March 1992

Classification of chest radiographs for pneumoconiosis: a comparison of two methods of reading

D. C. F. Muir
McMaster University, Hamilton, Ontario

Charles D. Bernholz
University of Nebraska-Lincoln, cbernholz2@unl.edu

W. K. C. Morgan
University of Western Ontario

J. O. Roos
Ministry of Labour, Toronto, Ontario, Canada

J. Chan
Ministry of Labour, Toronto, Ontario, Canada

See next page for additional authors

Classification of chest radiographs for pneumoconiosis: a comparison of two methods of reading

D C F Muir, C D Bernholz, W K C Morgan, J O Roos, J Chan, W Maehle, J A Julian, A Sebestyen

Variability between readers in the evaluation of the radiographic appearances of pneumoconiosis has been investigated by several authors.[1–3] To reduce inter and intrareader variability, standard reference films were introduced for epidemiological purposes.[4] Improved films and methods of classification were published in subsequent years.[5,6] The category of profusion of small opacities is determined by considering the profusion as a whole over the affected zones of the lung, and by comparing this with the standard radiographs. This requires a mental process of integrating profusion over the affected zones.

The purpose of the present study was to determine whether there would be any advantage, in terms of improved reader agreement, in reading and reporting the six lung zones separately rather than integrating all readings to produce a single classification for each film. The possible validity of evaluating the six zones separately was suggested by a previous study of the relation between radiographic appearances and underlying pathology in which a good correlation was found (kappa = 0·47, $p < 0·0001$).[7]

An opportunity arose to investigate this question during an epidemiological study of hard rock miners in Ontario.[8]

Material

The original cohort consisted of 2109 hard rock miners in Ontario. The criteria for selection have been reported elsewhere.[8] After preliminary examination 205 films from 48 miners were submitted in random order to five readers working independently. Each of the readers separately classified all six zones of the lung fields according to the prevalence and type of opacity using the International Labour Office (ILO) 1980 standard reference films.[6] These 205 films were again presented in random order to the five readers who classified the films according to the normal ILO (1980)[6] instructions so that each reader produced a single overall classification for each film. Several months elapsed between each reading trial.

Occupational Health Program, McMaster University, Hamilton, Ontario, Canada
D C F Muir, C D Bernholz, J A Julian, A Sebestyen
Department of Respiratory Diseases, University of Western Ontario, London, Ontario, Canada
W K C Morgan
Ministry of Labour, Toronto, Ontario, Canada
J O Roos, J Chan, W Maehle

Statistical analysis and results

The overall distribution of abnormality in the film set was estimated by averaging (after conversion to a 12 point numerical scale: 0/− = 1, 0/0 = 2, 0/1 = 3, . . . 3/+ = 12) the five independent classifications for each whole film. Table 1 shows the results. The average degree of abnormality in each zone was estimated in the same way. The mean values were between 0/1 and 1/0 and there was no obvious gradient in the mean value in each of the different horizontal lung regions. The mean values were 3·55, 3·70, and 3·32 for the upper, middle, and lower zones respectively. The distribution of round and irregular opacities was examined by summing the number of occasions when each reader recorded primary or secondary small opacities. The number of small, round opacities were 1451, 1259, and 822 in the upper, middle, and lower zones respectively and the number of small, irregular opacities were 317, 747, and 851 respectively. It is evident that round opacities predominated in the upper zones whereas round and irregular opacities were equally represented in the lower zones.

Two approaches were used to compare the agreement between readers when using the two different methods of classifying films.

In the first method, the readings for the whole lung were converted to the 12 point scale.[6] The variation of the five reader classifications recorded for each film was expressed by calculating the mean absolute deviation. The mean absolute deviation for each of the six zones on a film was calculated in the same manner. Finally, the average value for these six zonal scores was calculated. In this way an estimate of the

Table 1 The distribution of classifications in the set of 205 films

	ILO classification											
	0/−	0/0	0/1	1/0	1/1	1/2	2/1	2/2	2/3	3/2	3/3	3/+
Numerical scale	1	2	3	4	5	6	7	8	9	10	11	12
Number of films	0	5	58	85	27	17	6	6	1	0	0	0

overall variability associated with zonal classifications on a single film was obtained.

The next step was to compare the mean absolute deviations associated with each film by the two different reading methods. For this purpose, the non-parametric Wilcoxon matched pairs signed rank test[9] (adjusted for ties) was performed on the 205 pairs of mean absolute deviation scores from the whole lung readings, and the averaged zonal deviations. A comparison of the two methods resulted in a mean difference of −0·147, a 95% confidence interval of −0·213 to −0·093, and a significant difference ($p < 0·0001$). The results of this analysis are compatible with the hypothesis that less variability occurs between readers when classifying separate zones than when assigning an overall classification to the same films.

An alternative approach employing an unweighted kappa statistic was also used. This method makes no assumptions about the width of the categories. The generalised kappa analysis for multiple classifications by multiple readers described by Fleiss[10] was suitable. A program for the analysis capable of accepting up to 2000 sets of up to 20 classifications by any number of readers was prepared. There were 205 sets of whole film readings available for analysis and 192 sets for individual zones. Table 2 shows the results of the analysis.

Overall agreement among observers when reading individual zones was better than when assigning a single classification to each film. A formal test of the differences between these two kappa values gave a significant Z value of 6·305 ($p < 0·001$). This result is compatible with the hypothesis of a real difference in agreement between readers when using the two methods of reading films.

Discussion

Variability between classifications has previously been studied by considering agreement or lack of agreement between readers. This may be either in the simple form of percentage agreement or by means of chance corrected kappa statistics. In the present investigation we used two methods of comparing the amount of agreement between readers. The first approach may be regarded as a development of the classic paper by Fletcher and Oldham.[1] They quantified disagreement between readers by quoting the extreme range of opinions associated with a single film. We have extended this by calculating the mean of the absolute deviations from the mean for that film. A matched pairs test was used for comparing the deviations in the two methods of classifying radiographs so that allowance for individual film quality and level of abnormality was taken fully into account. A non-parametric test (Wilcoxon)[9] was chosen so as to avoid assumptions about the normality of the distributions that are required by the matched paired *t* test.

The second approach to examining agreement between readers using the two methods of classifying the radiographs assumes that the ratings are categorical. Several indices have been introduced for this purpose, the simplest being a straightforward statement of the proportion of agreement between two readers. Cohen[11] introduced the kappa statistic, which makes allowance for the proportion of agreement that would have been expected on the basis of chance alone and extended the kappa measure to attribute relative weights to the possible disagreement between raters.[12] Fleiss (1971)[13] extended the test further to include multiple ratings of the same set of subjects by the same raters and this forms

Table 2 Results of the agreement analysis

	No	*Generalised kappa*	*Standard error of kappa*	*Z score (H_0: kappa = 0)*
Whole lung	205	0·080303	0·01056	7·60118
Zones:				
Right upper	192	0·242521	0·01166	20·79929
Left upper	192	0·268191	0·01214	22·08723
Right middle	192	0·123671	0·01150	10·75401
Left middle	192	0·116720	0·01181	9·88205
Right lower	192	0·078224	0·01222	6·40002
Left lower	192	0·059931	0·01269	4·72100
All zones combined	1152	0·153920	0·00487	31·61346

the basis of the present study. An unweighted kappa was used in the analysis as the methodology of a weighted kappa for multiple ratings by several readers has not been developed.

The results are compatible with the hypothesis that overall agreement between readers is better when they read at a zonal level and that greater disagreement occurred when they assign a single overall classification to each film. The mean level of abnormality in the horizontal zones was nearly constant and the distribution of whole film classifications was such that most of the films were in a narrow range between categories 0/1 and 1/1. For this reason it was not possible to study the effect of degree of radiological abnormality on agreement between readers.

Limitations of the kappa statistic include its sensitivity to the prevalence of abnormality and to the number of categories into which the films are classified.[13 14] To illustrate the sensitivity of the kappa statistic to the number of categories, the whole film classification kappa analysis was repeated while the number of categories was collapsed from a 12 point scale to the simplified four point scale. The results of the whole film classification using the extended and the simplified scale were:

	Generalised kappa	Standard error
Twelve point scale	0·08030	0·01056
Four point scale	0·16194	0·01855

The increase in the value of kappa is evident.

Our results show better overall agreement between readers when classifying individual zones than when carrying out the mental process of integrating all zones so that a single classification is obtained for each film. There was a gradient in agreement from the upper to the lower zones. This is not likely to be related to gradients in the severity of abnormality in the zones as the overall level was almost constant. Round opacities were more preponderant in the upper zones, however, than in the lower zones where round and irregular opacities were equally represented. Observer agreement for round opacities has been reported as being better for round opacities[15] and this seems to be the likely explanation for our findings. We cannot exclude the possibility that opacities are most easily evaluated when they are in the apices.

Most of the films represented the earliest stages of pneumoconiosis. This is perhaps the most important stage from the clinical and epidemiological point of view. It will, however, be important to determine the extent to which our findings can be generalised to more advanced categories.

The practical utility of any improvement in agreement is not easy to quantify. It is important to note that the ILO system is essentially an epidemiological tool and was not designed for the purposes of establishing a diagnosis in individual cases. In the individual person, it may be more relevant to compare sequential films side by side rather than on a randomised basis.

It would also be necessary to consider the extra time and effort involved in reading at the zonal level.

The ultimate test of utility must be to determine whether zonal classifications correlate better with estimated exposure to dust or physiological parameters than do whole film readings. At this stage it is concluded only that classification of single zones is associated with better agreement between readers and may offer benefits for epidemiological investigations.

1 Fletcher CM, Oldham PD. The problem of consistent radiological diagnosis in coal miners pneumoconiosis. *Br J Ind Med* 1949;6:168–83.
2 Ashford JR. A problem of subjective classification in industrial medicine. *Applied Statistics* 1959;8:168–85.
3 Fay JWJ, Ashford JR. The study of observer variation in the radiological classification of pneumoconiosis. *Br J Ind Med* 1960;17:279–92.
4 Fletcher CM, Oldham PD. The use of standard films in the radiological diagnosis of coal workers pneumoconiosis. *Br J Ind Med* 1951;8:138–49.
5 International Labour Office. Classification of radiographs of the pneumoconiosis. *Occupational Health and Safety* 1959;9: 63–70.
6 International Labour Office. International classification of radiographs of pneumoconiosis. Geneva: ILO, 1980.
7 Verma DK, Muir DCF, Stewart ML, Julian JA, Ritchie AC. The dust content of the lungs of hard rock miners and its relationship to occupational exposure, pathological and radiological findings. *Ann Occup Hyg* 1982;26:401–9.
8 Muir DCF, Shannon HS, Julian JA, Verma DK, Sebestyen A, Bernholz CD. Silica exposure and silicosis among Ontario hard rock miners. I. Methodology. *Am J Ind Med* 1989;16: 5–11.
9 Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1945;1:80–3.
10 Fleiss JL. Measuring nominal scale agreement among many raters. *Psych Bull* 1971;76:378–82.
11 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
12 Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psych Bull* 1968;70:213–20.
13 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epididemiol* 1987;126:161–9.
14 Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;41:949–57.
15 Blair WG, Cochrane AL, Dick JA, *et al.* Radiological classification of the pneumoconiosis. *Arch Environ Health* 1966; 12:314–30.

Accepted 17 February 1992