Examiner training and reliability in two randomized clinical trials of adult dental caries

David W. Banting, DDS, MSc, PhD[1]; Bennett T. Amaechi, BDS, MS, PhD[2]; James D. Bader, DDS, MPH[3]; Peter Blanchard, DDS[4]; Gregg H. Gilbert, DDS, MBA[5]; Christina M. Gullion, PhD[6]; Jan Carlton Holland, RDH, MS[3]; Sonia K. Makhija, DDS, MPH[5]; Athena Papas, DMD, PhD[7]; André V. Ritter, DDS, MS[3]; Mabi L. Singh, DMD, MS[7]; William M. Vollmer, PhD[6]

1 Schulich School of Medicine & Dentistry, The University of Western Ontario, London, ON
2 University of Texas Health Science Center, San Antonio, TX
3 University of North Carolina School of Dentistry, Chapel Hill, NC
4 Dental Services of Massachusetts, Boston, MA
5 Department of General Dental Sciences, University of Alabama at Birmingham School of Dentistry, Birmingham, AL
6 Kaiser Permanente Center for Health Research, Portland, OR
7 Tufts University School of Dental Medicine, Boston, MA

Keywords
dental caries; early diagnosis; reproducibility of results; ICDAS-II; clinical trial; examiner training; examiner standardization.

Correspondence
Dr. David Banting, Schulich School of Medicine & Dentistry, The University of Western Ontario, London, ON N6A 5C1. Tel.: 519-661-2111 x86130; Fax: 519-661-3875; e-mail: dbanting@uwo.ca. Bennett T. Amaechi is with the University of Texas Health Science Center. James D. Bader, Jan Carlton Holland, and André V. Ritter are with the University of North Carolina School of Dentistry. Peter Blanchard is with the Dental Services of Massachusetts. Gregg H. Gilbert and Sonia K. Makhija are with the Department of General Dental Sciences, University of Alabama at Birmingham School of Dentistry. Christina M. Gullion and William M. Vollmer are with the Kaiser Permanente Center for Health Research. Athena Papas and Mabi L. Singh are with the Tufts University School of Dental Medicine.

Received: 11/2/2010; accepted: 7/8/2011.

doi: 10.1111/j.1752-7325.2011.00278.x

Abstract

Objectives: This report describes the training of dental examiners participating in two dental caries clinical trials and reports the inter- and intra-examiner reliability scores from the initial standardization sessions.
Methods: Study examiners were trained to use a modified International Caries Detection and Assessment System II system to detect the visual signs of non-cavitated and cavitated dental caries in adult subjects. Dental caries was classified as no caries (S), non-cavitated caries (D1), enamel caries (D2), and dentine caries (D3). Three standardization sessions involving 60 subjects and 3,604 tooth surface calls were used to calculate several measures of examiner reliability.
Results: The prevalence of dental caries observed in the standardization sessions ranged from 1.4 percent to 13.5 percent of the coronal tooth surfaces examined. Overall agreement between pairs of examiners ranged from 0.88 to 0.99. An intra-class coefficient threshold of 0.60 was surpassed for all but one examiner. Inter-examiner unweighted kappa values were low (0.23-0.35), but weighted kappas and the ratio of observed to maximum kappas were more encouraging (0.42-0.83). The highest kappa values occurred for the S/D1 versus D2/D3 two-level classification of dental caries, for which seven of the eight examiners achieved observed to maximum kappa values over 0.90. Intra-examiner reliability was notably higher than inter-examiner reliability for all measures and dental caries classifications employed.
Conclusion: The methods and results for the initial examiner training and standardization sessions for two large clinical trials are reported. Recommendations for others planning examiner training and standardization sessions are offered.

Introduction

Training dental examiners is an essential component of good quality control in dental research. However, little guidance is available with regard to how inter- and intra-examiner reliability should be reported and interpreted and what is an acceptable, or at least realistically achievable, level of performance. This report describes the procedures used for training dental examiners in the Prevention of Adult Caries Study (PACS) and the Xylitol for Adult Caries Trial (X-ACT) studies and provides inter- and intra-examiner reliability results from the initial standardization sessions. Both studies are multisite, randomized clinical trials funded by the National Institute for Dental and Craniofacial Research that evaluate

different interventions for preventing the development of coronal and root caries in adults. Both studies use the same protocol for classifying coronal and root dental caries and for training examiners. The approach and results presented should be instructive to others conducting clinical studies of dental caries involving the detection of cavitated and non-cavitated carious lesions.

Methods

Both the PACS and X-ACT trials employed a dental caries detection system that uses visual signs to detect dental caries. Tactile instrumentation, using a CPITN-E probe (Henry Schein, Inc., Melville, NY, USA), was used only to confirm the presence of cavitation, the width of the marginal gap when caries is associated with restorations and sealants and to gauge the depth of root caries. We combined the seven International Caries Detection and Assessment System (ICDAS)-II caries classification categories into just four categories (1,2). No attempt was made to distinguish between active and inactive lesions because the validity and reliability of methods for determining caries activity are rudimentary. In addition, we trained examiners to refer to these categories using the S, D1, D2, D3 nomenclature rather than the ICDAS-II numerical classifications (3). In addition, we used codes F, C, and P to denote the presence on a surface of a filling, crown, or pit and fissure sealant, respectively, M to denote a missing tooth, and Y to denote a partially erupted or otherwise non-scorable surface. When measuring examiner reliability, the S, F, C, and P codes were deemed to represent a "no caries" call and are hereafter collectively referred to simply as "S" calls, while the Dx, FDx, CDx, and PDx codes denote "dental caries" calls (with x taking values 1, 2, and 3 to reflect the three levels of dental caries severity) (Table 1).

Table 1 Coronal Dental Caries Nomenclature and Classification Categories

Code	Description
M	Tooth is missing for any reason
Y	Partially erupted tooth or unable to score a surface
S	Sound surface with no visible caries and no filling, crown, or sealant
F	Surface has a restoration present but no visible caries at the marginal interface
C	Surface is covered by a crown but shows no visible caries at the marginal interface
P	Surface has a pit and fissure sealant but no visible caries at the marginal interface
D1	Non-cavitated enamel caries (no restoration, crown, or sealant) • after 5 seconds of air-drying, a color change (opacity, brown/ black discoloration) is present that is consistent with the appearance of dental caries and the color change: • extends beyond the confines of the pits and fissures • is located in the gingival 1/3 area of coronal buccal and lingual surfaces • is located just under the contact area or on the gingival 1/3 of mesial and distal surfaces • is located on a root surface that can be visualized
D2	Cavitated enamel caries (no restoration, crown, or sealant) • D1 criteria + cavitation (loss of surface integrity) on coronal surfaces • D1 criteria + cavitation (loss of surface integrity ≥0.5 mm) on root surfaces
D3 (non-cavitated)	Non-cavitated dentine caries (no restoration, crown, or sealant) • D1 criteria (no cavitation) + an underlying dark shadow on coronal surfaces only
D3 (cavitated)	Cavitated dentine caries (no restoration, crown, or sealant) • D2 criteria + dentine is visible in the cavitation on coronal surfaces only
FD1, FD2, FD3	D1, D2, or D3 caries adjacent to a restoration
CD1, CD2, CD3	D1, D2, or D3 caries adjacent to a crown
PD1, PD2, PD3	D1, D2, or D3 caries adjacent to a sealant

A procedure manual was developed and distributed to the study examiners in advance of the training session. A PowerPoint (Microsoft Corporation, Redmond, WA, USA) presentation was used for training. Each initial training and standardization session took 4 days to complete and was composed of 2 days of instruction followed by hands-on training and 2 days for the standardization of the study examiners. For PACS, separate East and West Coast training and standardization sessions were held. All of the X-ACT examiners participated in a single training and standardization session.

The didactic portion of the training included a review of examiner and recorder roles, the examination procedure, and the detection and recording of dental caries. Emphasis was placed on identifying and interpreting the visual clinical appearance of dental caries rather than on determining a diagnosis or making a treatment decision. The positioning of examiners and recorders, the instrumentation to be used, communication between examiners and recorders, and the proper completion of recording forms were also discussed. Additional unstructured time was set aside for general discussion.

For the hands-on training, study examiners initially observed the examination of a volunteer subject by the trainer and then each examiner assessed the subject to verify the calls. Questions about the procedure or reasons for discrepancies in scores were discussed. Two more subjects were evaluated in the same manner followed by a group discussion of the clinical findings and any disparities in the interpretation of the study criteria. Differences in scoring were resolved

by consensus. Eleven additional subjects were then examined by the trainer and the study examiners together, and any differences or disagreements were reviewed and resolved.

Following this training period, we scheduled volunteer subjects for independent examination by the examiners to permit the calculation of inter-examiner reliability. Since the pattern of dental caries is largely bilateral and tends to be more similar within mouths than between mouths, we used half-mouth evaluations with prespecified, contralateral quadrants to allow for a larger number of different subjects to be examined. Since the examinations take roughly 15-20 minutes to complete, we conducted them in six 1.5-hour clinics with four volunteer subjects per clinic. Therefore, 24 subjects were required for each standardization session. In addition, we scheduled eight of those subjects to return for a repeat examination at least 24 hours later to assess intra-examiner reliability. Examiners worked in separate operatories and were blinded to the calls made by other examiners.

A single trainer was used for the examiner training and standardization. Since the purpose of the standardization was to determine how the study examiners compare with one another and not with a standard, sensitivity and specificity calculations are not presented. We report several measures of inter-examiner agreement. First, by converting the four categories of dental caries to a numerical scale (e.g., S = 0, D1 = 1, D2 = 2, D3 = 3) we calculate a mean dental caries score across all tooth surfaces for a given subject by a given examiner. The difference in the mean caries scores provides a measure of relative bias among examiners in terms of their dental caries calls. These numerical caries scores were also used to determine an intra-class correlation coefficient (ICC) that provides an estimate of inter-examiner agreement (4,5).

Since the PACS and X-ACT scoring systems do not distinguish between enamel and dentine cavitated (D2 and D3) lesions, we combined these into a single D2/D3 category. In addition, because PACS includes non-cavitated (D1) lesions in its primary outcome analysis and X-ACT does not, we assessed examiner reliability between the S versus D1/D2/D3 and the S/D1 versus D2/D3 two-level classifications of dental caries and the S versus D1 versus D2/D3 three-level classification. For each of these analyses, we reported the proportion of agreement between pairs of examiners. With the dichotomous classifications, we estimated reliability using a pairwise, unweighted kappa statistic (6). For the three-level comparison, we also calculated a weighted kappa using linear weights. Finally, we included the ratio of the observed kappa to its maximum theoretically possible value given the observed marginal totals (7).We used data from the subset of participants who returned for a repeat evaluation to calculate intra-examiner reliability using a methodology parallel to that for inter-examiner reliability.

For all of the aforementioned analyses, surfaces that were scored as missing (M) or unable to score (Y) by any of the examiners were removed from the analysis because the determination of "missing due to caries" is an unreliable measure as it depends on patient self-reporting or a subjective decision by the examiner. With longitudinal studies employing annual examinations, it is highly unlikely that a lesion will not be detected prior to extraction of a tooth because of caries. Teeth missing for other reasons are not central to the purpose of the PACS and X-ACT studies. Analyses were carried out in SAS Release 9.2 (SAS Institute Inc., Cary, NC, USA).

No definitive criteria exist for what are deemed to be acceptable levels of examiner agreement and reliability. For inter-examiner analyses, we considered the acceptable values to be greater than 0.90 for overall agreement, 0.60 for ICC and 2×2 unweighted kappa, and 0.65 for weighted kappa. Corresponding thresholds for intra-examiner kappa statistics were estimated to be higher, with the minimum acceptable level set to 0.70 for unweighted kappa and 0.75 for weighted kappa.

Results

Although back-up examiners participated in all three training and standardization sessions, only the results for the eight primary examiners are presented. The primary study examiners included three private group dental practitioners, a public health dentist and four dental school faculty members.

Consistent with study eligibility criteria, all volunteer subjects for the examiner training and standardization sessions had either been treated for active caries in the previous 12 months or had active, untreated dental caries. Due to the limited number of root caries detected during the standardization sessions, we only report findings for coronal surfaces.

The overall prevalence of coronal dental caries (D1, D2, and D3 lesions) detected by the study examiners was low, ranging from 1.4 percent to 13.5 percent of the surfaces examined, and the majority of these were non-cavitated (D1) (Table 2).

Analysis of the numerically coded scores revealed evidence of examiner variability, with one examiner (no. 3 at the PACS West Coast session) scoring somewhat lower and one examiner (no. 1 at the X-ACT session) scoring somewhat higher than the others. The acceptable ICC threshold of 0.60 was surpassed (range 0.67-0.87) for all pairs of examiners except those involving examiner no. 3 at the PACS West Coast session (range 0.41-0.47).

Overall agreement ranged from 0.88 to 0.99 across all pairs of examiners and all two-way and three-way comparisons. The highest levels of agreement were consistently achieved for comparisons of S/D1 versus D2/D3 calls (Table 3).

Unweighted kappa statistics ranged from 0.23 to 0.57 for all pairs of examiners and classifications of dental caries. When examiner no. 3 at the PACS West Coast session was excluded, the unweighted kappas ranged from 0.42 to 0.57. Kappas for the two-level classifications of dental caries were generally higher than for the three-level classification.

Table 2 Distribution of Tooth Surface Calls at Initial Standardization Sessions

Caries category	Examiner 1 (%)	Examiner 2 (%)	Examiner 3 (%)
PACS East Coast training (1,255 coded tooth surfaces from 19 subjects)			
S	1,154 (91.9)	1,155 (92.0)	
D1	89 (7.1)	77 (6.1)	
D2	6 (0.5)	15 (1.2)	
D3	6 (0.5)	8 (0.7)	
PACS West Coast (927 coded tooth surfaces from 17 subjects)			
S	875 (94.4)	888 (95.8)	914 (98.6)
D1	30 (3.2)	15 (1.6)	2 (0.2)
D2	14 (1.5)	15 (1.6)	11 (1.2)
D3	8 (0.9)	9 (1.0)	0 (0)
X-ACT (1,422 coded tooth surfaces from 24 subjects)			
S	1,230 (86.5)	1,305 (91.8)	1,302 (91.6)
D1	136 (9.6)	91 (6.4)	88 (6.2)
D2	33 (2.3)	17 (1.2)	18 (1.3)
D3	23 (1.6)	9 (0.6)	14 (0.9)

PACS, Prevention of Adult Caries Study; X-ACT, Xylitol for Adult Caries Trial.

The ratio of the observed to maximum kappa value ranged from 0.45 to 0.83. Nine (43 percent) of the inter-examiner kappa ratios were greater than 0.60. Examiner performance was about the same for all of the classifications of dental caries.

Weighted kappas were generally higher than the corresponding unweighted kappas, ranging from 0.34 to 0.53. When examiner no. 3 at the PACS West Coast session was excluded, weighted kappas ranged from 0.47 to 0.62, although only one examiner pair exceeded 0.60 reliability when using the S versus D1 versus D2/D3 caries classification (Table 3).

Intra-examiner agreement was excellent, ranging from 0.86 to 1.00, with only examiner no. 1 at the X-ACT session

Table 3 Inter-Examiner Reliability Results (E = Examiner)

		Agreement	K (SE)	max K	K/(max K)	Kw (SE)
PACS East Coast						
E1 × E2	S versus D1/D2/D3	0.92	0.45 (0.045)	0.99	0.46	
	S/D1 versus D2/D3	0.99	0.57 (0.101)	0.68	0.83	
	S versus D1 versus D2/D3	0.91	0.43 (0.044)	0.94	0.46	0.53 (0.053)
PACS West Coast						
E1 × E2	S versus D1/D2/D3	0.96	0.57 (0.063)	0.85	0.67	
	S/D1 versus D2/D3	0.98	0.51 (0.091)	0.96	0.53	
	S versus D1 versus D2/D3	0.95	0.46 (0.058)	0.83	0.56	0.62 (0.064)
E1 × E3	S versus D1/D2/D3	0.95	0.29 (0.073)	0.39	0.76	
	S/D1 versus D2/D3	0.98	0.35 (0.108)	0.66	0.53	
	S versus D1 versus D2/D3	0.95	0.23 (0.063)	0.39	0.60	0.38 (0.087)
E2 × E3	S versus D1/D2/D3	0.96	0.29 (0.083)	0.49	0.60	
	S/D1 versus D2/D3	0.98	0.33 (0.104)	0.62	0.53	
	S versus D1 versus D2/D3	0.96	0.26 (0.076)	0.49	0.52	0.34 (0.093)
X-ACT						
E1 × E2	S versus D1/D2/D3	0.90	0.46 (0.037)	0.73	0.63	
	S/D1 versus D2/D3	0.97	0.42 (0.069)	0.62	0.68	
	S versus D1 versus D2/D3	0.89	0.44 (0.035)	0.74	0.60	0.47 (0.044)
E1 × E3	S versus D1/D2/D3	0.89	0.46 (0.037)	0.74	0.62	
	S/D1 versus D2/D3	0.96	0.42 (0.067)	0.72	0.58	
	S versus D1 versus D2/D3	0.88	0.42 (0.035)	0.75	0.56	0.49 (0.042)
E2 × E3	S versus D1/D2/D3	0.92	0.48 (0.042)	0.99	0.49	
	S/D1 versus D2/D3	0.98	0.47 (0.082)	0.89	0.53	
	S versus D1 versus D2/D3	0.91	0.44 (0.040)	0.97	0.45	0.52 (0.048)

K, unweighted kappa; Kw, weighted kappa; max K, maximum possible K; SE, standard error; PACS, Prevention of Adult Caries Study; X-ACT, Xylitol for Adult Caries Trial.

Table 4 Intra-Examiner Reliability Results

		Agreement	K (SE)	max K	K/(max K)	Kw (SE)
PACS East Coast						
E1	S versus D1/D2/D3	0.96	0.75 (0.061)	0.98	0.77	
	S/D1 versus D2/D3	1.00	0.67 (0.315)	0.67	1.00	
	S versus D1 versus D2/D3	0.96	0.76 (0.060)	0.98	0.77	0.74 (0.073)
E2	S versus D1/D2/D3	0.96	0.68 (0.073)	0.89	0.78	
	S/D1 versus D2/D3	1.00	0.89 (0.111)	0.89	1.00	
	S versus D1 versus D2/D3	0.95	0.66 (0.073)	0.87	0.76	0.77 (0.063)
PACS West Coast						
E1	S versus D1/D2/D3	0.97	0.60 (0.115)	1.00	0.60	
	S/D1 versus D2/D3	0.99	0.66 (0.223)	0.66	1.00	
	S versus D1 versus D2/D3	0.97	0.56 (0.114)	0.92	0.61	0.67 (0.121)
E2	S versus D1/D2/D3	0.98	0.56 (0.132)	0.95	0.59	
	S/D1 versus D2/D3	0.99	0.66 (0.159)	1.00	0.66	
	S versus D1 versus D2/D3	0.97	0.46 (0.122)	0.95	0.49	0.69 (0.117)
E3	S versus D1/D2/D3	0.99	0.60 (0.184)	0.60	1.00	
	S/D1 versus D2/D3	0.99	0.75 (0.172)	0.75	1.00	
	S versus D1 versus D2/D3	0.99	0.60 (0.184)	0.60	1.00	0.70 (0.171)
X-ACT						
E1	S versus D1/D2/D3	0.88	0.53 (0.055)	0.64	0.83	
	S/D1 versus D2/D3	0.97	0.67 (0.090)	0.94	0.71	
	S versus D1 versus D2/D3	0.86	0.48 (0.053)	0.65	0.74	0.65 (0.055)
E2	S versus D1/D2/D3	0.92	0.65 (0.055)	0.97	0.67	
	S/D1 versus D2/D3	0.99	0.74 (0.101)	0.91	0.81	
	S versus D1 versus D2/D3	0.91	0.64 (0.054)	0.97	0.66	0.70 (0.060)
E3	S versus D1/D2/D3	0.91	0.62 (0.057)	0.85	0.73	
	S/D1 versus D2/D3	0.99	0.76 (0.105)	0.95	0.79	
	S versus D1 versus D2/D3	0.91	0.62 (0.056)	0.86	0.73	0.68 (0.064)

K, unweighted kappa; Kw, weighted kappa; max K, maximum possible K; SE, standard error; PACS, Prevention of Adult Caries Study; X-ACT, Xylitol for Adult Caries Trial.

falling below 0.90. Similar to inter-examiner agreement, intra-examiner agreement was best when comparing the S/D1 versus D2/D3 calls (Table 4).

Unweighted kappa values for intra-examiner comparisons were noticeably higher than for the inter-examiner comparisons, ranging from 0.46 to 0.76 for the three-level classification of dental caries (S versus D1 versus D2/D3), 0.53 to 0.75 for comparing S versus D1/D2/D3 calls, and 0.66 to 0.89 for comparing S/D1 versus D2/D3 calls. These kappas varied markedly across examiners but were highest for the two PACS East Coast examiners.

The ratios of the observed to maximum intra-examiner kappas ranged from 0.49 to 1.00 for all three dental caries classifications with 17 (71 percent) exceeding 0.70. For the S/D1 versus D2/D3 comparison, seven of the eight examiners (88 percent) had intra-examiner observed to maximum kappa values over 0.90 (Table 4).

The weighted intra-examiner kappa values for the S versus D1 versus D2/D3 classification of dental caries ranged from 0.65 to 0.77, but only one examiner exceeded the 0.75 reliability threshold (Table 4).

Discussion

Many measures are available to estimate examiner reliability, but unfortunately, no current consensus exists as to which measure is preferable when studying dental caries involving non-cavitated lesions. We have deliberately chosen to use and report simpler, descriptive statistical measures for this study because they are widely used and easily computed. There are, however, other more sophisticated methods that can be utilized (8).

The prevalence of dental caries observed at the three standardization sessions varied quite dramatically. Such large differences in dental caries prevalence rates suggest that cross-study comparisons should be interpreted with caution. The very high prevalence of surfaces without caries (the S classification used in this study) among our standardization subjects undoubtedly helped contribute to the high levels of overall inter- and intra-examiner agreement that we observed. This is an acknowledged limitation of this measure of examiner reliability as is the likelihood that some proportion of that agreement can occur by chance alone.

The ICC is commonly used as a measure of inter-examiner reliability for ordinal measures. We contend that the four-level (S, D1, D2, and D3) caries classification system used in the PACS and X-ACT studies is categorical, and therefore, the ICC is a valid measure of examiner reliability. In fact, several investigators consider the ICC to be superior to a weighted kappa when there are multilevel outcome measures (9,10). Fleiss suggests that ICC values between 0.40 and 0.75 represent fair to good reliability (11). In the PACS and X-ACT standardization sessions, ICC levels exceeded 0.60 for all pairs of examiners.

The (unweighted) kappa statistic addresses a perceived limitation of the proportion of overall agreement measure by removing the portion of agreement that is expected to occur by chance alone, thus making it congruent with the classic concept of reliability (i.e., ratio of true to observed variance). However, kappa has its own limitations. The maximum possible agreement is often less than unity because kappa values are influenced by the prevalence of the outcome under study and by the amount of bias present between examiners (12,13). We believe that the very low prevalence of dental caries observed in the standardization subjects, combined with the differences observed among examiners with respect to their mean dental caries scores served to suppress inter-examiner kappa scores to the point where they failed to achieve our predetermined levels of acceptable reliability. Furthermore, the kappa statistic is predicated on the assumption that examiners' calls are independent, and when not certain of a call, an examiner will simply guess, thus producing chance agreement. However, we maintain that the examiners' calls are not independent because they have all undergone a common training program and that totally random guessing would be an unlikely event (14).

Also, some investigators argue that the kappa statistic is interpretable only when the outcome is binary (15). When an outcome is classified into more than two levels and these levels have some natural ordering, not all disagreements may be deemed equally serious. For instance, in our three-level classification (S versus D1 versus D2/D3), paired ratings of S versus D2/D3 may be viewed as reflecting worse disagreement than ratings of S versus D1 or D1 versus D2/D3.

The weighted kappa statistic was designed to take the relative seriousness of examiner disagreements into consideration by giving "partial credit" (a weight between 0 and 1) to intermediate levels of agreement. Although the appropriate choice of weights is a matter of opinion, for our analyses, we assigned weights of 0.5 to ratings of both (S versus D1) and (D1 versus D2/D3). Weighted kappa is most appropriately used in the assessment of reliability for ordered classifications, but this measure should be interpreted with caution not only in light of the weighting issue but because it behaves more similar to a measure of association than an index of agreement (16).

One fairly consistent pattern that emerged from our analyses was that examiner reliability (whether expressed as intra- or inter-examiner agreement) was highest for comparisons of the two-level S/D1 versus D2/D3 classification and lowest for comparisons of the three-level S versus D1 versus D2/D3 classification. The latter finding is not surprising since it is to be expected that reliability will decrease when examiners are asked to make increasingly fine distinctions between disease categories. The fact that reliabilities were consistently higher for the S/D1 versus D2/D3 classification than for the S versus D1/D2/D3 classification suggests that the D1 state is closer in appearance to the sound category (S) than to the cavitated category(scores of D2 and D3).

The S/D1 versus D2/D3 classification of dental caries reflects the traditional North American approach to caries detection. However, despite optimizing kappa reliability scores between study examiners, this classification of dental caries may not necessarily be the most appropriate outcome measure for studies involving etiology and prevention. The S versus D1/D2/D3 dichotomization would make more scientific sense for these types of studies, while the S versus D1 versus D2/D3 classification might be most appropriate for studies that score transitions from S to D1, from D1 to D2, and from S to D2 separately as is the case for the PACS primary outcome analysis.

The threshold levels for acceptable inter-examiner reliability set for the PACS and X-ACT studies were based on Shrout's criteria (17). Accordingly, kappas should ideally exceed 0.80, but a range of 0.60-0.80 may more realistically represent acceptable, though moderate, reliability. However, the use of a universal cut point to define acceptable examiner reliability may not be an ideal or useful goal. Different types of studies involving dental caries may require different types of outcome categories depending on the scientific question(s) being asked. Our results clearly show that the level of reliability, as measured by kappa at least, is dependent on the classification of dental caries used as the outcome measure. More work needs to be done to define the optimal categorizations for different types of studies and to determine realistically achievable levels of examiner reliability for various classifications of dental caries. This is particularly true in the context of collaborative trials such as PACS and X-ACT, where the majority of examiners are practicing dentists, are not employed full time as study examiners, and have divergent educational qualifications and clinical experience.

A related, practical issue associated with examiner training and standardization is how to deal with poor examiner reliability in the context of a collaborative clinical trial. Such trials are conducted on a finite budget and typically with a very tight timeline, and it is often simply not logistically feasible to replace an examiner whose reliability statistics do not

meet preset standards. Inter-examiner reliability for one of the examiners (no. 3 at the PACS West Coast session) proved to be comparatively low. This examiner was subsequently provided additional training and underwent a second standardization session but was ultimately replaced during the study.

We can speculate as to some of the factors that may affect examiner reliability in clinical dental caries studies. For some examiners, a single initial training session may not be sufficient to overcome the "built-in" programming for dental caries detection developed through his or her dental education and experience. This embedded programming is difficult to overwrite with short-term training and it is, we suspect, what an examiner ultimately reverts to when faced with a dental caries call dilemma. Furthermore, if an examiner does not examine study subjects on a daily basis, reinforcement of study detection criteria and procedures through repetition does not take place, thus adversely affecting examiner reliability. In addition, many enamel anomalies mimic non-cavitated dental caries in adults and hence, may further confound the caries detection process. The differentiation between non-carious white spot lesions (dental fluorosis, developmental abnormalities, hypocalcification, etc.) and non-cavitated dental caries represents a particularly challenging aspect of examiner training. Finally, although treatment considerations should not influence the detection of dental caries, we suspect this happens to varying degrees, particularly with respect to cavitated lesions.

An independent, and longer, initial training session might have served to facilitate the learning process for the examiners not only with respect to the study criteria but also for the large number of rules that are needed to assist an examiner in making a dental caries call. A separate standardization session, with a longer schedule of examinations, could then have been organized to quickly follow (within a month) the training session, with a review and final reliability evaluation immediately preceding the baseline examinations. However, the practicality of a longer training period and additional standardization sessions prior to the launch of the clinical trial in terms of time, cost, examiner fatigue, subject recruitment, resources, and other considerations must be weighed against the expected improvement in the examiner reliability measures and the impact it would have on the study.

Only a small number of trials or epidemiologic studies have reported detailed reliability data for examinations that include non-cavitated dental caries (18-24). Comparisons between this study and those studies are difficult because in the other reported studies, different diagnostic systems were used, the prevalence of dental caries was quite high compared with this study, the other studies examined children and young adult subjects, whereas we examined only adult subjects and, in the other studies, the outcome classification was not always reported (e.g., whether binary or >2 categories were used). Nevertheless, the inter-examiner reliability kappa values for the PACS and X-ACT examiners tend toward the lower end of those reported in the literature, although the intra-examiner kappa values compare more favorably (Table 5).

The results presented in this report are limited to an initial standardization of study examiners. Ongoing standardization sessions occurred roughly every 9 months for PACS and every 12 months for X-ACT. The format of the ongoing sessions was similar to the initial session except that the didactic portion was truncated to allow the standardization of study examiners to be completed in 2 days.

Recommendations

The experience gained through training and standardizing examiners for the PACS and X-ACT dental caries studies naturally leads to observations and recommendations that may be of assistance to those planning future clinical dental caries studies involving non-cavitated dental caries lesions. Based on the PACS and X-ACT initial training and standardization sessions, we suggest the following observations merit consideration:

- The initial training session for study examiners should involve a minimum of 2 days if the examiners are not familiar with the coding.
- The initial training session should be followed immediately, or within a week, by at least a 2-day examiner standardization session composed of eight 1.5-hour clinic sessions.
- There should be an abbreviated review session and a second standardization session, totaling 2 days duration before the baseline examinations are begun.
- At least annually, and preferably semi-annually, examiner standardization sessions should be incorporated into the planning of clinical trials.
- Subjects recruited for examiner training and standardization sessions should be representative of the study population and screened for the presence of both non-cavitated and cavitated carious lesions.
- Ideally, four to six examiners should participate in a standardization session.
- Initially, it takes an examiner about 20 minutes to complete a half-mouth examination using the study procedures and criteria, but once the study examiners are familiar with the study criteria and procedures, full-mouth examinations can be utilized in subsequent (re-standardization) sessions.
- Four subjects can be comfortably examined by four to six examiners using a half-mouth (two diagonal quadrants) in a clinic session lasting 1.5 hours.
- At least the final two 1.5-hour clinic sessions of the standardization session should be devoted to determining intra-examiner reliability.

Table 5 Examiner Reliability Reported in Studies Using a Non-Cavitated Caries Category

Study	Number of examiners	Number of subjects	Mean age (years)	Examiner training	Examination conditions	Caries criteria	Caries prevalence	Reliability measure	Range of reported inter-examiner reliability	Range of reported intra-examiner reliability
Pitts and Fyffe, 1988 (3)	3	287	21	Details not reported	Dental operatory	WHO (1979) modified	Mean D1MFS 8.40; mean D2MFS 6.12; mean D3MFS 4.75	kappa	Not reported	0.80 (D1 lesions); 0.82 (D2 lesions); 0.80 (D3 lesions)
Ismail *et al*., 1992 (18)	3	91	8	15 days, 200 examinations	Portable equipment	special, with D1 calls		kappa	0.9	0.9
Nyvad *et al.*, 1999 (19)	2	350 (50 × 7 times over 3 yr)	9-14	Discussions and exercises over 1 month	Portable equipment	Nyvad *et al.*, 1999 (19)	Not reported	kappa	0.85-0.86 (sound/decayed) 0.88-0.93 (sound/cavitated)	0.81-0.90 (sound/decayed) 0.81-0.90 (sound/cavitated)
Fyffe *et al.*, 2000 (20)	20	43	13	2.5 days	Portable equipment, subject on table	DSTM	Mean decayed surfaces 10.2-20.4	kappa	0.47-0.61 (D1 lesions); 0.52-0.67 (D3 lesions)	Not reported
Assaf *et al*., 2006 (a, b) (21,22)	11	64	6-7	4 sessions, 28 hours total	Outdoor, natural light	WHO + IL	Mean dmft 4.0	kappa	0.11-0.65	Not reported
Ismail *et al*., 2007 (2)	4 4	292 338	Not reported	Details not reported	Not reported	ICDAS	Not reported	weighted kappa weighted kappa	0.63-0.84 0.68-0.85	0.59-0.82 0.73-0.85
Kühnisch *et al*., 2008 (23)	1	20 (molars only)	8-12	Details not reported	Mobile dental unit	ICDAS-II	Not reported	weighted kappa	0.90 (versus calibrator)	0.88
Braga *et al*., 2009 (24)	6	126	4	2 sessions, 8 hours total	Dental chair with conventional lighting	ICDAS-II	83%	kappa weighted kappa	0.50 (sound/decayed) 0.78	Not reported Not reported

dmft, decayed, missing and filled primary teeth; DSTM, Dundee Selectable Threshold Method; ICDAS, International Caries Detection and Assessment System; IL, initial lesions; WHO, World Health Organization.

- Twenty-four subjects, with eight of those subjects returning for a repeat examination, provide a sufficient number of surface observations for examiner reliability determination.
- It is most efficient to have four operatories or examination areas available for each clinic session.
- All study examiners should use lenses with the same magnification when examining study subjects.
- When the prevalence of dental caries is less than 10 percent, the ratio of the observed kappa statistic to the maximum possible kappa statistic permitted by the marginal totals should be reported.
- If the levels of acceptable examiner reliability determined for a study are not achieved, one or more of the examiners will require additional training and, if reliability does not improve, replacement.
- None of the measures of examiner reliability is without limitations and we are hard-pressed to recommend one over another at this time.
- When study participants will be examined by the same examiner at each follow-up visit, intra-examiner reliability is at least as important a metric as inter-examiner reliability.
- An international conference/workshop should be organized to thoroughly investigate, discuss, and reach a consensus on the best measure(s) to assess examiner reliability and the achievable and acceptable levels of reliability for different types of clinical studies using various classifications of dental caries.

Acknowledgment

The PACS and X-ACT clinical trials are supported by the NIDCR (ClinicalTrials.gov identifiers NCT00357877 and NCT00393055, respectively), with some additional funding and in-kind support of PACS by CHX Technologies.

References

1. Ismail AI, Banting D, Eggertsson H, Ekstrand K, Ferreira-Zandoná A, Longbottom C, Pitts NB, Reich E, Ricketts D, Selwitz R, Sohn S, Topping GVA, Zero D. Rational and Evidence for the International Caries Detection and Assessment System (ICDAS II). In: Stookey GK, editor. *Clinical Models Workshop: Remin-Demin, Precavitation, Caries*. Proceedings of the 7th Indiana Conference. Indiana University School of Dentistry, Indianapolis, Indiana; 2005. p. 161–21.
2. Ismail AI, Sohn W, Tellez M, Amaya A, Sen A, Hasson H, Pitts NB. The International Caries Detection and Assessment System (ICDAS): an integrated system for measuring dental caries. *Community Dent Oral Epidemiol.* 2007;**35**(3): 170-8.
3. Pitts NB, Fyffe HE. The effect of varying diagnostic thresholds upon clinical caries data for a low prevalence group. *J Dent Res.* 1988;**67**(3):592-6.
4. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;**86**(2):420-8.
5. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996;**1**: 30-46.
6. Cohen J. A coefficient of agreement of nominal scales. *Psychol Bull.* 1960;**20**:37-46.
7. Dunn G. *Design and analysis of reliability studies: the statistical evaluation of measurement errors*. London: Edward Arnold; 1989.
8. Dunn G. *Statistical evaluation of measurement errors*. London: Arnold, a Member of the Hodder Headline Group; 2004.
9. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as a measure of reliability. *Educ Psychol Meas.* 1973;**33**:613-7.
10. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol.* 1987;**126**(2):161-9.
11. Fleiss JL. *The design and analysis of clinical experiments*. New York: John Wiley & Sons; 1986.
12. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol.* 1993;**46**(5):423-9.
13. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther.* 2005;**85**(3):257-68.
14. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. *Psychol Bull.* 1987;**101**: 140-6.
15. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry.* 2002;**59**(10):877-83.
16. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. *J Clin Epidemiol.* 1993;**46**(9): 1055-62.
17. Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res.* 1998;**7**(3):301-17.
18. Ismail AI, Brodeur JM, Gagnon P, Payette M, Picard D, Hamalian T, Olivier M, Eastwood BJ. Prevalence of non-cavitated and cavitated carious lesions in a random sample of 7-9-year-old schoolchildren in Montreal, Quebec. *Community Dent Oral Epidemiol.* 1992;**20**(5): 250-5.
19. Nyvad B, Machiulskiene V, Baelum V. Reliability of a new caries diagnostic system differentiating between active and inactive caries lesions. *Caries Res.* 1999;**33**(4):252-60.
20. Fyffe HE, Deery C, Nugent ZJ, Nuttall NM, Pitts NB. Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). *Community Dent Oral Epidemiol.* 2000;**28**(1):42-51.
21. Assaf AV, Meneghim MC, Zanin L, Cortelazzi KL, Pereira AC, Ambrosano GM. Effect of different diagnostic thresholds on dental caries calibration. *J Public Health Dent.* 2006;**66**(1): 17-22.
22. Assaf AV, de Castro MM, Zanin L, Tengan C, Pereira AC. Effect of different diagnostic thresholds on dental caries

calibration – a 12 month evaluation. *Community Dent Oral Epidemiol.* 2006;**34**(3):213-9.

23. Kühnisch J, Berger S, Goddon I, Senkel H, Pitts N, Heinrich-Weltzien R. Occlusal caries detection in permanent molars according to WHO basic methods, ICDAS II and laser fluorescence measurements. *Community Dent Oral Epidemiol.* 2008;**36**(6):475-84.
24. Braga MM, Oliveira LB, Bonini GA, Bonecker M, Mendes FM. Feasibility of the International Caries Detection and Assessment System (ICDAS-II) in epidemiological surveys and comparability with standard World Health Organization criteria. *Caries Res.* 2009;**43**(4):245-9.