PUBLIC HEALTH IMPLICATIONS OF THE VARIABILITY IN THE INTERPRETATION OF "B" READINGS

DAVID L. PARKER, M.D., M.P.H. • Alan P. Bender, Ph.D., D.V.M. • Anita Barklind, M.S.

Minnesota Department of Health, Section of Chronic Disease and Environmental Epidemiology
717 Delaware Street S.E., P.O. Box 9441, Minneapolis, MN 55440, USA

BACKGROUND

Despite the care that has been given to interpretation, competent observers have repeatedly encountered difficulties in the consistent classification of radiographs. Researchers have been aware of the variability present in the interpretation of chest radiographs for many decades. As early as 1947, Birkelo published a paper evaluating the prevalence of tuberculosis observed in chest radiographs.[1]

Shortly after this study was conducted, Garland published a classic study on the scientific evaluation of diagnostic procedures. In this paper, Garland states that "though useful when, as occasionally happens, the chest radiograph is used as the sole examination, its reliability may be evanescent." He goes on to say in "nearly every activity that can be tested, it has been repeatedly demonstrated that humans, even experts in a given field, exhibit enormous variations in their ability to be consistent with themselves and others equally competent in applying to mass-survey work. ... Consequently, every day persons throughout the country are being informed that their chests are free from disease when, in point of fact, they probably are not (and visa versa). This results in false security on the one hand and needless alarm on the other hand."[2]

The purpose of this paper is to discuss the public health implications of the reliability of the "B" reading program, that is, the ability of different "B" readers to accurately and consistently reproduce findings during repeated examination of radiographs of people with disease of known or unknown status. For example, when a problem with pneumoconiosis is suspected, films may be submitted to one or several readers for interpretation. If multiple readers agree, then it is likely that their interpretation is correct. It is possible that readers may agree and still be incorrect in their interpretation. If agreement is low, then the usefulness of the interpretation is suspect.

METHODS

The issues on which this paper is based arose from a call received by the Minnesota Department of Health (MDH) in January, 1985, from a "B" reader and radiologist (Reader 1) in northern Minnesota. This is an area that has historically had many problems related to asbestos in mine tailings. The radiologist stated he had found diffuse and/or circumscribed pleural thickening in approximately 30% of 500 sequential chest radiographs taken during the preceding two months in his clinic practice. Subsequent review of the films by MDH staff led to consultation with the National Institute for Occupational Safety and Health (NIOSH).

At the request of the MDH, a "B" reader (Reader 2) from NIOSH came to Minnesota to review these findings. Reader 2 reviewed 259 films interpreted by Reader 1 and 310 films from other regional clinics. Reader 2 confirmed the apparent increase in pleural changes seen by Reader 1 and noted similar increases in other regional clinics. Because of the confirmed increase, two additional radiographic evaluations were arranged.

Five hundred and sixty-six films were transported to Reader 3, a pulmonary physician and experienced "B" reader in New York City. Following this third reading, the films were shipped to NIOSH in Morgantown, West Virginia. At NIOSH, the films were randomly allocated in equal numbers among ten blocks. Negative and positive control films were added so there were 100 films per block. Positive control films were selected for the presence of pleural changes. Films were then interpreted independently by three readers who had been selected from a panel of five readers. The trial was a randomized incomplete block design with each of five readers being assigned six blocks and each film read a total of three times. Films were read in Morgantown, and readers were unaware of the origin of the films. Except for Reader 1, all readers interpreted the films according to the 1980 International Labor Office (ILO) classification system. Reader 1 interpreted films only for pleural changes.

ANALYSIS AND RESULTS

A kappa statistic was used to measure concordance between Readers 1 through 3. Concordance was not measured in this way between members of the NIOSH panel because of the large number of possible combinations. This statistic measures agreement between readers and simultaneously accounts for agreement due to chance. A kappa statistic is continuous and ranges between -1 and $+1$. A statistic of 0 or less represents poor agreement and a statistic of $+1$ reflects complete agreement.[3]

As seen in Table I, concordance between Readers 1 and 2 was moderate (kappa = 0.58) for the presence of any pleural thickening. Readers 2 (NIOSH consultant) and 3 (New York reader) agreed on 70 films being positive for pleural changes and the kappa statistic was 0.39, once again indicating a moderate degree of concordance (Table II). However, when

Table I
Presence of Pleural Thickening: Concordance Between Readers 1 and 2

Pleural Thickening (First Reader)	Pleural Thickening (Second Reader) Absent	Present	Total
Absent	97 (37.4)*	23 (8.9)	120 (46.3)
Present	32 (12.4)	107 (41.3)	139 (53.7)
Total	129 (49.8)	130 (50.2)	259 (100.0)

*Percent
KAPPA = 0.58

Table II
Presence of Any Pleural Changes: Concordance Between Readers 2 and 3

Pleural Thickening (Second Reader)	Pleural Thickening (Third Reader) Absent	Present	Total
Absent	360(63.5)*	13 (2.3)	364(65.8)
Present	119(21.5)	70(12.7)	189(34.2)
Total	470(85.0)	83 (15.0)	553**

* Percent
** Excludes 13 films rated as not readable.
KAPPA = 0.39

pleural plaquing and diffuse pleural thickening were examined by side (Tables III and IV), concordance was poor, with a kappa statistic of 0.26 and 0.20 respectively. Thus under more stringent criteria, agreement appeared to diminish considerably.

According to Readers 1 and 2, approximately 70% of males and 25% of females from Reader 1's clinic had pleural abnormalities. The proportion of males and females read as positive varied considerably between Readers 2 and 3. Overall, Reader 2 noted 54% of males and 15% of females had pleural abnormalities. Reader 3 found 25% of males and 5% of females had pleural abnormalities. The NIOSH readers found 8% of males and less than 1% of females positive for pleural changes. These differences between readers were statistically significant using McNemar's test.[3]

Table V shows the number of Minnesota films read as positive by zero, one, two, or three of the NIOSH (Morgantown) readers. A total of 24 (4.2%) of the Minnesota films were read as positive by at least two readers. In addition, the number of positive control films (n=34) read as positive (i.e., sensitivity) was approximately 55% but varied slightly from reader to reader. The number of negative control films (n=400) read as negative (i.e., specificity) was 98% or more for all readers.

Data from the control films were used to estimate the conditional probability of a film being positive given zero, one, two, or three positive readings (Table VI). The value for "Π" represents the approximate probability of a film being positive if it was drawn at random from the batch of all Minnesota films. The value for "p1" represents sensitivity and the value for "p2" represents specificity. For this trial, we see that the conditional probability of a film being positive for any pleural changes given zero, one, two, or three positive readings (under the conditions of this trial) was

Table III
Pleural Plaquing*: Concordance Between Reader 2 and Reader 3 Bilaterally

Pleural Plaquing (Second Reader)	Pleural Plaquing (Third Reader) None	Unilateral Left	Unilateral Right	Bilateral	Total
None	393	1	6	1	401
Unilateral Left	17	3	0	0	20
Unilateral Right	17	1	4	1	23
Bilateral	77	8	10	14	109
Total	504	13	20	16	553

* Only includes plaques noted on the chest wall
KAPPA = 0.26

Table IV
Diffuse Pleural Thickening: Concordance Between Reader 2 and Reader 3 Bilaterally

Diffuse Thickening (Second Reader)	Diffuse Thickening (Third Reader) None	Unilateral Left	Unilateral Right	Bilateral	Total
None	503	4	1	1	509
Unilateral Left	10	4	0	0	14
Unilateral Right	7	0	1	1	9
Bilateral	19	1	1	0	21
Total	539	9	3	2	553

KAPPA = 0.20

Table V

Number of Films with Zero, One, Two, or Three Positive Readings for Pleural Changes*

	Category of Film Reading			
	Zero	One	Two	Three
Number	436	44	14	10

*NIOSH readers only.
Sixty two unread films are not counted.
(N = 566, including unread films.)

Table VI

Probability of a Film Being Positive Given Zero, One, Two, or Three Positive Readings for Pleural Changes*

	Number of Positive Readings			
	0	1	2	3
$\Pi = 0.09$				
$P_1 = 0.55$	0.01	0.36	0.97	~1.0
$P_2 = 0.98$				

NIOSH readers only.

0.01, 0.36, 0.97, and approximately 1.0 respectively. It should be noted that values for Π, sensitivity, and specificity are dependent upon the mix of positive and negative radiograph readings.

In order to further evaluate the reasons for the variability observed in the NIOSH trial, logistic regression procedures were used with the absence and presence of pleural changes coded 0 and 1 respectively. Independent variables used in the prediction equation were age ($<60, \geq 60$), sex, parenchymal opacity profusion (two levels, $\leq 0/1$ and $\geq 1/0$), and the presence of other pulmonary abnormalities (two levels: none, any). The regression model fit well and there were no significant interaction terms. Assumptions required for logistic regression were satisfied. The summary odds ratios and 95% confidence intervals for each of five NIOSH readers are presented in Table VII. For example, Reader 1 was 5.5 times more likely to find evidence of pleural changes if the film being interpreted had evidence of parenchymal opacities of 1/0 or greater compared to films with opacities rated 0/1 or less. As seen in this table, age and sex did not influence radiograph interpretation for pleural changes. However, for some readers, profusion and/or the presence of other diseases

Table VII

Odds Ratio and 95% Confidence Intervals for Four Factors Used in Predicting the Presence or Absence of a Positive Reading for Pleural Disease

	Factor			
Reader	Age	Sex	Profusion	Other Abnormalities
1	NS*	NS	5.5 (1.6, 18.6)	4.6 (1.5, 13.9)
2	NS	NS	NS	NS
3	NS	NS	NS	3.4 (1.2, 10.0)
4	NS	NS	26.1 (7.7, 88.3)	NS
5	NS	NS	NS	4.7 (1.2, 18.2)

* Not significant.

appeared to exert a moderate to strong influence on the interpretation of films for the presence of pleural abnormalities.

DISCUSSION

Many studies have been published evaluating factors that affect the interpretation of radiographs. These factors include: film quality, subject age and weight, presence of disease, and reader.[4-10]

Liddell found that film quality tended to be higher for radiographs of men with no evidence of coal worker's pneumoconiosis and to decrease with increasing chest wall thickness. The subject's age was not found to substantially affect film quality.[5] Pearson et al. found that the proportion of unsatisfactory films increased with increasing values of the ratio of weight to sitting height.[8] These findings are of interest because it has been demonstrated that technical faults are, in general, randomly distributed and attributable to errors in taking and processing films rather than in differences between subjects even though there may be a slight tendency for the proportion of unsatisfactory films to increase with increasing weight.[8]

Further, Liddell found film quality introduced only slight biases into the reading of pneumoconioses although readers tended to find more parenchymal abnormalities in overexposed films and fewer parenchymal abnormalities in underexposed films when compared to good films.[5] Other investigations, however, have found that readers tend to read more abnormalities in underexposed films and less abnormality in overexposed films.[6,7] In Minnesota, film quality was adequate for all but a handful of radiographs. For this reason, it seems unlikely that film quality affected the results of the Minnesota study.

Reader experience also plays a role in the evaluation of radiographs. Different readers appear to compensate differently for changes in film quality. Reger et al. found that

experienced readers were better able to compensate for changes in film quality. In addition, certain readers either consistently find more abnormalities or less abnormalities on films compared with their colleagues.[9] Felson et al. found that readers with minimal training tended to find more cases of coal workers with pneumoconiosis than experienced readers. Felson attributed the differences between readers found in his study to several factors: 1) inherent interobserver disagreement; 2) lack of experience with the classification system in use; and 3) lack of familiarity with the radiographic manifestations of coal workers' pneumoconiosis.[10]

The problems encountered during the MDH investigation were in many ways similar to those described above. The percentage of films interpreted as abnormal varies among readers. These readers appeared to have been influenced by factors such as the presence of disease, and anecdotally, reader experience may have played a major role. The original readers were both newly certified "B" readers and were not experienced in interpreting films with asbestos-related disorders. These were also the readers who found the highest percent of individuals with pleural changes.

Two years after the original investigation was completed, eight radiographs that the investigators thought were "definitely positive" were sent to a pulmonary physician for review. After reviewing the medical records and films, this physician felt that the pleural and/or parenchymal abnormalities in six of the cases (75%) could be best explained by the presence of diseases unrelated to the pneumoconioses.

This finding, in part, led the MDH to once again evaluate the original data and develop the logistic regression model described above. This model confirms that the reading of radiographs for the presence of pleural abnormalities is at times strongly influenced by the presence of parenchymal opacities and/or diseases; however, it was not possible to define the nature of this relationship.

The magnitude of inter-reader agreement has undergone considerable scrutiny. Early studies on this problem were conducted by Birkelo, Garland, Fletcher, and Yerushalmy.[1,2,10,11] In 1970, Reger and Morgan had 2,337 radiographs evaluated by 4 readers. The percent of films interpreted as having complicated coal workers' pneumoconiosis ranged from 8.0% to 22.5%.[9] In only slightly more than one half (56.7%) of these films was there agreement between readers. Felson et al. evaluated inter-reader agreement for 3 readers. For films read as normal, pairs of readers agreed with each other between 10.1% and 68.9% of the time. For abnormal films, agreement ranged between 5.5% and 10.2%.[10]

Several studies have examined the variability in the radiographic assessment of pleural changes. In a review of 674 radiographs of naval dockyard workers, Sheers et al. found the prevalence of pleural changes to range between 14% and 30%.[12] Reger et al. evaluated inter-reader variability in the radiographic detection of pleural changes in 555 radiographs.[13] Radiographs were evaluated twice for each worker—first using a posterior-anterior (PA) film and then using PA plug oblique films. The prevalence of pleural abnormalities in this study ranged between 40% and 81% and a higher detection rate was found with the use of additional radiographs. Using PA films only, the kappa statistic for inter-reader agreement for the presence of pleural plaques averaged 0.33 and for diffuse pleural thickening 0.43. The addition of oblique films caused a decrement of the kappa statistic to 0.23 and 0.25 for pleural plaques and pleural thickening respectively.[13]

A higher detection rate of pleural abnormalities using three radiographs (left anterior, oblique, right anterior oblique and PA) compared with PA only was also shown by Baker and Green.[14] The high detection rate, however, appears to be at the expense of sensitivity, specificity and reliability.[13] The number of positive control films read as positive (i.e., sensitivity) was only 55% in the MDH study. It seems that any further decrement in sensitivity resulting from the use of oblique films would, in most instances, be unwarranted.

Green et al. examined the effect of using a broad (any pleural thickening) versus a strict criterion (pleural thickening of 2 mm or greater) on the prevalence of pleural changes in high risk (asbestos exposed) and low risk (no or little asbestos exposure) groups. Using a broad criterion, prevalence ranged from 45.1% (low risk) to 40.9% (high risk), and, using a strict criterion, prevalence ranged from 2.6% (low risk) to 9.4% (high risk).

Depending upon the number of positive readings and the readers selected, the percent of Minnesota films positive for pleural changes varied between 2% and 38% (Table VIII). Thus, we were faced with a problem where "case definition" was highly dependent upon the judgement of the investigators and it was not clear which was the best set of interpretations to use. We do not feel the results of the MDH study support the use of a specific (e.g., 2 mm) threshold criteria. However, we concur with the conclusion of Green et al. that there is a "great need for specific criteria and uniform methodology" in the interpretation of pleural findings.

The low sensitivity and high inter-reader variability present in the evaluation of films for asbestos-related pleural or parenchymal changes could significantly influence the results of an epidemiologic study. Readers 1 and 2 found a large

Table VIII

Number of Positive Pleural Readings by Sex
(N = 2755 Readings)*

Number of Positive Readings	Sex		Percent Of Total	Total
	Male	Female		
0	122	231	62	353
1	82	38	21	120
2	38	8	8	46
3	23	5	5	28
4	10	0	2	10
5	8	1	2	9
Total	283	283		566

* Five B Readers per film (10 not read by reader 2, 3 not interpreted by reader 3, and 62 not interpreted by members of the NIOSH panel.

number of abnormalities in both men and women indicating what appeared to be a generalized environmental exposure to asbestos. Subsequent investigation revealed widespread steam tunnels to many regional homes. These tunnels, as well as the pipes within homes, appeared to be asbestos-lined. Another possible source of exposure was piles of taconite mine tailings near or within town limits. Because of concern about environmental exposures, the third and subsequent readings were done. In later readings, when substantially fewer abnormalities were found in women, it was thought that the problem was probably occupational rather than environmental in origin.

It is felt that the low sensitivity of the interpretation of radiographic changes of the pleural should be more widely recognized among those involved in occupational disease surveillance. The impact of this variability in radiographic readings on public health decisions was illustrated in Minnesota and, to date, the significance of these apparent abnormalities is still difficult to evaluate.

Based on these findings and a review of the epidemiologic literature, we feel further consideration should be given to resolving the issues presented here. We would like to make the following recommendations to optimize the use of information found on the chest radiograph:

1. A threshold for determining the presence or absence of pleural changes should be developed. In part, the problem encountered by the MDH arose because of the ambiguity in defining pleural changes. Dr. E. Nicholas Sargent (personal communication) recommends the use of a scoring system similar to that used for parenchymal changes (e.g., 0/0, 0/1, 1/0, 1/1) with 0/0 indicating a high degree of certainty that a particular shadow does not represent a pleural abnormality (e.g., muscle, fat) and 1/1 indicating a high degree of certainty that a shadow does represent a pleural abnormality (e.g., plaque);

2. Experiments should be conducted in which the "B" reader is asked to interpret films with and without an abbreviated medical history. At the end of each reading, the interpreter should be asked to conclude if, given the patient's (worker's) medical history, any changes seen are most likely due to a pneumoconiosis, other disease, both, or if such a determination cannot be made;

3. It appears that the interpretation of pleural changes may be too complex. This complexity makes the interpretation of inter-reader agreement difficult. If possible, the reading of pleural changes should be simplified;

4. One third of the "B" reading form is devoted to interpreting changes of the pleura. However, there are very few films in the set of ILO standard films devoted to these changes. These films should be enhanced to reflect the degree and nature of changes that are presented on the ILO-NIOSH "B" reading form; and

5. "B" readers, in the course of their training, should be cautioned about the implications and utility of "B" reading. Knowledge of the problems involved in the epidemiologic use of radiographs should be a routine part of the "B" reader examination and/or course of study.

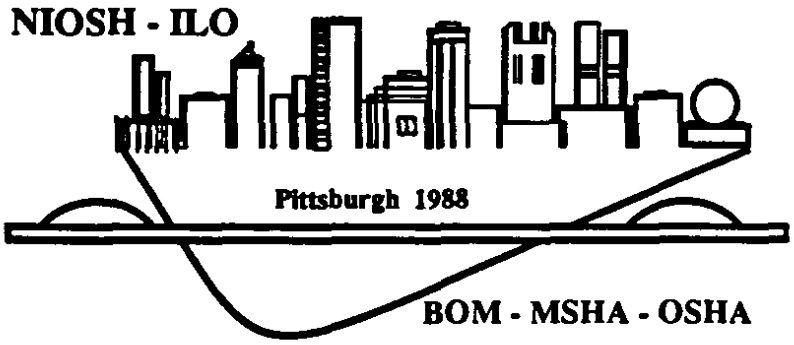
Inter-reader variability in the interpretation of radiographs has been evaluated in the past. This is the first instance known to the authors where this problem has had a direct impact on public health. When initially presented with this problem, the authors (DP and AB) consulted national experts on asbestos-related disorders; all agreed that we might have a major public health problem related to environmental asbestos exposure. As our investigation evolved, it appeared that this was not really an environmental problem at all, but was due to inter-reader variability in the interpretation of radiographs, thus substantiating previous studies on the problem of inter-reader variability.

REFERENCES

1. Birkelo, C.C., Chamberlain, W.W., Phelps, P.S., Schools, P.E., Zachs, D., Yerushalmy, J.: Tuberculosis case finding—A comparison of the effectiveness of various roentgenographic and photofluorographic methods. *J.A.M.A.* 133:359-367 (1947).
2. Garland, L.H.: On the scientific value of diagnostic procedures. *Radiology*. 52:309-327 (1948).
3. Fleiss, J.L.: *Statistical Methods for Rates and Proportions*. John Wiley & Sons, Inc., New York (1981).
4. Fletcher, C.M., Oldham, P.D.: The problem of the consistent radiological diagnosis in coal workers' pneumoconiosis. *Br. J. Ind. Med.* 6:168-182 (1949).
5. Liddell, F.D.K.: The effect of film quality on reading radiographs of simple pneumoconiosis in a trial of x-ray sets. *Br. J. Ind. Med.* 18:165-174 (1961).
6. Reger, R.B., Smith, C.A., Kibelstis, J.A., Morgan, W.K.G.: The effect of film quality and other factors on the roentgenographic categorization of coal workers' pneumoconiosis. *Am. J. Roentgenol.* 115:462-472 (1972).
7. Wise, M.E., Oldham, P.D.: Effect of radiographic technique on readings of categories of simple pneumoconiosis. *Br. J. Ind. Med.* 20:145-153 (1963).
8. Pearson, N.G., Ashford, J.R., Morgan, D.C., Pasqual, R.S.H., Rae, S.: Effect of quality of chest radiographs of coal workers' pneumoconiosis. *Br. J. Ind. Med.* 22:81-92 (1965).
9. Reger, R.B., Morgan, W.K.G.: On the factors influencing consistency in the radiographic diagnosis of pneumoconiosis. *Am. Rev. Respir. Dis.* 102:905-915 (1970).
10. Felson, B., Morgan, W.K.G., Bristol, L.J., Pendergrass, E.P., Dessen, E.L., Linton, O.W., Reger, R.B.: Observations on the results of multiple readings of chest films in coal workers' pneumonconiosis. *Radiology*. 109:19-23 (1973).
11. Yerushalmy, J.: Reliability of chest radiography in the diagnosis of pulmonary lesions. *Am. J. Surg.* 89:231-240 (1955).
12. Sheers, G., Rossiter, C.E. Gilson, J.C., Mackenzie, F.A.F.: U.K. Naval dockyards asbestos study: Radiological methods in the surveillance of workers exposed to asbestos. *Br. J. Ind. Med.* 35:195-203 (1978).
13. Reger, R.B., Ames, R.G., Merchant, J.A., Polakoff, P.P., Sargent, E.N., Silbiger, M., Whitlesey, P.: The detection of thoracic abnormalities using posterior-anterior (PA) vs PA and oblique roentgenograms. *Chest* 81:290-295 (1982).
14. Baker, E.L., Greene, R.: Incremental value of oblique chest radiographs in the diagnosis of asbestos-related pleural disease. *Am. J. Ind. Med.* 3:17-21 (1982).
15. Green, R., Boggis, C., Jantsch, H.: Asbestos-related pleural thickening: Effect of threshold criterion on interpretation. *Radiology*. 152:569-573 (1984).
16. Sargent, E.N., Boswell, W.D., Rall, P.W., Markovitz, A.: Subpleural fat pads in patients exposed to asbestos: Distinction from non-calcified pleural plaques. *Radiology*. 152:273-279 (1984).

Proceedings of the VIIth International Pneumoconioses Conference | *Part*
Transactions de la VIIe Conférence Internationale sur les Pneumoconioses | *Tome*
Transaciones de la VIIa Conferencia Internacional sobre las Neumoconiosis | *Parte*

I



Pittsburgh, Pennsylvania, USA—August 23–26, 1988
Pittsburgh, Pennsylvanie, Etats-Unis—23–26 aout 1988
Pittsburgh, Pennsylvania EE. UU—23–26 de agosto de 1988

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control
National Institute for Occupational Safety and Health

CDC
CENTERS FOR DISEASE CONTROL

Sponsors

International Labour Office (ILO)
National Institute for Occupational Safety and Health (NIOSH)
Mine Safety and Health Administration (MSHA)
Occupational Safety and Health Administration (OSHA)
Bureau of Mines (BOM)

September 1990

DISCLAIMER

Sponsorship of this conference and these proceedings by the sponsoring organizations does not constitute endorsement of the views expressed or recommendation for the use of any commercial product, commodity, or service mentioned.

The opinions and conclusions expressed herein are those of the authors and not the sponsoring organizations.

DHHS (NIOSH) Publication No. 90-108 Part I