2006

Accuracy and Consistency of Radiographic Interpretation Among Clinical Instructors in Conjunction with a Training Program

Sharon K. Lanning
Virginia Commonwealth University, sklanning@vcu.edu

Al M. Best
Virginia Commonwealth University, albest@vcu.edu

Henry J. Temple
University of Michigan

See next page for additional authors

Reprinted by permission of Journal of Dental Education, Volume 70, 5 (May 2006).

Authors
Sharon K. Lanning, Al M. Best, Henry J. Temple, Philip S. Richards, Alllison Carey, and Laurie K. McCauley

Accuracy and Consistency of Radiographic Interpretation Among Clinical Instructors in Conjunction with a Training Program

Sharon K. Lanning, D.D.S.; Al M. Best, Ph.D.; Henry J. Temple, D.D.S.; Philip S. Richards, D.D.S., M.S.; Allison Carey, B.S., M.P.H.; Laurie K. McCauley, D.D.S., Ph.D.

Abstract: There are inaccuracies and inconsistencies of radiographic interpretation among clinical instructors. The purpose of this investigation was to determine if a training program could improve the accuracy and consistency of instructors' ratings of bone loss. A total of thirty-five clinical instructors consisting of periodontal faculty (periodontists and general dentists), dental hygiene faculty, and periodontal graduate students viewed projected digitized radiographic images and quantified bone loss for twenty-five teeth into four descriptive categories. Ratings of bone loss were made immediately before (pretest) and after (post-test 1) initiation of the training program and then again three months later (post-test 2). Ratings were compared to the correct choice categories as determined by direct measurement using the Schei ruler. Overall agreement with the correct choice improved over time (from 64.5 percent to 85.2 percent) with the greatest change from pretest (64.5 percent) to post-test 1 (76.5 percent). Mean and absolute differences improved in three of the four categories, but worsened in one from pretest to post-test 1. This category returned to its original high value at post-test 2. The greatest improvement in consistency among instructors' ratings was seen in one of the four categories, which was "none" (no bone loss). Extension of the training program may further enhance the accuracy and consistency of instructors' radiographic interpretation.

Dr. Lanning is Assistant Professor, Department of Periodontics, Virginia Commonwealth University School of Dentistry and formerly Clinical Assistant Professor, University of Michigan School of Dentistry; Dr. Best is Associate Professor, Department of Biostatistics, Virginia Commonwealth University; Dr. Temple is Clinical Instructor, Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry; Dr. Richards is Clinical Associate Professor, Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry; Ms. Carey is Research Assistant, Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry; and Dr. McCauley is Professor and Chair, Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry. Direct correspondence and requests for reprints to Dr. Sharon Lanning, Virginia Commonwealth University School of Dentistry, 520 North 11th Street, P.O. Box 980566, Richmond, VA 23298-0566; 804-828-4867 phone; 804-828-0657 fax; sklanning@vcu.edu.

Key words: radiographic interpretation, periodontology, faculty development, dental faculty, dental hygiene faculty, educational research, student assessment

Submitted for publication 11/7/05; accepted 1/24/06

Dental school faculty vary in the interpretation of diagnostic tests. Lewis et al.[1] reported low agreement among dentists in evaluating study casts for occlusal stability and tissue loss in malocclusion cases. Disagreement among dental faculty in judging gold-plated tooth models for severity of bruxism has been described.[2] Mileman et al.[3] found considerable variation among clinical instructors in assessing bitewing radiographs for the presence and depth of interproximal carious lesions. Shetty et al.[4] reported variability among oral and maxillofacial surgeons in judging the severity of mandibular fractures using extraoral radiographs. Low agreement among dentists has been noted in evaluating radiographic alveolar bone levels at implant fixtures.[5]

Previous work revealed inaccuracy and variability among periodontal and preventive faculty in rating radiographic bone loss.[6,7] Radiographic findings are important adjuncts to clinical examinations in establishing periodontal diagnosis, prognosis, and long-term evaluation of the periodontium.[8] The position of the alveolar bone crest and its relationship to the tooth's cemento-enamel junction (CEJ) and apex can be used to determine the linear degree of interproximal bone loss.[9] The percent of alveolar bone loss, in conjunction with clinical parameters, is commonly used to determine the presence, degree, and extent of periodontitis.[10]

Inaccurate and inconsistent assessment of percent bone loss among clinical instructors is particularly problematic in an academic environment. Mul-

tiple instructors commonly oversee the diagnosis and treatment of dental school patients. Varied or inaccurate assessment of radiographs could lead to misdiagnosis, over- or undertreatment, or inadequate longitudinal evaluation of patients' periodontal conditions. Additionally, clinical instructors are responsible for teaching and assessing students' abilities to interpret radiographic findings. Clinical instructors' inaccurate and inconsistent evaluations of radiographs may be detrimental to student learning, assessment of student performance, and teaching effectiveness.[11]

A structured training program may improve accuracy and consistency among clinical instructors' ratings of percent bone loss. In this investigation, existing plain film radiographs meeting specific criteria were digitized and displayed by LCD projector. The use of a single method for projecting digitized images offered the advantage of standardized image projection for training large groups. The purpose of this investigation was to determine the accuracy and consistency of clinical instructors' ratings of percent bone loss for a series of digitized intraoral radiographic images in conjunction with a structured training program.

Methods

After obtaining approval from the university's Institutional Review Board, twenty-five digital radiographic images were obtained for the purposes of this study.[7] Radiographs had distinct enamel caps and pulpal chambers, molar cusps with little or no occlusal surface showing, and interproximal contacts free of overlap. Additionally, the CEJ and apex or apices of the study teeth were clearly visible. Radiographs were duplicated using Kodak duplicating film and processor Rp X-OMAT Model M7B with EK Developer Solution and SUREX RP Fixer. Twenty-five radiographs of acceptable quality were randomly selected. Forty percent of the test teeth were anterior and 60.0 percent were posterior, 44.0 percent were maxillary and 56.0 percent were mandibular, and 36.0 percent were single-rooted and 64.0 percent were multirooted.

Radiographs were prepared for projection by scanning them using a flatbed Microtek ScanMaker 8700 scanner and software ScanWizard Pro 7.0, which used a scanning resolution of 300 pixels per inch. Digitized images were imported into Microsoft PowerPoint and projected via LCD projector using a resolution of 1024 x 768 in a dimly lit room. Two of the authors (SKL and HJT) judged the digitized radiographic images to be of acceptable quality after minor grey scale adjustments.

The "actual" amount of bone loss was determined independently by three of the authors (SKL, HJT, and PSR), as described previously.[7] These authors viewed the duplicated plain film radiographs on standard view box separately, without consultation with one another, in an artificially lit room using a Schei ruler to the nearest 5 percent.[12] The Schei ruler used was a plastic transparent ruler with a 2 mm thick marking at its margin and a series of equidistant lines radiating from a center point each representing 5 percent bone loss. The 2 mm thick marking was placed on the tooth's CEJ, and one of the radiating lines was placed on the tooth's apex or most apically positioned apices. The "actual" amount of bone loss was determined by identifying the position of the alveolar bone crest relative to the ruler's markings. One discrepancy in rating bone loss occurred among the authors and was discussed until consensus was reached. Twenty-four percent of test teeth had no bone loss, 24 percent had <15 percent bone loss, 28 percent had between 15 and 30 percent bone loss, and 24 percent had >30 percent bone loss. Two of the authors (SKL and HJT) verified the correct choice categories, using the LCD projector and a computer-generated grid that was superimposed on study teeth.

Clinical instructors from the University of Michigan School of Dentistry including full- and part-time dental hygiene faculty, periodontal faculty (periodontists and general dentists), and periodontal graduate students were recruited into this investigation. These faculty members and graduate students will be collectively referred to as "clinical instructors." Clinical instructors simultaneously completed a twenty-seven-item pretest (referred to as pretest 1) immediately prior to a training program on radiographic interpretation (Figure 1). Question 1 asked clinical instructors to identify themselves as a dental hygiene faculty member, graduate student, or periodontal faculty member. Question 2 asked clinical instructors to describe their years of clinical experience as <5, 5-10, or >10 years. Questions 3-27 asked clinical instructors to rate percent bone loss for indicated teeth while simultaneously viewing magnified digitized radiographic images using an LCD projector by selecting one of the following categories: none, <15 percent, 15-30 percent, and >30 percent. Choices were based on American Dental Association (ADA) and American Academy of Periodontology (AAP)[13-15] guidelines as outlined in the school's clinic manual

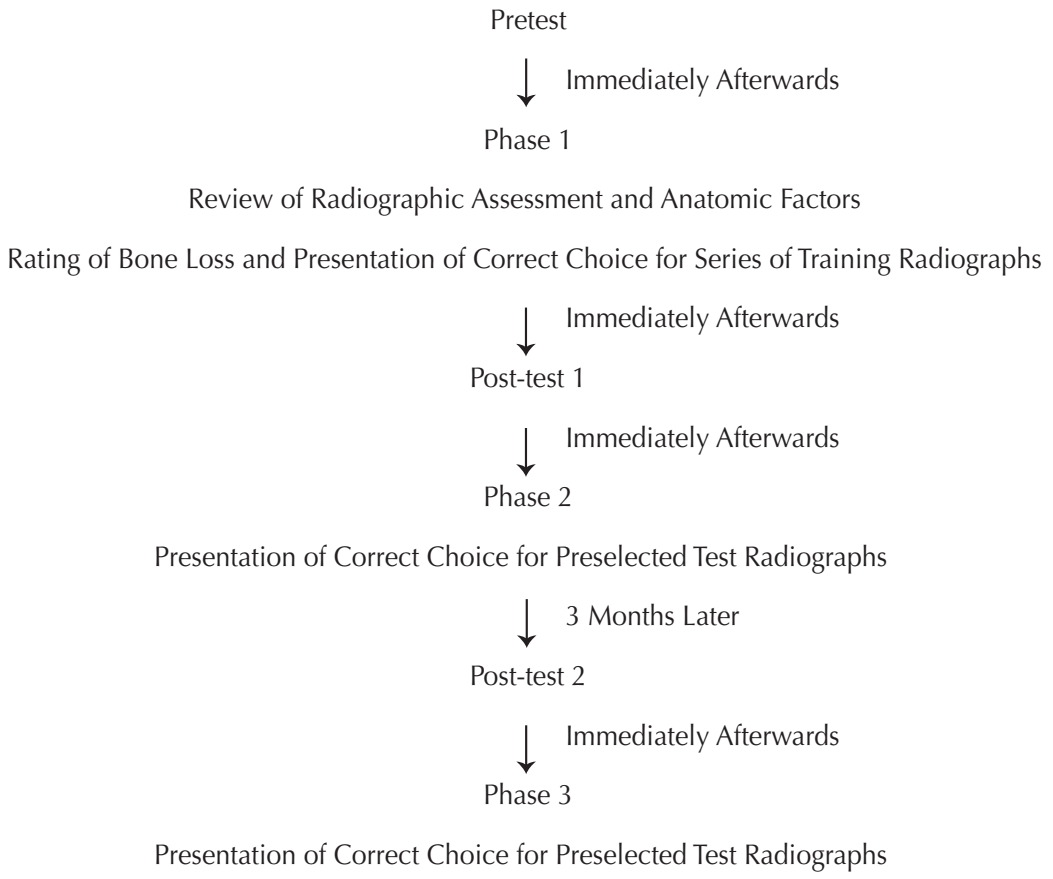

Figure 1. Timeline of training program

for gingivitis and mild, moderate, and severe periodontitis, respectively. For the purpose of statistical analysis, numbers were assigned to each bone loss category as follows: none=(1), <15 percent=(2), 15-30 percent=(3), and >30 percent=(4). Written and verbal instructions were given to ensure consistent viewing practices among clinical instructors. Specifically, they were asked to rate percent bone loss 2 mm apical from the CEJ to the root apex, and teeth with mesial and distal percent bone loss discrepancies were to be rated by the greater percentage of the two. For each question, clinical instructors were given at least thirty seconds to rate percent bone loss, record their response on the questionnaire, and transmit their response via wireless remote. The wireless remote was part of an audience response system (ARS) that allowed "real-time" display of responses during phase one of the training program. However, during the pretest and post-tests it was used for data collection only. Discrepancies between written and transmitted responses were omitted from the research database.

The clinical instructors then participated in the first phase of the training program where one author (PSR) led a one-hour interactive session in which radiographic technique and key anatomical factors of periodontal radiographic interpretation were reviewed. Clinical instructors then viewed six projected training radiographs and rated percent bone loss for indicated teeth, as described above. A graph was prepared showing the percent of clinical instructors' ratings for each bone loss category per tooth using the ARS software and presented to the group. Addition-

ally, a computer-generated grid representing <15 percent, 15-30 percent, and >30 percent bone loss was superimposed on each tooth to display the correct choice category (Figure 2). Anatomical factors were identified, and accuracy and consistency of ratings among clinical instructors were discussed for each training radiograph.

Clinical instructors completed a post-test (referred to as post-test 1) immediately following the first phase of the training program. The twenty-five digitized radiographic images used in the pretest were randomly projected, and clinical instructors rated percent bone loss as before. Immediately following post-test 1, the second phase of the training program began where five preselected digitized radiographic images of the twenty-five used in the pretest and post-tests were reviewed with the clinical instructors. Radiographs were selected since they represented bone loss in each category with two examples coming from category 15-30 percent bone loss. Once again, a graph displaying the clinical instructors' ratings and a grid representing bone loss categories were generated for each test tooth and shared with the clinical instructors. Trends in radiographic interpretation, accuracy, and consistency of ratings among clinical instructors were discussed. The pretest, phases one and two of the training program, and post-test 1 were conducted on the same day during a two-hour session. Three months after this session, clinical instructors completed a second post-test (referred to as post-test 2) using the same twenty-five digitized radiographic images and rated percent bone loss as described above. Immediately afterwards, the third phase of the training program began when five preselected digitized radiographic images of the twenty-five used in the pretest and post-tests (but not used in phase two of the program) were reviewed with the clinical instructors. Again, graphs were presented showing a summary of clinical instructors' ratings, and a grid was superimposed over each test tooth displaying the correct choice category. A discussion regarding the accuracy and consistency of ratings among clinical instructors took place.

Data collected from the pretest and post-tests were analyzed for accuracy and consistency among clinical instructors. Sensitivity and specificity are usually used as indices of accuracy, yet they are not defined in situations with more than two categories.

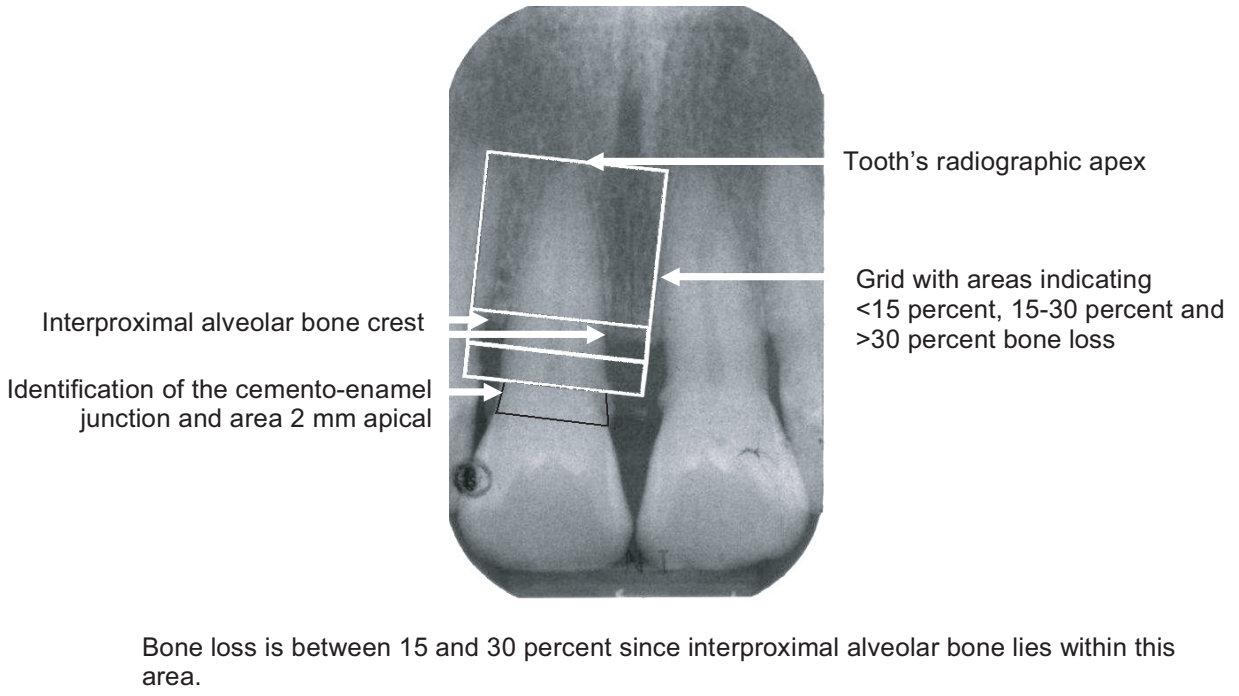

Figure 2. Grid representing varying degrees of alveolar bone loss

Therefore, Kappa coefficient described both agreements between the three occasions (pretest, post-test 1, and post-test 2) and accuracy defined as agreement with the correct choice. Accuracy was also measured by differences from the correct choice in two ways. One dependent variable was the difference between the clinical instructors' ratings and the correct choice; this variable is indicated as "difference" in all tables. This difference is thus the signed rater error and reflects net deviation from the correct choice in one direction. A positive difference indicates an overestimation of bone loss, and a negative difference indicates underestimation of bone loss. The second dependent variable used in the final analysis was the absolute value of this difference. A zero indicates a correct choice, and a positive value reflects overall deviation from the correct choice in either direction. This variable is indicated as "absolute" in all tables. Both the arithmetic difference and absolute difference are necessary because there may be zero average difference while the absolute difference is non-zero, and if there is non-zero absolute difference, it is necessary to describe the direction of the difference. Disagreement was analyzed using repeated-measured, mixed-models analysis with the following independent variables in the ANOVA model: three clinical instructor groups, four correct choice categories, twenty-five radiographs, three occasions, and all possible two-way interactions of these effects. These analyses allowed for dependency of the ratings done by the same clinical instructor across both the multiple radiographs and the three occasions.

Accurate ratings are consistent since they all center on the correct choice. Where ratings are not accurate, they may be consistent—centering around an inaccurate value with little variability—or they may be inconsistent—varying widely. Consistency is thus measured by the standard deviation (SD) of the ratings (square root of the squared difference between the ratings minus the mean of all the ratings provided). To look for differences in consistency, a mixed-model, heterogeneous-variance analysis tested for standard deviation differences between the three clinical instructor groups, the four correct choice categories, and the three occasions.

Results

Thirty-five clinical instructors completed the pretest. The instructors were six dental hygiene faculty members, sixteen graduate students, and thirteen periodontal faculty members (Table 1). All of the dental hygiene faculty and most of the periodontal faculty had ten or more years of clinical experience whereas all but one of the graduate students had less than five years of clinical experience. Discrepancies were noted between written and transmitted responses for 1.8 percent of ratings; these ratings were omitted from the database. The upper panel of Table 2 presents rated bone loss for each correct choice category. For teeth with no bone loss, 63.3 percent (131/207) of the clinical instructors' ratings were accurate. Fifty-five percent, 48.8 percent, and 94.1 percent of the clinical instructor's ratings were accurate for categories <15 percent, 15-30 percent, and >30 percent bone loss, respectively. Overall, clinical instructors' agreement with the correct choice was 64.5 percent. When corrected for chance agreement, this agreement was Kappa=52.7 percent (SE=2.2 percent).

Table 1. Number of clinical instructors for each of the three occasions

Group	Number of Raters		
	Pretest	Post 1	Post 2
Dental Hygiene Faculty	6	4	3
Graduate Students	16	8	5
Periodontal Faculty	13	10	9
Total	35	22	17

Twenty-two clinical instructors completed post-test 1. The instructors consisted of four dental hygiene faculty members, eight graduate students, and ten periodontal faculty members (Table 1). There was no change in years of clinical experience for the instructors who completed the pretest as compared to those who completed post-test 1. Discrepancies were noted between written and transmitted responses for 1.6 percent of ratings; these ratings were omitted from the database. Rated bone loss and comparisons to correct choice are shown in the middle panel of Table 2. For teeth with no bone loss, 93.2 percent (123/132) of the clinical instructors' ratings were accurate. Seventy-five percent, 60.1 percent, and 80.8 percent of the clinical instructors' ratings were accurate for categories <15 percent, 15-30 percent, and >30 percent bone loss, respectively. Overall, clinical instructors' agreement with the correct choice was 76.5 percent. When corrected for chance agreement, this agreement was Kappa=68.7 percent (SE=2.4 percent).

Table 2. Accuracy of the rating across the three occasions

Comparing the pretest rating to the correct choice	Correct Choice				
Pretest	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	131	9	0	0	140
Less than 15% (2)	71	112	42	1	226
Between 15 and 30% (3)	4	77	117	11	209
Greater than 30% (4)	1	6	81	190	278
Total	207	204	240	202	853
Accuracy=	0.633	0.549	0.488	0.941	0.645
(SE)=	(0.034)	(0.035)	(0.032)	(0.017)	(0.016)
			Kappa=0.527		(SE 0.022)

Comparing the first post-test rating to the correct choice	Correct Choice				
Post-test 1	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	123	16	2	0	141
Less than 15% (2)	9	97	37	2	145
Between 15 and 30% (3)	0	16	92	23	131
Greater than 30% (4)	0	1	22	105	128
Total	132	130	153	130	545
Accuracy=	0.932	0.746	0.601	0.808	0.765
(SE)=	(0.022)	(0.038)	(0.040)	(0.035)	(0.018)
			Kappa=0.687		(SE 0.024)

Comparing the second post-test rating to the correct choice	Correct Choice				
Post-test 2	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	94	6	0	0	100
Less than 15% (2)	8	83	20	0	111
Between 15 and 30% (3)	0	12	89	10	111
Greater than 30% (4)	0	0	6	92	98
Total	102	101	115	102	420
Accuracy=	0.922	0.822	0.774	0.902	0.852
(SE)=	(0.027)	(0.038)	(0.039)	(0.029)	(0.017)
			Kappa=0.803		(SE 0.023)

Seventeen clinical instructors completed post-test 2. The instructors were three dental hygiene faculty members, five graduate students, and nine periodontal faculty members (Table 1). There was no change in years of clinical experience for the instructors who completed the pretest as compared to those who completed post-test 2. Discrepancies were noted between written and transmitted responses for 1.3 percent of ratings; these ratings were omitted from the database. For teeth with no bone loss, 92.2 percent (94/102) of the clinical instructors' ratings were accurate. Eighty-two percent, 77.4 percent, and 90.2 percent of the clinical instructors' ratings were accurate for categories <15 percent, 15-30 percent, and >30 percent bone loss, respectively. Overall, clinical instructors' agreement with the correct choice was 85.2 percent. When corrected for chance agreement, this agreement was Kappa=80.3 percent (SE=2.3 percent).

Twenty-two clinical instructors provided ratings for both the pretest and post-test 1. The twenty-two clinical instructors consisted of four dental hygiene faculty members, eight graduate students, and ten periodontal faculty members. Their ratings were directly compared, and agreement was 67.3 percent (Kappa=56.5 percent, SE=2.7 percent) (Table 3, upper panel). Seventeen clinical instructors provided ratings during both post-tests 1 and 2. The seventeen clinical instructors consisted of three dental hygiene faculty members, five graduate students, and nine pe-

riodontal faculty members. Their ratings were directly compared, and agreement was 76.7 percent (Kappa=68.9 percent, SE=2.8 percent) (Table 3, middle panel). Seventeen clinical instructors provided ratings during the pretest and post-test 2. The seventeen clinical instructors were three dental hygiene faculty members, five graduate students, and nine periodontal faculty members. Their ratings were directly compared, and agreement was 67.8 percent (Kappa=57.1 percent, SE=3.1 percent) (Table 3, bottom panel). As accuracy improved from pretest to post-test 1 (Kappa=52.7 percent to 68.7 percent), agreement between these two occasions was relatively low (67.3 percent). Subsequently, as accuracy improved slightly from post-test 1 to post-test 2 (Kappa=68.7 percent to 78.7 percent), agreement between these two occasions was higher (76.7 percent).

Overall, the mean difference and absolute difference of the rated response compared with the correct choice were similar. For both measures, the difference between the rated choice and the correct choice was different between the three occasions (p=0.0001), but the amount of change was not consistent across the four correct choice categories (p=0.0001). For the mean difference, there was significant change from the pretest to post-test 1, but the amount of change differed between the four correct choice categories. That is, within categories none and <15 percent bone loss, differences changed from approximately 0.4 to approximately 0.06 (Table 4). Within category 15-30 percent bone loss, the difference changed from approximately 0.2 to approximately –0.1. Lastly, within category >30 percent bone loss, the change was from a near-zero difference of –0.05 towards a larger difference of –0.2. Similar changes from the pretest to post-test 1 were also observed for the absolute difference (Table 4, absolute difference column). The absolute difference significantly decreased in correct choice categories none, <15 percent, and 15-30 percent bone loss, but significantly increased in category >30 percent bone loss. In this category, the difference and absolute difference approached baseline values at the second post-test. There were no significant changes in the mean and absolute differences between post-test 1 and post-test 2 in any of the correct choice catego-

Table 3. Agreement of the ratings between the three occasions

			Pretest		
Post-test 1	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	89	49	2	0	140
Less than 15% (2)	2	82	53	5	142
Between 15 and 30% (3)	0	14	70	44	128
Greater than 30% (4)	0	0	5	117	122
Total	91	145	130	166	532
	Agreement=0.673		Kappa=0.565		(SE 0.027)

			Post-test 1		
Post-test 2	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	92	7	0	0	99
Less than 15% (2)	12	83	15	1	111
Between 15 and 30% (3)	1	27	68	14	110
Greater than 30% (4)	0	1	19	76	96
Total	105	118	102	91	416
	Agreement=0.767		Kappa=0.689		(SE 0.028)

			Post-test 2		
Pretest	None (1)	Less than 15% (2)	Between 15 and 30% (3)	Greater than 30% (4)	Total
None (1)	131	9	0	0	140
None (1)	66	3	0	0	69
Less than 15% (2)	31	65	13	1	110
Between 15 and 30% (3)	2	38	60	6	106
Greater than 30% (4)	0	2	35	85	122
Total	99	108	108	92	407
	Agreement=0.678		Kappa=0.571		(SE 0.031)

Table 4. Mean rater error for each clinical instructor group, correct choice category, and occasion

Correct Choice[1]	Occasion	n	Disagreement with Correct Choice: Difference[2] Mean	SD	95% CI	Absolute[3] Mean	SD	95% CI
Dental Hygiene Faculty								
1	Pre test	36	0.39	0.49	(0.23, 0.55)	0.39	0.49	(0.23, 0.55)
	Post test 1	24	0.04	0.20	(-0.04, 0.12)	0.04	0.20	(0, 0.12)
	Post test 2	18	0.06	0.24	(-0.05, 0.16)	0.06	0.24	(0, 0.16)
2	Pre test	36	0.64	0.68	(0.42, 0.86)	0.64	0.68	(0.42, 0.86)
	Post test 1	24	0.17	0.48	(-0.03, 0.36)	0.25	0.44	(0.07, 0.43)
	Post test 2	18	0.11	0.47	(-0.11, 0.33)	0.22	0.43	(0.02, 0.42)
3	Pre test	41	0.39	0.59	(0.21, 0.57)	0.49	0.51	(0.33, 0.64)
	Post test 1	28	0.00	0.67	(-0.25, 0.25)	0.36	0.56	(0.15, 0.56)
	Post test 2	20	-0.10	0.31	(-0.23, 0.03)	0.10	0.31	(0, 0.23)
4	Pre test	36	0.00	0.00	NA	0.00	0.00	NA
	Post test 1	24	-0.13	0.45	(-0.30, 0.05)	0.13	0.45	(0, 0.30)
	Post test 2	18	-0.06	0.24	(-0.16, 0.05)	0.06	0.24	(0, 0.16)
all	Pre test	149	0.36	0.56	(0.27, 0.45)	0.38	0.54	(0.30, 0.47)
	Post test 1	100	0.02	0.49	(-0.08, 0.12)	0.20	0.45	(0.11, 0.29)
	Post test 2	74	0.00	0.33	(-0.08, 0.08)	0.11	0.31	(0.04, 0.18)
Graduate Student								
1	Pre test	94	0.45	0.62	(0.32, 0.57)	0.45	0.62	(0.32, 0.57)
	Post test 1	48	0.04	0.20	(-0.02, 0.10)	0.04	0.20	(0, 0.10)
	Post test 2	30	0.03	0.18	(-0.03, 0.10)	0.03	0.18	(0, 0.10)
2	Pre test	91	0.37	0.63	(0.24, 0.50)	0.48	0.54	(0.37, 0.60)
	Post test 1	47	-0.09	0.58	(-0.25, 0.08)	0.34	0.48	(0.20, 0.48)
	Post test 2	29	0.00	0.38	(-0.14, 0.14)	0.14	0.35	(0.01, 0.27)
3	Pre test	109	0.15	0.72	(0.01, 0.28)	0.53	0.50	(0.44, 0.63)
	Post test 1	55	-0.16	0.71	(-0.35, 0.03)	0.49	0.54	(0.35, 0.63)
	Post test 2	34	-0.24	0.55	(-0.42, -0.05)	0.35	0.49	(0.19, 0.52)
4	Pre test	90	-0.10	0.30	(-0.16, -0.04)	0.10	0.30	(0.04, 0.16)
	Post test 1	46	-0.15	0.36	(-0.26, -0.05)	0.15	0.36	(0.05, 0.26)
	Post test 2	30	-0.13	0.35	(-0.26, -0.01)	0.13	0.35	(0.01, 0.26)
all	Pre test	384	0.22	0.63	(0.15, 0.28)	0.40	0.53	(0.35, 0.45)
	Post test 1	196	-0.09	0.52	(-0.16, -0.02)	0.27	0.45	(0.20, 0.33)
	Post test 2	123	-0.09	0.41	(-0.16, -0.02)	0.17	0.38	(0.10, 0.24)
Periodontal Faculty								
1	Pre test	77	0.34	0.50	(0.23, 0.45)	0.34	0.50	(0.23, 0.45)
	Post test 1	60	0.10	0.30	(0.02, 0.18)	0.10	0.30	(0.02, 0.18)
	Post test 2	54	0.11	0.32	(0.03, 0.20)	0.11	0.32	(0.03, 0.20)
2	Pre test	77	0.30	0.56	(0.17, 0.42)	0.40	0.49	(0.29, 0.51)
	Post test 1	59	0.03	0.49	(-0.09, 0.16)	0.20	0.45	(0.09, 0.32)
	Post test 2	54	0.07	0.43	(-0.04, 0.19)	0.19	0.39	(0.08, 0.29)
3	Pre test	90	0.08	0.71	(-0.07, 0.22)	0.50	0.50	(0.40, 0.60)
	Post test 1	70	-0.14	0.60	(-0.28, 0.00)	0.37	0.49	(0.26, 0.49)
	Post test 2	61	-0.07	0.44	(-0.18, 0.05)	0.20	0.40	(0.10, 0.30)
4	Pre test	76	-0.05	0.28	(-0.12, 0.01)	0.05	0.28	(0, 0.12)
	Post test 1	60	-0.28	0.49	(-0.41, -0.16)	0.28	0.49	(0.16, 0.41)
	Post test 2	54	-0.09	0.29	(-0.17, -0.01)	0.09	0.29	(0.01, 0.17)
all	Pre test	320	0.16	0.56	(0.10, 0.22)	0.33	0.48	(0.28, 0.38)
	Post test 1	249	-0.08	0.51	(-0.14, -0.01)	0.24	0.45	(0.19, 0.30)
	Post test 2	223	0.00	0.39	(-0.05, 0.06)	0.15	0.36	(0.10, 0.19)

ries except for a decrease in the absolute difference in correct choice category 15-30 percent bone loss.

There was no evidence of a rater group difference on either the difference or absolute difference ($p>0.5$), but there was no group to correct choice interaction ($p>0.1$). The dental hygiene faculty had the nominally highest difference (least square mean=0.1, SE=0.07, 95 percent CI=–0.04±0.24), the graduate students had the middle difference value (least square mean=0.07, SE=0.05, 95 percent CI=–0.03±0.17), and the periodontal faculty had the smallest difference (least square mean=0.03, SE=0.05, 95 percent

CI=–0.07±0.13). There was some evidence that the amount of change for the difference was not consistent between the three clinical instructor groups over time (p=0.09). That is, the change in the difference for the periodontal faculty group was smaller than for the other two groups during the course of the study.

Using a mixed-model heterogeneous-variance analysis, it was determined that the variability of the difference (clinical instructors' ratings minus the mean of the ratings provided) did not depend upon the three clinical instructor groups, but did depend upon the four correct choice categories and three occasions (LR chi square=215, df=11, p<0.0001). That is, there was more consistency (less variability) for correct choice categories none and >30 percent bone loss across time. However, this trend was not observed in the middle two categories (p<0.0001). That is, within both categories <15 percent and 15-30 percent bone loss, consistency of clinical instructors' responses remained unchanged across time (typical SD was approximately 0.40). Whereas within category >30 percent bone loss, consistency decreased from pretest (typical SD=0.16) to post-test 1 (typical SD=0.26) and then increased at post-test 2 (typical SD=0.18). The predominant increase in consistency was in correct choice category none, where the SD decreased from 0.34 at the pretest to 0.14 at both post-test occasions.

Overestimation of bone loss occurred during the pre-test more often than underestimation as indicated by positive mean differences for categories none, <15 percent, and 15-30 percent bone loss (Table 4, difference column). In the category <15 percent, 37.2 percent of clinical instructors' ratings were given as 15-30 percent bone loss, and only 4.2 percent were given as no bone loss. Similarly, in category 15-30 percent, 34.2 percent of clinical instructors' ratings were given as >30 percent bone loss, and only 17.3 percent were given as <15 percent bone loss. From the pretest to post-test 1, accuracy of ratings in categories <15 percent and 15-30 percent increased, and overestimation of bone loss decreased by half. There was an increase in underestimation of bone loss (decrease in accuracy) in category >30 percent between the pretest and post-test 1, but by the second post-test, accuracy had returned to its original high level. The increase in accuracy from post-test 1 to post-test 2 is particularly evident in categories <15 percent and 15-30 percent, where underestimation and overestimation of bone loss decreased, respectively.

Discussion

Inaccuracy and inconsistency among clinicians have been well documented in both medicine and dentistry.[6,7,16-28] Attempts to train or calibrate clinicians to enhance accuracy and/or inter-rater agreement have shown mixed outcomes. Roy et al.[26] demonstrated that a computerized self-instructional program increased family physicians' cardiac auscultation skills. Dahlstrom et al.[27] reported that a training program intended to calibrate examiners of temporomandibular disorders (TMD) resulted in an increase in recognizing signs of TMD; however, it was not sufficient to create reliability among multiple examiners. Robertello et al.[28] found that their brief training program improved the reliability of examiners' assessments of the clinical serviceability of amalgam restorations, although the authors noted that the "gain was not to the level commonly accepted by the literature."

Our results show clinical instructors' agreement with the correct choice overall improved with time. The greatest improvement was seen immediately after the first phase of the training program, yet accuracy continued to get better from post-test 1 to post-test 2. The mean difference and absolute difference improved in categories none, <15 percent, and 15-30 percent bone loss, yet worsened in category >30 percent bone loss immediately after the first phase of the training program. In this category, the difference and absolute difference improved from post-test 1 to post-test 2. Additionally, consistency of clinical instructors' responses initially decreased and then increased in category >30 percent bone loss. That is, the accuracy and consistency of ratings worsened immediately following phase one of the training program. Participation in this component of the training program may have been detrimental to clinical instructors' ability to judge bone loss >30 percent. Improvement in accuracy and consistency among clinical instructors' ratings was noted from post-test 1 to post-test 2 as agreement with the correct choice approached its initial high value. It may be that stressing the underestimation of bone loss in this category during the second phase of the program addressed any weakness of the training program. It is also possible that clinical instructors went back to judging severe bone loss in the manner they were accustomed to before participating in the program. Furthermore, it may be that the decrease in the initially high accu-

racy and consistency among clinical instructors was due to regression towards the mean.

The amount of error varied between the four bone loss categories. The greatest improvement of accuracy and consistency among clinical instructors' ratings occurred in correct choice category none. Greater inaccuracies and inconsistencies are not unexpected in categories <15 percent and 15-30 percent bone loss since errors can occur on both sides of these middle categories. Although, as suggested previously, it may be that bone loss of <15 percent and 15-30 percent was more difficult to assess than none or >30 percent or teeth, and the actual amounts of bone loss selected for this study could have contributed to greater errors observed in these two categories.[7]

Previous work found periodontal faculty members to have significantly less error than dental hygiene faculty members in categories <15 and 15-30 percent bone loss.[7] There is some evidence that the amount of change in rater error, as a result of the training program, was not consistent for the three clinical instructor groups. Since the periodontal faculty began with nominally more accurate ratings, the amount of improvement possible was smaller than the other two groups. It is not unexpected that dental hygiene faculty members' accuracy rates were initially lower than the other two groups since they are not diagnosticians nor do they routinely perform in-depth clinical assessments on a vast array of periodontal patients. In general, rater error could occur due to poor digitized radiographic image quality, use of a projector for displaying these magnified images, indistinguishable or difficulty in recognizing anatomical landmarks, or rating bone loss from a distance less than or greater than 2 mm apical from the CEJ as elaborated on earlier.[7] Rater error could have persisted throughout the duration of this study for any of these reasons or may be a result of clinical instructors holding onto strongly held beliefs[29] or a reflection of the training program's effectiveness and duration. Our results show an improvement in the difference and absolute difference between the three occasions. It may be that extending the program and concentrating on areas where errors persist could further improve accuracy and consistency of clinical instructors' responses.

Overestimation of radiographic bone loss has been reported previously where the "gold standard" for which clinicians' ratings were compared was direct surgical or Schei rule measurements.[7,30-32] Immediately after phase one of the training program, overestimation of percent bone loss decreased by half in categories <15 percent and 15-30 percent bone loss, resulting in an improvement in accuracy. However, in category >30 percent there was an increase in underestimation of bone loss, resulting in a decrease in accuracy. At the second post-test, there was less underestimation of bone loss, and the accuracy and consistency of clinical instructors' responses returned to their originally high values.

The percent of alveolar bone loss is an important component in establishing a diagnosis of periodontitis and managing the disease over time.[10] Categories of bone loss used in this investigation (none, <15 percent, 15-30 percent, and >30 percent) help establish diagnoses of gingivitis and mild, moderate, and severe periodontitis, respectively. These categories make clinicians aware of and sensitive to all diagnostic findings and treatment needs. For example, progression of bone loss from 15 to 30 percent carries with it the potential for more complex treatment and/or potential specialty referral in order to achieve therapeutic goals. Accurate and consistent radiographic interpretation coupled with clinical findings is essential for establishing initial periodontal diagnosis and long-term follow-up of a patient.[33] In a dental school setting, where multiple instructors participate in the care of a single patient, inaccurate and inconsistent ratings of percent bone loss could be particularly problematic. That is, differences among clinical instructors could lead to a variety of periodontal diagnoses, prognoses, and treatment recommendations, which ultimately could result in over- or undertreatment. Inaccuracies and inconsistency among clinical instructors may also influence students' abilities to correctly rate radiographic bone loss or relate these findings to clinical findings, which are needed to adequately diagnosis and manage periodontal patients. Furthermore, variations among clinical instructors could negatively influence assessment of student performance and teaching effectiveness. Clinical instructors in most educational programs are considered content experts and evaluate students based on their ability to generate an answer consistent with theirs. If the said experts' opinions are different on different occasions, then the ability to reliably assess student performance and evaluate teaching programs is lost.

Clinicians' ratings of radiographic bone loss should ideally be consistently accurate; however, this goal was not reached during the course of this study. This must be taken into consideration when teaching and assessing students' abilities to judge percent bone loss. It may be that the best way to ensure that

students and clinical instructors alike are consistently rating percent bone loss accurately is to use the Schei ruler to verify the actual amount of percent bone loss especially when it is thought to be between <15 percent and 30 percent or when the amount of bone loss is in question. The Schei ruler has been found to be accurate in determining bone loss as compared to surgical measurement, and it is efficient and easy to use.[12] Additionally, computer-assisted radiography has been shown to improve the accuracy of detecting changes in alveolar bone.[34-36] This technology could be an asset in the teaching and learning of radiographic assessment in dental and dental hygiene education. Unfortunately, it is not available in all dental schools or clinical practices.

The training program was designed to be relatively brief and ongoing and to provide immediate feedback to clinical instructors on their assessment of radiographic bone loss. This was made possible by utilizing a single projection system, the ARS, for displaying responses in "real-time" during phase one of the training program and reviewing test radiographs immediately after each post-test. A second post-test was administered to determine clinical instructors' recall of information some time after the initiation of the training program. The three-month interval was thought to be appropriate in testing clinical instructors' recall of information, and it was most convenient based on their other professional obligations. The improvement in accuracy and consistency among clinical instructors seen immediately following phase one of the program was seen again after three months. Therefore, the skills that clinical instructors gained as part of the program were sustained over time. It is important to note that the first phase of the program may have contributed to clinical instructors' inability to correctly judge percent bone loss of >30 percent since accuracy rates and consistency among clinical instructors worsened immediately afterwards. Possible reasons for this have been discussed earlier.

The number of clinical instructors participating in this study decreased over time, and nonparticipation could lead to sampling bias. The difference in pretest and post-test 1 response rates was influenced by the number of clinical instructors eligible to participate in each of these tests. Seven "new" clinical instructors (five graduate students and two periodontal faculty members) joined the department between the third occasion (post-test 2) and the first two occasions (the pretest and post-test 1). Under all testing conditions, clinical instructors viewed and rated digitized radiographic images and responded to test questions simultaneously yet independently, without consulting with one another. Since these "new" clinical instructors were tested under the conditions just described and had not previously viewed the radiographic images nor participated in the training program, their responses were incorporated into the pretest data set. Sessions were offered once, and scheduling conflicts could have prevented clinical instructors from participating in a session or in a session's entirety given their other teaching, research, and clinical responsibilities. Differences in the number of clinical instructors participating in the training program are likely a result of scheduling conflicts. However, changes in response rates could be a reflection of clinical instructors' beliefs that the program was redundant or not useful or that their individual accuracy was adequate and participation in the program was no longer needed. Providing an opportunity for clinical instructors to critique the training program may have provided insight into further reasons for nonparticipation.

This program has other limitations. Digitized radiographic images were scanned using a relatively low resolution and displayed by a fixed-pixel projector. These images compared to plain films likely differed in resolution, contrast, grey-scale manipulation, and magnification. This could have affected image quality and thus impacted the results of this investigation since clinicians' responses were compared to correct choice categories as determined by viewing plain films on a view box. However, it is important to note that two of the authors (SKL and HJT) independently confirmed correct choice categories using the LCD projector prior to the clinical instructors' viewing of digitized radiographic images. Clinical instructors may have discussed radiographs and their ratings of percent bone loss with one another throughout the course of this investigation, which could have influenced the results. It could also be argued that multiple viewings of the same radiographs could have contributed to the increase in accuracy and consistency of clinical instructors' responses reported here. However, radiographs were randomly viewed at each occasion, and three months separated post-test 1 and post-test 2, making it difficult for clinical instructors to base their ratings of bone loss on familiarity of the radiographs alone. It is more likely that the skills the clinical instructors gained as part of this program were applied during these post-tests.

It may be acceptable to have inconsistencies among clinical instructors when there are a number

of subjective elements that go into making clinical decisions as long as the decisions are based on evidence or accepted practice guidelines. Making determinations of bone loss is based on relationships between anatomical landmarks, which can actually be measured. Therefore, determining percent bone loss is less subjective than interpretation of other clinical findings that can not be directly measured, and inconsistencies among clinical instructors are less expected and less acceptable. Further attempts at training and calibrating instructors are needed so that the accuracy and consistency of their ratings can be enhanced and teaching effectiveness and students' abilities can be adequately evaluated. The training program resulted in a general improvement in accuracy for most categories; however, greater improvement in accuracy and consistency among clinical instructors may be possible with extension of the program to include more radiographs. An additional next step would be to determine if the gains in accuracy and consistency of clinical instructors' assessments of percent bone loss could be "transferred" to plain films. Previous work showed rating percent bone loss by viewing projected digitized radiographic images was only slightly different in terms of accuracy and consistency as compared to viewing plain films via view box.[7] Therefore, skills learned as part of this training program should be easily applied to plain film viewing.

Conclusion

The overall agreement with the correct choice improved somewhat as a result of the training program, with the greatest change observed immediately after the first phase. Mean and absolute differences improved in categories none, <15 percent, and 15-30 percent bone loss. For category >30 bone loss, accuracy and consistency among clinical instructors worsened immediately after the first phase of the training program yet returned to their original high values at post-test 2. The greatest improvement in consistency among instructors' ratings was seen in the correct choice category none. Extension of the training program to include more radiographs may further enhance the accuracy and consistency of clinical instructors' radiographic interpretations. Accurate and consistent assessment of radiographs among clinical instructors is necessary for adequate evaluation of patient care, student performance, and teaching effectiveness.

Acknowledgments

The authors would like to thank Mrs. Barbara Wolfgang and Mrs. Beverly Sutton for their administrative assistance, Drs. Roger Hill and Stephen Soehren for assistance with radiographic selection, and Dr. Charles Janus for assistance with manuscript preparation. Additionally, the authors would like to acknowledge the periodontal and preventive clinical faculty and graduate students at the University of Michigan, School of Dentistry Department of Periodontics and Oral Medicine (formerly the Department of Periodontics/Prevention/Geriatrics) for their participation in and dedication to this project.

REFERENCES

1. Lewis EA, Albino JE, Cunat JJ, Tedesco LA. Reliability and validity of clinical assessment of malocclusion. Am J Orthod 1982;81(6):473-7.
2. Marbach JJ, Raphael KG, Janal MN, Hirschkorn-Roth R. Reliability of clinician judgments of bruxism. J Oral Rehabil 2003;30(2):113-8.
3. Mileman PA, Pudell-Lewis DJ, van der Weele LT. Variation in radiographic caries diagnosis and treatment decision among university teachers. Community Dent Oral Epidemiol 1982;10(6):329-44.
4. Shetty V, Atchison K, Der-Martirosian C, Wang J, Belin TR. Determinants of surgical decisions about mandible fractures. J Oral Maxillofac Surg 2003;61(7):808-13.
5. Grondahl K, Suden S, Grondahl HG. Inter- and intraobserver variability in radiographic bone level assessment at Branemark fixtures. Clin Oral Implants Res 1998;9(4):243-50.
6. Lanning SK, Pelok SD, Williams BC, Richards PS, Sarment DP, Oh TJ, et al. Variation in periodontal diagnosis and treatment planning among clinical instructors. J Dent Educ 2005;69(3):325-37.
7. Lanning SK, Best AM, Temple HJ, Richards PS, Carey A, McCauley LK. Accuracy and consistency of radiographic interpretation among clinical instructors using two viewing systems. J Dent Educ 2006;70(2):149-59.
8. Carranza FA, Takei HH. Radiographic aids in the diagnosis of periodontal disease. In: Newman MG, Takei HH, Carranza FA, eds. Carranza's clinical periodontology. Philadelphia: W.B. Saunders Co., 2002:454-68.
9. Albandar JM. Validity and reliability of alveolar bone level measurements made on dry skulls. J Clin Periodontol 1889;16(9):575-9.
10. Greenfield DS, Williams RC, Goldhaber P. Radiographic measurement of alveolar bone loss: a perspective in vitro. J Clin Periodontol 1981;8(6):474-80.
11. Scruggs RR, George MC. Faculty calibration in clinical education: a review of the literature. Educ Dir Dent Hyg 1985;10(4):15-21.
12. Schei O, Waerhaug J, Lovdal A, Arno A. Alveolar bone loss as related to oral hygiene and age. J Periodontol 1959;30:7-16.

13. Perry DA, Beemsterboer P, Taggart EJ. Periodontics for the dental hygienist. 2nd ed. Philadelphia: W.B. Saunders, 2001:108-9, 180.
14. Armitage GC. Development of a classification system for periodontal diseases and conditions. Ann Periodontol 1999;4(1):1-6.
15. American Academy of Periodontology. Parameters of care. J Periodontol 2000;71(5):849-83.
16. Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med 1982;307(10):588-96.
17. Boom R, Gonzalez C, Fridman L, Ayala JF, Realpe JL, Morales P, et al. Looking for "indicants" in the differential diagnosis of jaundice. Med Decis Making 1986; 6(1):36-41.
18. Todd BS, Stamper R. Limits to diagnostic accuracy. Med Inform 1993;18(3):255-70.
19. Mileman PA, Pudell-Lewis DJ, van der Weele LT. Effect of variation in caries diagnosis and degree of caries on treatment decisions by dental teachers using bitewing radiographs. Community Dent Oral Epidemiol 1983; 11(6):356-62.
20. Mileman PA, Purdell-Lewis DJ, Dummer P, van der Weele LT. Diagnosis and treatment decisions when using bitewing radiographs: a comparison between two dental schools. J Dent 1985;13(2):140-51.
21. Espelid I, Tveit AB, Fjelltveit A. Variations among dentists in radiographic detection of occlusal caries. Caries Res 1994;28(3):169-75.
22. Mileman PA, van der Weele LT. Accuracy in radiographic diagnosis: Dutch practitioners and dental caries. J Dent 1996;18(3):130-6.
23. Tongsong T, Iamthogin A, Wanapirak C, Piyamongkol W, Sirichotiyakul S, Boonyanurak P, et al. Accuracy of fetal heart-rate variability interpretation by obstetricians using the criteria of the National Institute of Child Health and Human Development compared with computer-aided interpretation. J Obstet Gynaecol Res 2005;31(1):68-71.
24. Balabanova Y, Coker R, Fedorin I, Zakharova S, Plavinskij S, Krukov N, et al. Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. BMJ 2005;13:379-82.
25. Moore JH, Goss DL, Baxter RE, DeBerardino TM, Mansfield LT, Fellows DW, et al. Clinical diagnostic accuracy and magnetic resonance imaging of patients referred by physical therapists, orthopaedic surgeons, and nonorthopaedic providers. Ortho Sports Phys Ther 2005; 35(2):67-71.
26. Roy D, Sargeant J, Gray J, Hoyt B, Allen M, Fleming M. Helping family physicians improve their cardiac auscultation skills with an interactive CD-ROM. Contin Educ Health Prof 2002;22(3):152-9.
27. Dahlstrom L, Keeling SD, Fricton JR, Galloway Hilsenbeck S, Clark GM, Rugh JD. Evaluation of a training program intended to calibrate examiners of temporomandibular disorders. Acta Odontol Scand 1994; 52(4):250-4.
28. Robertello FJ, Pink FE. The effect of a training program on the reliability of examiners evaluating amalgam restorations. Oper Dent 1997;22(2):57-65.
29. Bader JD, Shugars DA. Agreement among dentists' recommendation for restorative treatment. J Dent Res 1993;72(5):891-6.
30. Khocht A, Janal M, Harasty L, Chang KM. Comparison of direct and conventional intra-oral radiographs in detecting alveolar bone loss. J Am Dent Assoc 2003; 134(11):1468-75.
31. Nair MK, Ludlow JB, Tyndall DA, Platin E, Denton G. Periodontitis detection efficacy of film and digital images. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 1998; 85(5):608-12.
32. Furkart AJ, Dove SB, McDavid WD, Nummikoski P, Matteson P. Direct digital radiography for the detection of periodontal bone lesions. Oral Surg Oral Med Oral Pathol 1992;74(5):652-60.
33. Hull PS, Hillman DG, Beal JF. A radiographic study of the prevalence of chronic periodontitis in 14-year-old schoolchildren. J Clin Periodontol 1975;2(4):203-10.
34. Cury PR, Araujo NS, Bowie J, Sallum EA, Jeffcoat MK. Comparison between subtraction radiography and conventional radiographic interpretation during long-term evaluation of periodontal therapy in Class II furcation defects. J Periodontol 2004;75(8):1145-9.
35. Kullendorff B, Grondahl K, Rohlin M, Nilsson M. Subtraction radiography of interradicular bone lesions. Acta Odontol Scand 1992;50(5):259-67.
36. Chai-U-Dom O, Ludlow JB, Tyndall DA, Webber RL. Comparison of conventional and TACT (Tuned Aperture Computed Tomography) digital subtraction radiography in detection of pericrestal bone-gain. J Periodontal Res 2002;37(2):147-53.