, and high observer-reference reliability for the observers, with 'very good' agreement for all variables, except for the number of teeth with periapical lesions. For the observers, this was true before as well as after calibration. The number of remaining teeth and the number of root-filled teeth were expected to be more or less consistent and the results were in accordance with a previous study specifically assessing PRs [1]. The results confirm that the method of calibration investigated in this study significantly reduces observer variation in diagnosing endodontic inflammatory disease, particularly periapical lesions, in PRs. However, as some variation persisted, intervention by calibration failed to achieve complete reliability.

Different methods have been applied to the calibration of observers in order to improve reliability in the examination of intra-oral radiographs. These include using reference images [13], having the observers themselves establish the criteria for the radiographic assessment [7,12,16] or presenting the results from a baseline assessment and requesting the observers to reduce false positive diagnoses in subsequent assessments [8,11]. The present study utilised ideas from the latter study, in conjunction with allowing experienced observers state a reference standard for the material being investigated.

Table 3. Observer-reference reliability for the variables DMFT-score and Remaining teeth.

	Observer 1		Observer 2		Observer 3	
	Diff. Reference standard Mean (SD) Median (min; max)	ICC (95% CI) Limits of agreement	Diff. Reference standard Mean (SD) Median (min; max)	ICC (95% CI) Limits of agreement	Diff. Reference standard Mean (SD) Median (min; max)	ICC (95% CI) Limits of agreement
Assessment 1						
DMFT-score	−0.27 (1.08) 0.00 (−4; 2)	0.98 (0.97; 0.99) −2.40–1.85	−0.53 (2.62) 0.00 (−2; 2)	0.89 (0.84; 0.93) −5.66-4.60	−0.16 (1.23) 0.00 (−4; 2)	0.98 (0.96; 0.98) −2.57–2.25
Remaining teeth	0.15 (0.66) 0.00 (−1; 4)	0.99 (0.99; 0.99) −1.14–1.44	0.42 (2.45) 0.00 (−1; 24)	0.88 (0.83; 0.92) −4.38–5.22	0.22 (0.63) 0.00 (−1; 4)	0.99 (0.98; 0.99) −1.02–1.45
Assessment 2						
DMFT-score	−0.20 (1.10) 0.00 (−4; 3)	0.98 (0.97; 0.99) −2.36–1.96	−0.31 (1.22) 0.00 (−4; 3)	0.97 (0.96; 0.98) −2.70–2.08	−0.33 (1.20) 0.00 (−4; 3)	0.97 (0.96; 0.98) −2.68–2.02
Remaining teeth	0.16 (0.66) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.14–1.46	0.17 (0.62) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.05–1.39	0.18 (0.63) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.05–1.41
Assessment 3						
DMFT-score	0.11 (1.11) 0.00 (−4; 4)	0.98 (0.97; 0.99) −2.06–2.28	−0.27 (1.31) 0.00 (−5; 3)	0.97 (0.96; 0.98) −2.84–2.30	−0.40 (1.16) 0.00 (−4; 3)	0.98 (0.96; 0.98) −2.67–1.86
Remaining teeth	0.12 (0.61) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.07–1.31	0.16 (0.62) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.05–1.37	0.18 (0.64) 0.00 (−2; 4)	0.99 (0.99; 0.99) −1.08–1.44

For comparison between assessments, the Wilcoxon signed rank test was used.

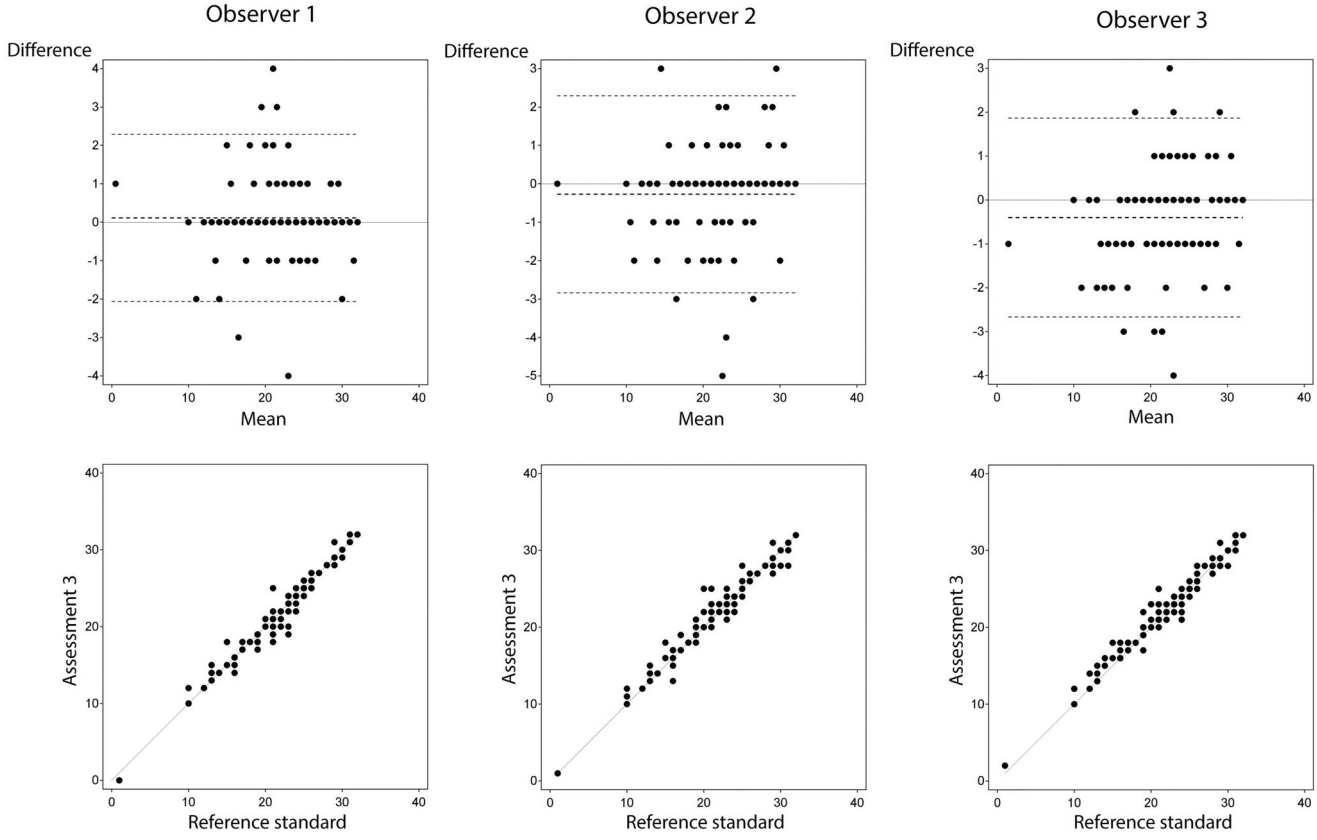

Figure 2. Bland–Altman difference plots (above) and scatter plots (below) using the variable DMFT-score for assessment 3 and the reference standard for observers 1, 2 and 3. In the Bland–Altman difference plots, the centre dashed line represents the overall mean of the differences and the outer dashed lines show the range containing the mean of the differences ±1.96 standard deviations, referred to as the limits of agreement. The scatter plots illustrate the correlation between assessment 3 and the reference standard.

Initial assessment of teeth with periapical lesions showed 'good' intra-observer agreement for both the endodontist and the radiologist ($\kappa = 0.69$ and 0.72). Inter-observer agreement, however, was 'moderate' ($\kappa = 0.53$) and lower than reported by a previous study on reliability of periapical diagnosis using PR ($\kappa = 0.63$, 0.71 and 0.77) [5]. The two specialists compiled the reference standard by consensus prior to the initial diagnostic session by three less experienced observers.

After calibration, all three observers achieved significant improvement in agreement with the reference standard ($\kappa = 0.59$, 0.64 and 0.80) for registrations of teeth with periapical lesions. These results are comparable to previous studies evaluating other types of calibration methods. Eckerbom et al. [7] calibrated two observers by discussing contradictory findings in a separate set of radiographs and disclosed an agreement, presented as Scott's π, of $\pi = 0.64$. Reit [11] studied two methods, one similar to Eckerbom et al. [7] and

Table 4. Observer-reference reliability for the variables root-filled teeth and teeth with periapical lesions.

	Observer 1		Observer 2		Observer 3	
	Diff. Reference standard Mean (SD) Median (min; max)	Weighted Kappa (95% CI) Percent agreement	Diff. Reference standard Mean (SD) Median (min; max)	Weighted Kappa (95% CI) Percent agreement	Diff. Reference standard Mean (SD) Median (min; max)	Weighted Kappa (95% CI) Percent agreement
Assessment 1						
Root-filled teeth	−0.03 (0.22) 0.00 (−1; 1)	0.98 (0.96; 1.00) 95.0%	−0.01 (0.23) 0.00 (−1; 1)	0.98 (0.96; 1.00) 95.0%	0.00 (0.27) 0.00 (−1; 1)	0.98 (0.95; 0.99) 94.0%
Teeth with periapical lesions	0.16 (0.72) 0.00 (−2; 2)	0.60 (0.49; 0.72) 67.0%	−0.98 (0.97) −1.00 (−3; 1)	0.30 (0.21; 0.40) 29.0%	−1.30 (1.05) −1.00 (−4; 2)	0.22 (0.13; 0.30) 19.0%
Assessment 2						
Root-filled teeth	0.01(0.23) 0.00 (−1; 1)	0.98 (0.96; 1.00) 95.0%	−0.03 (0.22) 0.00 (−1; 1)	0.98 (0.96; 1.00) 95.0%	−0.03 (0.22) 0.00 (−1;1)	0.98 (0.96; 1.00) 95.0%
Teeth with periapical lesions	−0.15 (0.56) 0.00 (-3; 1)	0.80 (0.72; 0.88) 81.0%	0.08 (0.68) 0.00 (-2; 2)	0.64 (0.52; 0.77) 69.0%	0.01 (0.72) 0.00 (-2; 2)	0.59 (0.47; 0.71) 61.0%
Assessment 3						
Root-filled teeth	−0.02 (0.14) 0.00 (−1; 0)	0.99 (0.98; 1.00) 98.0%	−0.03 (0.22) 0.00 (−1; 1)	0.98 (0.96; 1.00) 95.0%	0.00 (0.20) 0.00 (−1; 1)	0.98 (0.96; 1.00) 96.0%
Teeth with periapical lesions	−0.06 (0.53) 0.00 (−1; 1)	0.75 (0.66; 0.83) 72.0%	0.10 (0.69) 0.00 (−2; 2)	0.59 (0.47; 0.71) 61.0%	0.03 (0.78) 0.00 (−3; 2)	0.54 (0.41; 0.98) 59.0%

For comparison between assessments, sign test was used.

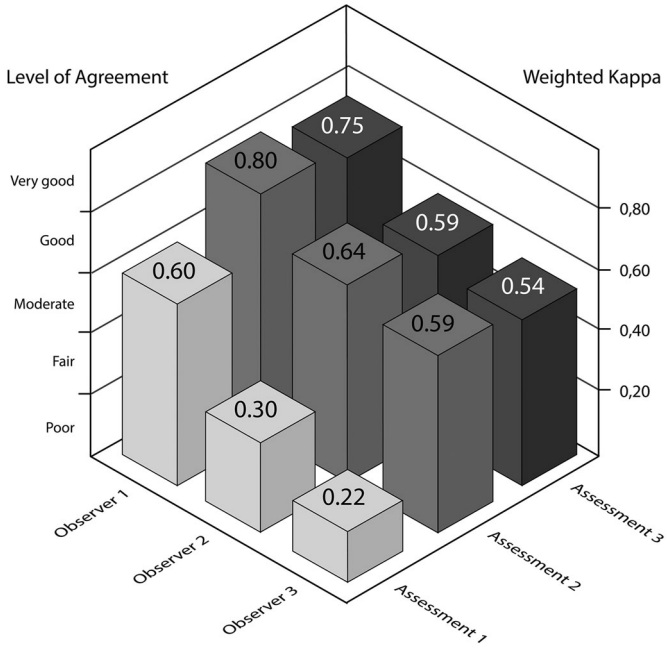

Figure 3. Bar chart, illustrating the Level of Agreement to the left and Weighted Kappa-values to the right, for the three assessments by observers 1, 2 and 3 for the variable teeth with periapical lesions.

one intended to reduce false positive diagnoses. Although both methods achieved improved agreement, the changes were significant only with respect to scores for indisputably healthy teeth. In a study by Saunders et al. [12], two external observers established a gold standard: three observers were subsequently calibrated by discussing contradictory results in the initial assessment. After calibration, the Kappa increased from 0.59 to 0.64.

In the present study, a third assessment was undertaken three months after calibration and with the sample concealed in the entire material. Compared with assessment 2, all observers exhibited a reduction in agreement with the reference standard ($\kappa = 0.54$, 0.59 and 0.75). This is in accordance with the conclusions by Reit [11], questioning the long-term benefits of calibration. In his study a third assessment was undertaken six months after calibration: the results more closely resembled those of the first assessment, prior to calibration.

Various measures have been proposed to standardise radiographic diagnosis. Viewing conditions can impact diagnosis [17]. Observer-related factors, e.g. mental state and fatigue have been implicated. Moreover, if the observers themselves have treated the teeth being assessed, they tend to score them differently [6]. In the present study, all the PRs were assessed under standardised condition, i.e. on the same high-resolution screen in a dimly lit room. The observers were instructed to assess no more than 25 PRs in the same session. All images were unidentifiable and none of the participants had been treated by the observers.

Alternatives to PR include full-mouth surveys with intra-oral radiographs or CBCT. Full-mouth surveys may require up to 20 radiographic images for a complete view of the oral condition of the patient. Though providing images with excellent detail, full-mouth surveys are also more time-consuming. It may be argued that they cause greater inconvenience for both researchers and patients participating in large cohort studies involving many medical, as well as oral examinations [14]. Also, the issue of radiation is worth addressing. However, the understanding that PR is equivalent to just a few intra-oral images has recently been questioned [18]. An updated method for calculating effective dose that, unlike before, includes salivary glands and oral mucosa, has resulted in reassessment of the effective dose from common dental radiographic examinations. Although the difference in measured effective dose has reduced, PR still results in a lower radiation dose (14.2-24.3 μSv) compared to full-mouth surveys (34.9 μSv) [19]. The advent of CBCT has led to major changes in endodontic diagnostics in daily practice, but has also called into question the scientific basis for evaluating treatment results [20]. However, recent studies in diagnostic accuracy using histological specimens as a reference have highlighted that even with this highly sensitive radiographic tool, misinterpretations of the nature of observed lesions may frequently occur, especially in root-filled teeth [21]. The

application of CBCT in comprehensive large-scale studies is also limited because several exposures are needed to fully image both jaws with adequate image quality. This not only increases costs and the time required for each examination, but also significantly increases the radiation dose [20].

The present study confirms that the reliability of the diagnostic procedure can be improved. However, the ambiguity of the findings on PRs still challenges the validity of the results. This becomes apparent when also the reference examiners, two experienced specialists, disclosed substantial variation in their separate assessments of periapical lesions. There is to date no evidence-based, practical, ethical method to determine the 'true' biological status of the periapical region. However, despite the limited radiographic diagnostic accuracy, relative differences between two groups, e.g. patients and healthy controls, can still be used to study potential associations.

In this context, Reit and Gröndahl demonstrated the importance of maintaining strict criteria for diagnoses and minimising false positive diagnoses, in order to reduce observer variation [8]. In their study six observers used a five-grade rating scale based on diagnostic certainty, the periapical probability index, when assessing periapical disease on intraoral radiographs. A high percentage of true positive registrations was consistently coupled with a higher percentage of false positive registrations. Interobserver variation was attributable to the fact that each observer adopted different criteria for diagnosis of periapical disease. The authors demonstrated that relative differences between groups could best be disclosed when strict criteria were applied to determine disease and positive findings were recorded only in cases of absolute certainty. Applying less stringent criteria risks obscuring any relative difference between groups by frequent reporting of false positive findings. The importance of avoiding false positive diagnoses is also affected by the prevalence. The more unusual the disease, the more important it is to maintain strict criteria. In the present study, all observers were instructed to register disease only when absolutely certain and the periapical status was dichotomised (lesion present or not present). The present study demonstrates improved observer reliability after calibration. This method could be applied to explore a possible relationship between periapical pathology and potential systematic differences between two groups of subjects, e.g. patients with myocardial infarction and a group of matched controls [14].

Conclusions

The study confirms that variables such as the number of remaining teeth, the number of root-filled teeth and the DMFT-score can be reliably assessed on PR. There was initially high variability in recording of teeth with periapical lesions. However, after calibration against a reference standard, agreement improved in all observers, resulting in increased reliability in detection of periapical lesions on PRs.

Acknowledgements

This research was possible due to collaboration with PAROKRANK steering committee, Karolinska Institutet, Solna, Sweden.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This particular study was supported by generous grants from European Society of Endodontology as well as funding from Västra Götalandsregionen, Public Dental Health, Sweden and University of Gothenburg, Sweden.

ORCID

Dan Sebring http://orcid.org/0000-0002-9008-0408

References

[1] Ahlqwist M, Halling A, Hollender L. Rotational panoramic radiography in epidemiological studies of dental health. Comparison between panoramic radiographs and intraoral full mouth surveys. Swed Dent J. 1986;10(1-2):73–84.
[2] Molander B, Ahlqwist M, Gröndahl HG, et al. Comparison of panoramic and intraoral radiography for the diagnosis of caries and periapical pathology. Dentomaxillofac Radiol. 1993;22(1):28–32.
[3] Akkaya N, Kansu O, Kansu H, et al. Comparing the accuracy of panoramic and intraoral radiography in the diagnosis of proximal caries. Dentomaxillofac Radiol. 2006;35(3):170–174.
[4] Estrela C, Bueno MR, Leles CR, et al. Accuracy of cone beam computed tomography and panoramic and periapical radiography for detection of apical periodontitis. J Endod. 2008;34(3):273–279.
[5] Nardi C, Calistri L, Pradella S, et al. Accuracy of Orthopantomography for apical periodontitis without endodontic treatment. J Endod. 2017;43(10):1640–1646.
[6] Goldman M, Pearson AH, Darzenta N. Endodontic success–who's reading the radiograph? Oral Surg Oral Med Oral Pathol. 1972; 33(3):432–437.
[7] Eckerbom M, Andersson JE, Magnusson T. Interobserver variation in radiographic examination of endodontic variables. Endod Dent Traumatol. 1986;2(6):243–246.
[8] Reit C, Gröndahl HG. Application of statistical decision theory to radiographic diagnosis of endodontically treated teeth. Scand J Dent Res. 1983;91(3):213–218.
[9] Tewary S, Luzzo J, Hartwell G. Endodontic radiography: who is reading the digital radiograph? J Endod. 2011;37(7):919–921.
[10] Parker JM, Mol A, Rivera EM, et al. Cone-beam computed tomography uses in clinical endodontics: observer variability in detecting periapical lesions. J Endod. 2017;43(2):184–187.
[11] Reit C. The influence of observer calibration on radiographic periapical diagnosis. Int Endod J. 1987;20(2):75–81.
[12] Saunders MB, Gulabivala K, Holt R, et al. Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study. Int Endod J. 2000;33(3):272–278.
[13] Orstavik D, Kerekes K, Eriksen HM. The periapical index: a scoring system for radiographic assessment of apical periodontitis. Endod Dent Traumatol. 1986;2(1):20–34.
[14] Ryden L, Buhlin K, Ekstrand E, et al. Periodontitis increases the risk of a first myocardial infarction: a report from the PAROKRANK study. Circulation. 2016;133:576–583.
[15] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174.

[16] Gröndahl HG. Some factors influencing observer performance in radiographic caries diagnosis. Swed Dent J. 1979; 3(5):157–172.
[17] Patel N, Rushton VE, Macfarlane TV, et al. The influence of viewing conditions on radiological diagnosis of periapical inflammation. Br Dent J. 2000;189(1):40–42.
[18] Granlund C, Thilander-Klang A, Ylhan B, et al. Absorbed organ and effective doses from digital intra-oral and panoramic radiography applying the ICRP 103 recommendations for effective dose estimations. Br J Radiol. 2016;89(1066): 20151052.
[19] Ludlow JB, Davies-Ludlow LE, White SC. Patient risk related to common dental radiographic examinations: the impact of 2007 International Commission on Radiological Protection recommendations regarding dose calculation. J Am Dent Assoc. 2008;139(9):1237–1243.
[20] Patel S, Brown J, Pimentel T, et al. Cone beam computed tomography in endodontics – a review of the literature. Int Endod J. 2019;52(8):1138–1152.
[21] Kruse C, Spin-Neto R, Evar Kraft DC, et al. Diagnostic accuracy of cone beam computed tomography used for assessment of apical periodontitis: an ex vivo histopathological study on human cadavers. Int Endod J. 2019;52(4):439–450.