Comparing Film and Digital Radiographs for Reliability of Pneumoconiosis Classifications:

A Modeling Approach

Ananda Sen, PhD, Shih-Yuan Lee, MS, Brenda W. Gillespie, PhD, Ella A. Kazerooni, MD, Mitchell M. Goodsitt, PhD, Kenneth D. Rosenman, MD, MPH, James E. Lockey, MD, MS, Cristopher A. Meyer, MD, E. Lee Petsonk, MD, Mei Lin Wang, MD, Alfred Franzblau, MD

Rationale and Objectives: The International Labour Office (ILO) system for classifying chest radiographic changes related to inhalation of pathogenic dusts is predicated on film-screen radiography. Digital radiography has replaced film in many centers. Digital images can be printed on film ("hard copy") or can be viewed at a computer workstation ("soft copy"). The goal of the present investigation was to compare the inter-reader and intra-reader agreement of ILO classifications for pneumoconiosis across image formats.

Materials and Methods: Traditional film radiographs, hard copy digital images, and soft copy digital images from 107 subjects were read by six B readers. A multiple reader version of the inter-reader kappa statistic was compared across image formats. Intra-reader kappa comparisons were carried out using an iterative least-squares approach (unadjusted analysis) as well as a two-stage regression model adjusting for readers and subject-level covariates.

Results: There were few significant differences in the inter-reader and intra-reader agreement across formats. For parenchymal abnormalities, inter-reader and intra-reader kappa values ranged from 0.536 to 0.646, and 0.65 to 0.77, respectively. In the covariate-adjusted analysis film-screen radiography was generally associated with a numerically greater reliability (ie, higher kappa values) than the other image formats, although differences were rarely statistically significant.

Conclusion: Film-screen radiographs, hard copy digital images, and soft copy digital images yielded similar reliability measures. These findings provide further support to the recommendation that soft copy digital images can be used for the recognition and classification of dust-related parenchymal abnormalities using the ILO system.

Key Words: Digital radiography; pneumoconiosis; B readers; ILO System; kappa.

Acad Radiol 2010; 17:511–519

From the Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI (A.S.), Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI (S.-Y.L., B.W.G), Department of Radiology, University of Michigan Medical School, Ann Arbor, MI (E.A.K., M.M.G.), Department of Internal Medicine, Michigan State University Medical School, East Lansing, MI (K.D.R.), Departments of Environmental Health (J.E.L.) and Radiology (C.A.M.), University of Cincinnati Medical School, Cincinnati, OH, National Institute for Occupational Safety and Health, Morgantown, WV (E.L.P., M.L.W.), Department of Environmental Health Sciences, University of Michigan School of Public Health, 109 South Observatory Street, Ann Arbor, MI 48109-2029 (A.F.). Received October 8, 2009; accepted December 4, 2009. Support for this research was provided by grant #S2200-22/23 from the Association of Schools of Public Health (ASPH) and the Centers for Disease Control and Prevention (CDC)/National Institute for Occupational Safety and Health (NIOSH). Mention of commercial products is not an endorsement by the authors, ASPH, CDC, NIOSH, or the ILO. The findings and conclusions contained in this report are those of the authors and do not necessarily reflect those of ASPH, CDC, NIOSH, or the ILO. **Address correspondence to:** A.F. e-mail: afranz@umich.edu

doi:10.1016/j.acra.2009.12.003

Since the early 20th century, standard posteroanterior (PA) film-screen chest radiography (FSR) has been the primary method for screening, diagnosis, medical monitoring, and epidemiological study of the pneumoconioses. In the 1930s, the International Labour Office (ILO) developed a scoring system for standardizing the classification of radiographs of pneumoconioses (1). The system has undergone several revisions, most recently in 2000 (2). The ILO system remains the most widely used method for classifying chest radiographs for pleural and parenchymal abnormalities related to inhalation of pathogenic dusts (3,4).

During the past three decades, numerous medical centers around the globe have adopted different forms of digital x-ray imaging into clinical practice. The widespread adoption of digital x-ray technology has numerous implications for use of the ILO system. The ILO system is dependent on the use of traditional FSR that is no longer available in many parts of United States or some other countries. Thus

a need and preference for moving towards digital technology is sharply on the rise.

Few investigations have attempted to examine equivalence between FSR and digital images in terms of identification and quantification of parenchymal and pleural changes because of dust inhalation, or have evaluated inter-reader or intra-reader agreement across image formats (5,6). Note that digital images can be presented in two different ways, namely "hard copy" (HC), which is laser printed on film and appears much like FSR, and "soft copy" (SC), which can be viewed on a computer workstation.

The goal of the present investigation was to compare the impact of chest radiograph image formats (FSR, HC, SC) on the results of ILO classifications performed by experienced readers among individuals with abnormalities of the lung parenchyma or pleura that may have resulted from dust inhalation. We recently reported similarity of results using either SC digital radiography or traditional FSR in the recognition and quantification of parenchymal abnormalities associated with pneumoconiosis and other forms of fibrotic lung disease.[7] The focus of the present report is on comparing both the inter-reader and the intra-reader agreement across the three formats using the kappa statistic, where each subject's chest image is read by all readers.

MATERIALS AND METHODS

Detailed methods on study design and data collection have been previously described (7). Briefly, subjects were recruited from two sources: patients seen or referred to the University of Michigan Medical Center and patients listed in the Michigan or Ohio Silicosis Registries (8,9). Subjects gave written informed consent that was approved by the Medical Institutional Review Board of the University of Michigan Medical Center and complied with Health Insurance Portability and Accountability Act regulations. All subjects completed a survey that included questions in the following areas: demographics (age, gender, race/ethnicity); smoking history; occupational and dust exposure history; and medical history, particularly related to lung and pleural disease (10). Height and weight were measured. A total of 107 subjects with a range of parenchymal and pleural abnormalities was recruited to participate in the study. For each subject a standard PA film radiograph and a PA digital radiograph (DR) were obtained on the same day. DR images were captured on the flat-panel amorphous selenium digital detector of the Hologic DR 1000C system (Hologic, Inc, Bedford, MA). Standard PA chest film-screen technique was employed: 125 kVp, 150 mA, wall unit, 72" (183 cm) source-to-image distance, all three phototimer sensors, Agfa UVC film (Agfa-Gevaert Group, Wilmington, DE) in Agfa UV Super Rapid Screen Cassette, Normal "0" density setting (the speed of the screen-film system was 200). DR images were printed on a Fuji FM-DPL high-quality laser printer (Fujifilm Medical Systems USA, Inc, Stamford, CT) using Fuji film. One DR image was lost in the picture archiving and communication system and was not recoverable, and one FSR was lost in the radiology file room, each on different subjects. Therefore, the final study group included 318 images, 106 each for FSR, HC, and SC, respectively, but these were based on 107 subjects. All images were read twice, in random order, by six B readers using the 2000 revision of the ILO classification system (2). SC DR images were read using four different picture archiving and communication systems. To ensure that all B readers employed comparable high-quality display monitors when reading SC images, they were required to use a high-resolution physician-quality diagnostic workstation similar to that employed in the radiology department at the University of Michigan at the time of the study (ie, Siemens Sienet Magic View 1000 system).

Statistical analyses were performed using SAS for Windows version 9.1 (11).

The most widely used statistical tool for assessing rater agreement using a categorical rating scale is the kappa statistic (12). In the case of two readers (or two rounds of readings by the same reader) rating n subjects into one of m mutually exclusive and exhaustive categories, let p_{ij} be the probability that a given subject is assigned to the i^{th} and j^{th} categories by readers 1 and 2, respectively. Consequently, $p_{i.} = \sum_{j=1}^{m} p_{ij}, p_{.j} = \sum_{i=1}^{m} p_{ij}$ denote the marginal probabilities of ratings in the i^{th} and j^{th} categories by raters 1 and 2, respectively. In this case, the kappa statistic is defined by

$$k = \frac{P_0 - P_c}{1 - P_c}, \tag{1}$$

where $P_0 = \sum_{i=1}^{m} p_{ii}$ denotes the observed probability of agreement and $P_c = \sum_{i=1}^{m} p_{i.}p_{.i}$ is the probability of agreement that would be expected purely by chance. Thus kappa measures agreement in excess of what would be expected by chance alone and higher kappa values imply greater reliability. Values of kappa greater than 0.75 are considered excellent, values between 0.40 and 0.75 are fair to good, and values less than 0.40 represent poor agreement beyond chance alone (13). Cohen recommended using the above definition for kappa for ratings in a nominal scale and suggested using weighted kappa when ordinal rating categories are used (12). For dichotomous rating scales, all weighted versions reduce to the unweighted version of kappa.

Analyzing Inter-reader Reliability

Inter-reader kappa statistics were calculated for each image format and round for each outcome measure. With six different readers, a set of $(6 \times 5)/2 = 15$ pairwise inter-reader kappa statistics is available. Because the between-format comparison and not the between-reader comparison is of primary interest, it is reasonable to work with a kappa measure that is pooled across all readers. An overall agreement measure among readers was computed for each image format and each round using the multi-reader version of the kappa statistic (12). This version is applicable to outcomes with dichotomous

as well as multiple ordinal/nominal categories. In the case of dichotomous outcomes, kappa has an attractive interpretation as an intra-class correlation coefficient. For the case where the number of rating categories is more than two (eg, small opacities), overall kappa can be expressed as a weighted average of category-specific kappa statistics. The SAS macro "MAGREE" was used to compute the overall multireader kappa value. The standard error for the multireader kappa is available only for the case when the true value of kappa equals zero and should only be used for testing whether kappa is significantly different from zero. To date, no general formula for the standard error for multirater kappa is available in the literature. To overcome this limitation, standard errors were calculated using a bootstrap method based on 2000 replications. The bootstrap samples were generated in SAS by drawing random samples with replacement from the given data of readings from all readers and each round. Each bootstrap replication consisted of as many observations as the original data. The multiple reader version of kappa was then calculated from each replication, and the 95% confidence interval was estimated as the interval between the 2.5th and the 97.5th sample percentiles from these 2000 kappa values.

Analysis of Intra-reader Reliability

Intra-reader kappa values, both within and between formats, were computed using equation 1.

The statistical methodologies used to carry out the inference on intra-reader agreement is largely nonstandard and requires modeling techniques that account for clustering effects within observations made on the same subject by different readers using different image formats. Although such modeling is routinely pursued in analyzing averages or probabilities, adaptation to the case of agreement measures is not straightforward, and required substantial programming effort for implementation.

When comparing intra-reader reliability between different image formats, the corresponding intra-reader kappa statistics are dependent, because they are produced from the same pool of patients. As a consequence, the usual testing procedure associated with comparing independent kappa statistics is no longer applicable. Barnhart and Williamson present a weighted least squares algorithm to compare correlated kappa statistics that uses an estimating equations approach for the inference and can be implemented using the capabilities of PROC CATMOD in SAS (14). We used this approach to compare the intra-reader (within format) kappa statistics across the three image formats for each B reader. The computational burden of this approach increases rapidly with an increase in the number of rating categories.

Regression Analysis of Intra-reader Reliability

Although an intra-reader kappa is a within-reader comparison and does not inherently require further adjustment for subject-level characteristics, it is still of interest to investigate the potential effect of subject-level information on these reliability measures. For example, it is possible that age, gender, or body mass index may affect the interpretation of the image, thereby affecting the reliability.

In the past decade or so, modeling agreement has been an active research area. Among the many methods proposed, the approach suggested by Lipsitz et al was adopted here primarily due to the simplicity of its implementation (15). This approach is applicable to any number of rating categories and can account for subject-level clustering. The basic approach is to fit a linear regression model of the type

$$\text{kappa} = \beta_0 \sum \beta_k W_K \tag{2}$$

where kappa is a subject-specific version of intra-reader kappa statistics, the β_k are regression coefficients, and the W_k are independent covariates, such as age, gender, body mass index, pack-years of smoking and individual B readers. Because models of kappa are not directly implementable in standard software packages, a multistage approach was adopted to fit such a model, as described in the following section.

First, considering only binary rating scales for modeling the intra-reader kappa statistics, rearranging terms in equation 1, and indexing the terms by i to indicate the i^{th} subject, a subject-level agreement statistic, κ_i, can be formulated as

$$P_{oi} = P_{ci} + (1 - P_{ci})k_i, \tag{3}$$

where $P_{oi} = \Pr(X_1 = 0, X_2 = 0) + \Pr(X_1 = 1, X_2 = 1)$ is the probability of exact agreement in rating, and $P_{ci} = \Pr(X_1 = 0)\Pr(X_2 = 0) + \Pr(X_1 = 1)\Pr(X_2 = 1)$, X_1 and X_2 being the ratings by the same reader in the two rounds, respectively. Here κ_i is the subject-level version of the intra-reader agreement statistic kappa as defined in equation 1. The basic strategy consists of a two-stage process as follows.

Stage 1

The marginal probability of "success" [$\Pr(X_1 = 1)$ and $\Pr(X_2 = 1)$] was modeled in a logistic regression framework, with image format and reader as factors, and age, gender, body mass index, and smoking history as subject-level covariates. For outcomes other than image quality, the models were further adjusted for median rating of image quality. A generalized estimating equation approach was used to account for the clustering of image readings for the same subject.

Stage 2

In the second stage, the probability of exact agreement, denoted by P_{oi} in equation 3, is modeled again in a clustered binomial logistic regression framework (with identity link), with the regression structure of equation 2 directly imposed on κ_i. Further, P_{ci} appearing on the right side of equation 3

is estimated by using the values obtained from the first stage model. It is clear from equation 3 that by treating P_{ci} as known constants, κ_i is simply a linear transform of P_{oi}, and so a linear regression fit to one of them uniquely corresponds to a linear regression fit to the other.

The above describes the approach in the context of binary rating scales. The extension to the case of multiple categories is straightforward. The algorithm used herein is a slight modification to that proposed by Lipsitz et al and is implemented in SAS 9.1 (11,15).

RESULTS

The mean age of the 107 subjects was 64.7 years. Most were male (80%), most reported a history of pathogenic dust exposure (56%), most were former or current smokers (64%), and most were overweight or obese (74%) (Table 1).

A total of 3816 readings were completed for the study (106 images × 3 formats × 6 readers × 2 rounds). All of the major small opacity profusion categories were represented in the study group (43%, 30%, 21%, and 6%, respectively, for ILO major profusion categories 0, 1, 2, and 3 based on FSR). There was also a reasonable representation of both small rounded (34%) and small irregular opacities (66%), based on reading of FSR images. Fifteen percent of readings of FSR images indicated the presence of large opacities, and 37% indicated the presence of pleural abnormalities.

TABLE 1. Subject Characteristics

	Frequency	Percent
Gender		
Male	86	80
Female	21	20
Body mass index (BMI, kg/m^2)		
BMI <25 (normal)	28	26
25 ≤BMI <30 (overweight)	45	42
30 ≤BMI (obese)	34	32
Ever smoked		
No	39	36
Yes	68	64
Current smoking		
No	97	91
Yes	10	9
History of dust exposure		
No	47	44
Yes	60	56
Dust exposure type (n = 60)*		
Silica	34	57
Asbestos	28	47
Other/unknown	12	20
	Mean (SD)	Median (range)
Age (y)	64.7 (11.9)	65 (31–91)
BMI (kg/m^2)	28.5 (5.2)	28.1 (19.5–48.8)
Pack-years		
All subjects (n = 107)	19.5 (24.1)	12 (0–96)
Ever smoked (n = 68)	30.7 (23.8)	23.5 (1–96)

*Some subjects reported more than one type of dust exposure, so the numbers sum to more than 60.

Inter-reader Reliability

Results comparing the multi-reader version of the inter-reader kappa statistic across image formats are displayed in Table 2A for both rounds. The table reports three image quality measurements: one with a four-point ordinal scale, the other two with dichotomous classifications of categories 1–2 versus categories 3–4, and category 1 versus 2–4. For round 1, only three pairwise comparisons, all related to image quality, exhibited a significant difference. In these instances, FSR images had significantly better inter-reader reliability than readings of HC and SC images (Table 2B). None of the inter-reader kappa comparisons between image formats in round 2 were statistically significant. Note that although the bootstrap confidence intervals based on the percentile method are generally known to be quite conservative (wide), in the present case they turned out to be close to the normal approximation–based confidence intervals calculated as: kappa ± 1.96 * (standard error). The latter is an indication that the bootstrapped distributions of the estimated kappa were reasonably symmetric.

Intra-reader Reliability

Table 3 describes two types of intra-reader kappa values for each of the variables considered. The first type–labeled "within format"–documents the mean and range of intra-round kappa for the six B readers, computed for each format (ie, between rounds, within format). The "between format" kappa signifies the kappa computed for all pairs of formats, the mean and range presenting the summary over all readers and rounds (ie, between formats, within rounds). The within format kappa values were generally higher than the between format values for the same outcomes; these differences may reflect the fact that the between format kappa values incorporate the impact of image format and intra-reader variation, whereas within format kappa values primarily reflect intra-reader variation. Overall agreement for the image quality variables tended to be lower and less consistent compared to other outcomes. For the variables measured on ordinal scales with more than two categories, kappa values were also calculated based on different weighting schemes. Not surprisingly, these were larger in magnitude compared to the corresponding unweighted kappas. However, the relative rankings were generally unchanged (data not shown). The rest of the analysis focused on comparing the within-format kappa values across the three different image formats.

In Table 3, the within-format kappa values for FSR in general appear to be slightly higher than for HC or SC. To formally compare the kappas across the image formats, we

TABLE 2A. Inter-reader Kappa Values of Reader Agreement by Image Format

	Inter-reader Kappa Round 1			Inter-reader Kappa Round 2		
	FSR	HC	SC	FSR	HC	SC
1.A: Image Quality (IQ: 4-point ordinal scale)	0.318	0.163	0.206	0.225	0.191	0.193
1.A: IQ (Category 1 versus 2–4)	0.338	0.270	0.242	0.245	0.266	0.244
1.A: IQ (Categories 1–2 versus 3–4)	0.432	0.215	0.182	0.305	0.218	0.087
2.A: Any parenchymal abnormalities (yes/no)	0.595	0.553	0.629	0.623	0.536	0.646
2.B: Small opacities (12-point scale)	0.288	0.225	0.271	0.258	0.240	0.256
2.B: Small opacities (4-point scale)	0.468	0.425	0.443	0.452	0.449	0.458
2.C: Large opacities (4-point scale)	0.463	0.506	0.463	0.453	0.531	0.447
2.C: Large opacities (yes/no)	0.622	0.701	0.621	0.597	0.718	0.635
3.A: Pleural abnormalities (yes/no)	0.415	0.524	0.499	0.462	0.421	0.437
3.C: Costophrenic angle obliteration (yes/no)	0.566	0.567	0.500	0.531	0.370	0.367
3.D: Diffuse pleural thickening (yes/no)	0.599	0.557	0.559	0.619	0.416	0.454

TABLE 2B. Significant Differences Among Inter-reader Kappa Values for Round 1*

	Kappa Difference	SE	Lower CI	Upper CI
Kappa difference image quality (4-point ordinal scale): FSR versus HC	0.155	0.052	0.046	0.250
Kappa difference image quality (categories 1–2 versus 3–4): FSR versus HC	0.217	0.096	0.010	0.376
Kappa difference image quality (categories 1–2 versus 3–4): FSR versus SC	0.250	0.100	0.032	0.417

*Standard errors and 95% confidence intervals for inter-reader kappa were calculated using bootstrap method (2000 replications, see text).

adjusted for the clustering effect within each subject. Table 4 reports kappa measures for dichotomous outcomes for each reader, as well as P values for the pairwise differences across image formats, using the weighted least-squares approach. A more detailed comparison including variables with polytomous ($\geq$2) rating scales is presented in Table 5 under the regression framework. The unadjusted comparisons in Table 4 generally show little difference across image formats with perhaps the exception of image quality. Although a few significant differences in intra-reader kappa across image formats are found for most readers, there is no discernible pattern among readers in terms of one format being consistently more reliable than the others. Importantly, none of the intra-reader kappa values for small parenchymal abnormalities or large opacities differed significantly. For Readers 3 and 5, there were some significant differences in intra-reader kappa values for pleural abnormalities, particularly for FSR vs. HC and HC vs. SC, but not for FSR vs. SC; none of the other readers demonstrated any significant differences in kappa values for any of the pleural outcomes.

In the regression analysis of intra-reader kappa, once again most comparisons produced statistically nonsignificant results (Table 5). The exceptions were image quality (categories 1–2 versus 3–4), where HC was associated with higher agreement than SC; pleural abnormality, where HC was associated with lower agreement than SC; and large opacities on a four-point scale, where FSR was associated with higher agreement than HC. With respect to image quality (category 1 versus 2–4), FSR was associated with higher agreement than SC with a borderline significance (P = .05). A trend is observed (.05 < P value < .1) both for image quality (category 1–2 versus 3–4) as well as large opacities (4-point scale), in which FSR was associated with a higher reliability compared to SC. In most instances, FSR was associated with a numerically higher intra-reader agreement than the other formats, although differences were rarely significant. In the regression model for pleural abnormalities, both greater age and median image quality were significantly associated with an overall higher reliability, while number of pack years was negatively associated. The reliability was significantly higher for females compared to males for the variable indicating presence or absence of parenchymal abnormalities. On the other hand, the overall reliability with respect to image quality (category 1 versus 2–4) was significantly higher for younger ages. No other significant associations with the covariates were obtained in the regression analysis. Note that due to the complexity of the regression modeling framework, it is difficult to interpret the regression coefficients beyond association.

TABLE 3. Intra-reader Kappa Values within and between Formats*

Variables	Within Format†			Between Format‡		
	FSR	HC	SC	FSR and HC	HC and SC	FSR and SC
1.A: Image quality (IQ: 4-point ordinal)	0.49 (0.31, 0.68)	0.41 (0.29, 0.52)	0.42 (0.21, 0.65)	0.13 (-0.11, 0.45)	0.28 (0.18, 0.50)	0.18 (0.05, 0.41)
1.A: IQ (Category 1 vs. 2–4)	0.54 (0.30, 0.76)	0.45 (0.07, 0.64)	0.48 (0.23, 0.67)	0.16 (-0.16, 0.53)	0.34 (0.17, 0.59)	0.21 (0.03, 0.48)
1.A: IQ (Categories 1–2 vs. 3–4)	0.51 (0.34, 0.65)	0.38 (-0.01, 0.52)	0.29 (-0.03, .49)	0.16 (-0.02, 0.38)	0.23 (-0.02, 0.51)	0.13 (-0.04, 0.36)
2.A: Parenchymal abnormalities (yes/no)	0.77 (0.71, 0.82)	0.75 (0.66, 0.79)	0.72 (0.61, 0.86)	0.65 (0.54, 0.75)	0.68 (0.52, 0.81)	0.70 (0.57, 0.80)
2.B: Small opacities (12-point scale)	0.39 (0.29, 0.46)	0.36 (0.26, 0.48)	0.37 (0.26, 0.51)	0.29 (0.22, 0.38)	0.33 (0.25, 0.40)	0.35 (0.27, 0.41)
2.B: Small opacities (4-point scale)	0.61 (0.49, 0.67)	0.60 (0.47, 0.70)	0.59 (0.50, 0.68)	0.48 (0.40, 0.59)	0.53 (0.47, 0.56)	0.56 (0.41, 0.64)
2.C: Large opacities (4-point scale)	0.70 (0.60, 0.86)	0.70 (0.62, 0.77)	0.61 (0.48, 0.83)	0.63 (0.55, 0.83)	0.58 (0.50, 0.77)	0.58 (0.44, 0.76)
2.C: Large opacities (yes/no)	0.81 (0.67, 0.96)	0.81 (0.77, 0.85)	0.75 (0.69, 0.88)	0.74 (0.60, 0.92)	0.74 (0.64, 0.83)	0.72 (0.57, 0.88)
3.A: Pleural abnormalities (yes/no)	0.69 (0.54, 0.83)	0.66 (0.47, 0.85)	0.69 (0.59, 0.84)	0.59 (0.49, 0.64)	0.66 (0.58, 0.73)	0.56 (0.49, 0.68)
3.C: Costophrenic angle Obliteration (yes/no)	0.73 (0.48, 0.85)	0.52 (0, 0.85)	0.59 (0, 0.79)	0.46 (-0.02, 0.65)	0.69 (0.49, 0.93)	0.48 (0.19, 0.62)
3.D: Diffuse pleural thickening (yes/no)	0.71 (0.48, 0.92)	0.58 (0, 1)	0.65 (0, 1)	0.54 (-0.02, 0.79)	0.72 (0.49, 1)	0.57 (0.19, 0.79)

*Table entries are mean kappa values with the associated range across the six readers.
†Within format: intra-round kappa values.
‡Between format: intra-format kappa values.

DISCUSSION

The present study assessed the impact of radiographic image format on inter-reader and intra-reader agreement. Few of the inter-reader (Tables 2A, 2B) or intra-reader comparisons (Tables 3, 5) differed significantly, and most of the few cases that were statistically significant involved outcomes related to classification of image quality.

Zähringer et al compared the prevalence of abnormalities using HC images obtained with a selenium drum detector (Thoravision: Philips Medical Systems, Hamburg, Germany) to traditional FSR from 50 miners (exposed to uranium and quartz dust, not asbestos) interpreted by four readers using the ILO system (5). Agreement of readings in different formats was assessed via comparison of the percent agreement among readers, but no kappa coefficients or similar chance-corrected statistics were presented.

The study by Takashima et al compared storage phosphor computed radiographs, flat panel detector digital radiographs, and conventional FSR for detection of small parenchymal opacities (6). Large opacities and pleural abnormalities were not reported as outcomes, and all digital images were viewed as HC only. The study involved a total of 90 readings, based on 30 subjects and 3 readers. Each reader compared the three images for each subject side-by-side at the same time; images were not reread, so no intra-reader kappa coefficients could be calculated. It was not stated whether inter-reader kappa values were calculated using dichotomized small opacity profusion scores (ie, yes/no), small opacity profusion scores based on the 4-point ILO scale, or small opacity profusion scores based on the 12-point ILO scale, which hampers direct comparison with results in other studies. The reported kappa values were: storage phosphor computed radiographs (kappa = 0.64); flat panel detector digital radiographs (kappa = 0.62); and, conventional film-screen radiographs (kappa = 0.55). The authors commented that "Overall inter-reader agreement of the profusion of small opacities on radiographs was high," but no between-format statistical comparison of kappa values was presented. Though the conclusions of the authors appear to be in general agreement with the current study, the study by Takashima et al was limited in power (ie, fewer subjects, readers, and total number of readings), scope (no SC, no analyses related to large opacities or pleural abnormalities), and statistical analyses (no formal statistical comparisons of kappa values between image formats). Also, it is unclear how the side-by-side viewing methodology may have influenced results.

Disagreement among readers has been a feature of all studies that have assessed inter-reader or intra-reader agreement of ILO classifications. Beginning with studies by Musch et al, inter-reader agreement has usually been quantified using some version of kappa (16-22). Direct comparisons of kappa statistics between studies are often hampered because of differences in study design/structure, and the precise manner of calculation of kappa often has varied (eg, weighted versus unweighted, or collapsing ILO small opacity profusion scores into idiosyncratic categories for calculation of kappa). The

TABLE 4. Intra-Reader Comparison of Kappas (Unadjusted for Covariates)

	FSR Kappa	FSR SE	HC Kappa	HC SE	SC Kappa	SC SE	FSR vs. HC *P* Value	FSR vs. SC *P* Value	HC vs. SC *P* Value
Reader 1									
Image quality (Category 1 vs. 2–4)	0.30	(0.11)	0.41	(0.12)	0.61	(0.09)	.485	.020	.158
Image quality (Categories 1–2 vs. 3–4)	0.43	(0.18)	0.52	(0.10)	0.29	(0.18)	.626	.609	.248
Parenchymal abnormalities (yes/no)	0.70	(0.08)	0.79	(0.08)	0.67	(0.09)	.416	.780	.285
Large opacities (yes/no)	0.70	(0.11)	0.77	(0.10)	0.68	(0.10)	.673	.910	.476
Pleural abnormalities (yes/no)	0.55	(0.09)	0.48	(0.08)	0.62	(0.08)	.549	.467	.162
Costophrenic angle obliteration (yes/no)	0.48	(0.22)	0	NA	0	NA	NA	NA	NA
Diffuse pleural thickening (yes/no)	0.48	(0.22)	0	NA	0	NA	NA	NA	NA
Reader 2									
Image quality (Category 1 vs. 2–4)	0.76	(0.07)	0.63	(0.08)	0.68	(0.07)	.233	.431	.580
Image quality (Categories 1–2 vs. 3–4)	0.43	(0.16)	0.49	(0.11)	0.43	(0.14)	.731	.995	.661
Parenchymal abnormalities (yes/no)	0.74	(0.07)	0.77	(0.06)	0.74	(0.07)	.727	.990	.641
Large opacities (yes/no)	0.96	(0.04)	0.85	(0.07)	0.81	(0.09)	.113	.119	.723
Pleural abnormalities (yes/no)	0.54	(0.07)	0.64	(0.08)	0.67	(0.08)	.287	.174	.781
Costophrenic angle obliteration (yes/no)	0.73	(0.09)	0.72	(0.11)	0.72	(0.10)	.910	.920	.986
Diffuse pleural thickening (yes/no)	0.66	(0.12)	0.63	(0.13)	0.67	(0.13)	.843	.937	.813
Reader 3									
Image quality (Category 1 vs. 2–4)	0.56	(0.08)	0.55	(0.08)	0.23	(0.09)	.894	.007	.007
Image quality (Categories 1–2 vs. 3–4)	0.64	(0.15)	0.38	(0.12)	0.19	(0.19)	.154	.023	.340
Parenchymal abnormalities (yes/no)	0.84	(0.06)	0.79	(0.07)	0.87	(0.05)	.479	.632	.295
Large opacities (yes/no)	0.88	(0.06)	0.84	(0.06)	0.88	(0.06)	.394	.917	.669
Pleural abnormalities (yes/no)	0.67	(0.08)	0.85	(0.06)	0.60	(0.10)	.046	.502	.021
Costophrenic angle obliteration (yes/no)	0.85	(0.11)	0.59	(0.17)	0.65	(0.19)	.083	.259	.766
Diffuse pleural thickening (yes/no)	0.92	(0.08)	0.65	(0.19)	0.74	(0.18)	.113	.345	.740
Reader 4									
Image quality (Category 1 vs. 2–4)	0.49	(0.11)	0.07	(0.11)	0.25	(0.11)	.008	.098	.232
Image quality (Categories 1–2 vs. 3–4)	0.34	(0.11)	0.50	(0.10)	0.39	(0.10)	.315	.764	.336
Parenchymal abnormalities (yes/no)	0.75	(0.07)	0.72	(0.09)	0.71	(0.07)	.755	.605	.952
Large opacities (yes/no)	0.84	(0.08)	0.84	(0.07)	0.72	(0.10)	.948	392	.316
Pleural abnormalities (yes/no)	0.75	(0.07)	0.65	(0.08)	0.67	(0.09)	.254	.423	.830
Costophrenic angle obliteration (yes/no)	0.75	(0.14)	0.85	(0.15)	0.74	(0.18)	.625	.946	.305
Diffuse pleural thickening (yes/no)	0.74	(0.18)	1.00	NA	0.79	(0.20)	NA	.836	NA
Reader 5									
Image quality (Category 1 vs. 2–4)	0.55	(0.08)	0.61	(0.09)	0.67	(0.07)	.632	.284	.606
Image quality (Categories 1–2 vs. 3–4)	0.65	(0.19)	0.42	(0.14)	0.49	(0.31)	.268	.657	.807
Parenchymal abnormalities (yes/no)	0.77	(0.06)	0.77	(0.06)	0.79	(0.06)	.998	.822	.807
Large opacities (yes/no)	0.73	(0.13)	0.80	(0.10)	0.70	(0.14)	.689	.879	.533
Pleural abnormalities (yes/no)	0.83	(0.06)	0.59	(0.09)	0.84	(0.06)	.008	.868	.016
Costophrenic angle obliteration (yes/no)	0.79	(0.14)	0.20	(0.18)	0.79	(0.14)	.002	.998	.002
Diffuse pleural thickening (yes/no)	0.88	(0.11)	0.23	(0.20)	0.74	(0.18)	.001	.524	.021
Reader 6									
Image quality (Category 1 vs. 2–4)	0.55	(0.10)	0.46	(0.12)	0.48	(0.10)	.550	.599	.921
Image quality (Categories 1–2 vs. 3–4)	0.58	(0.19)	0.17	(0.11)	0.07	(0.05)	.059	.007	.379
Parenchymal abnormalities (yes/no)	0.81	(0.06)	0.65	(0.09)	0.61	(0.09)	.074	.053	.678
Large opacities (yes/no)	0.85	(0.06)	0.80	(0.07)	0.72	(0.10)	.532	.140	.513
Pleural abnormalities (yes/no)	0.79	(0.06)	0.74	(0.08)	0.76	(0.07)	.669	.734	.911
Costophrenic angle obliteration (yes/no)	0.83	(0.10)	0.78	(0.12)	0.68	(0.13)	.778	.321	.508
Diffuse pleural thickening (yes/no)	0.65	(0.19)	1.00	NA	1.00	NA	NA	NA	NA

FSR, film-screen chest radiography; SE, standard error; HC, hard copy; SC, soft copy.

Table values show kappa estimates for each image format (FSR, SC, and HC), the respective SE, and the *P* values for the pairwise comparisons for each image format separately for each of the six B readers.

NA = not available due to the fact that either (a) kappa = 1.00 (boundary value) or (b) all ratings in one of the rounds fell in a single category.

TABLE 5. Difference in Intra-reader Reliability across Image Formats (Covariate Adjusted)*

	FSR vs. HC	FSR vs. SC	HC vs. SC
1.A: Image quality (4-pt scale)	0.0180 (0.89)	−0.0009 (0.99)	−0.0189 (0.89)
1.A: Image quality (Category 1 vs. 2–4)	0.0351 (0.58)	0.1103 (0.05)	0.0752 (0.19)
1.A: Image quality (Categories 1–2 vs. 3–4)	−0.0100 (0.93)	0.2086 (0.07)	0.2186 (0.01)
2.A: Parenchy abnormalities (yes/no)	0.0165 (0.74)	0.0451 (0.17)	0.0286 (0.51)
2.B: Small opacities (4-pt scale)	0.0044 (0.90)	0.0261 (0.36)	0.0218 (0.49)
2.C: Large opacities (yes/no)	0.0768 (0.15)	0.1161 (0.11)	0.0393 (0.61)
2.C: Large opacities (4-pt scale)	0.0982 (0.046)	0.1502 (0.06)	0.0520 (0.48)
3.A: Pleural abnormalities (yes/no)	0.0428 (0.37)	−0.0738 (0.16)	−0.1166 (0.03)
3.C: Costophrenic angle obliteration (yes/no)	0.0851 (0.34)	0.0693 (0.39)	−0.0158 (0.81)
3.D: Diffuse pleural thickening (yes/no)	0.0596 (0.63)	0.0307 (0.78)	−0.0290 (0.83)

The model for small opacities based on the 12-point scale is not presented due to the sparsity of the underlying 12 × 12 matrix.

*Table entries are differences of the estimated regression model coefficients (*P* value) associated with image formats. All models are adjusted for covariates (age, gender, body mass index, pack-years of smoking, and individual B readers). Models other than image quality are also adjusted for median image quality.

multireader inter-reader kappa values for image classifications shown in Table 2A are not identical to, but would be most comparable to the multireader kappa results reported by Musch et al and Lawson (19-21). Overall, inter-reader agreement for FSR in the present study compares favorably to inter-reader agreement based on kappa reported in other studies.

Fewer published studies have evaluated intra-reader agreement of ILO interpretations. The report by Impivaara et al involved two readers and presented results of intra-reader agreement for small opacity profusion scores and pleural abnormalities based on FSR (18). Intra-reader kappa results were based on ratings by two readers of 597 subjects. For small opacity profusion rated on a four-point scale they achieved unweighted intra-reader kappa values of 0.22 and 0.27, and weighted intra-reader kappas of 0.55 and 0.54, respectively. However, the four-point scale used by Impivaara et al, for calculating weighted kappas corresponded to 0/0, 0/1, 1/0, and ≥1/1, rather than the four traditional major ILO profusion categories (0, 1, 2, and 3), which prevents direct comparison with our findings for small opacities. In the present study the results shown in Tables 4 and 5 for FSR for parenchymal abnormalities would be most comparable. Intra-reader agreement for the six B readers in the present study appears to compare favorably with the results reported by Impivaara et al (18).

There are a number of practical challenges related to the use of digital imaging for pneumoconiosis that are outside the scope of the present study. The present study captured digital images using a Hologic DR system, but there are other technologies in use (eg, computed radiography), and new ones are likely to evolve. It is unclear how different digital technologies may impact image appearance and interpretation. Most ILO classifications are not performed at the institution or facility where the radiographs were obtained, rather they are transferred, usually on CDs. There is considerable variability of viewing software among commercial CD programs (eg, use of image compression, tools for windowing or magnification). The digital versions of the ILO standard radiographs used in the present study are only available as a research tool; no digital ILO standards are available for everyday use. There is a lack of standardization of digital image numerical processing among systems. These are just a few examples of practical issues that need to be addressed before digital imaging techniques can fully integrated with use of the ILO system.

Overall, FSR, HC digital images, and SC digital images yield similar reliability measures for parenchymal and pleural changes, even when adjusted for reader and subject-level covariates. And, based on results of FSR, the level of inter-reader and intra-reader agreement among the six readers in the present study compares favorably with previous studies that have quantified reader agreement using kappa. In a previous study we found that: readings of FSR and SC images were equivalent for prevalence of small opacity profusion; readings of HC images yielded significantly greater prevalence of small opacities compared to FSR and SC (71% vs. 65% vs. 67%, respectively); and the prevalence of pleural abnormalities differed significantly among all three image formats, with FSR greater than HC greater than SC (37% vs. 31% vs. 27%, respectively) (7). These findings, taken with the current results, provide reassurance that use of high quality SC images with the ILO pneumoconiosis classification system will not result in any substantial change in the recognition or quantification of small pneumoconiotic opacities or the reliability of the ratings, compared to traditional film images.

ACKNOWLEDGMENTS

We wish to thank James Good and Janis Huff in the University of Michigan Radiology Department for their diligence, support, and good cheer. Emmanuel Christodoulous, PhD, of the University of Michigan Radiology Department assisted with the digitization of the ILO standard radiographs. We are grateful to Drs. Kurt Hering and Igor Fedotov at the International Labour Organization for giving permission to digitize the ILO standard films for use in this study. Mary Jo Reilly at the Michigan Silicosis Registry and Ed Socie at the Ohio

Silicosis Registry provided invaluable assistance with recruitment of subjects. We thank Bethany Baker for her commitment to this project, and for being the glue that helped to hold it together and keep it moving forward. Drs. Michael Attfield and Bill Miller at NIOSH reviewed the manuscript and provided many thoughtful and helpful suggestions.

REFERENCES

1. Musch DC. Interobserver variation in classifying radiographs for the pneumoconioses. Ann Arbor, MI: University of Michigan Doctoral Dissertation, 1981.
2. International Labour Organization (ILO). Guidelines for the use of ILO international classification of radiographs of pneumoconioses, Revised edition 2000. Geneva: International Labour Office, 2002.
3. Pham QT. Chest radiography in the diagnosis of pneumoconiosis. Int J Tuberc Lung Dis 2001; 5:478–482.
4. Mulloy KB, Coultas DB, Samet JM. Use of chest radiographs in epidemiological investigations of pneumoconioses. Br J Indust Med 1993; 50: 273–275.
5. Zähringer M, Piekarski C, Saupe M, et al. Comparison of digital selenium radiography with an analog screen-film system in the diagnostic process of pneumoconiosis according to ILO classification [in German]. Fortschr Röntgenstr 2001; 173:942–948.
6. Takashima Y, Suganuma N, Sakurazawa H, et al. A flat-panel detector digital radiography and a storage phosphor computed radiography: screening for pneumoconioses. J Occup Health 2007; 49: 39–45.
7. Franzblau A, Kazerooni EA, Sen A, et al. Comparison of digital radiographs with film radiographs for the classification of pneumoconiosis. Acad Radiol 2009; 16:669–677.
8. Rosenman KD, Reilly MJ, Kalinowski DJ, et al. Silicosis in the 1990's. Chest 1997; 111:779–786.
9. Reilly MJ, Rosenman KD, Watt FC, et al. Silicosis surveillance - Michigan, New Jersey, Ohio, and Wisconsin, 1987–1990. MMWR 1993; 42:23–28.
10. American Thoracic Society. Epidemiology standardization project: recommended respiratory disease questionnaires for use with adults and children in epidemiological research. Am Rev Resp Dis 1978; 118:7–53.
11. SAS Institute Inc., SAS/STAT user's guide, version 9.1. Cary, NC: SAS Institute Inc, 2002.
12. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960; 20:37–46.
13. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. New York: John Wiley & Sons, 2003.
14. Barnhart HX, Williamson JM. Weighted least-squares approach for comparing correlated kappa. Biometrics 2002; 58:1012–1019.
15. Lipsitz SR, Williamson J, Klar N, et al. A simple method for estimating a regression model for κ between a pair of readers. J Royal Stat Soc Series A 2001; 164:449–465.
16. Huuskonen O, Kivisaari L, Zitting A, et al. High-resolution computed tomography classification of lung fibrosis for patients with asbestos-related disease. Scand J Work Environ Health 2001; 27:106–112.
17. Impivaara O, Zitting AJ, Kuusela T, et al. Observer variation in classifying chest radiographs for small lung opacities and pleural abnormalities in a population sample. Am J Indust Med 1998; 34:261–265.
18. Lawson CC, LeMasters MK, Lemasters GK, et al. Reliability and validity of chest radiograph surveillance programs. Chest 2001; 120:64–68.
19. Musch DC, Landis R, Higgins ITT, et al. An application of kappa-type analyses to interobserver variation in classifying chest radiographs for pneumoconiosis. Stat Med 1984; 3:73–83.
20. Musch DC, Higgins ITT, Landis JR. Some factors influencing interobserver variation in classifying simple pneumoconiosis. Brit J Ind Med 1985; 42: 346–349.
21. Naidoo RN, Robins TG, Soloman A, et al. Radiographic outcomes among South African coal miners. Int Arch Occup Environ Health 2004; 77: 471–481.
22. Welch LS, Hunting KL, Balmes J, et al. Variability in classification of radiographs using the 1980 International Labor Organization Classification for Pneumoconioses. Chest 1998; 114:1740–1748.