A COMPARISON OF SOME ALTERNATIVE PROCEDURES IN THE CLASSIFICATION OF CHEST RADIOGRAPHS FOR COAL-WORKERS' PNEUMOCONIOSIS

BY

S. RAE, J. R. ASHFORD, D. C. MORGAN, R. S. H. PASQUAL, and N. G. PEARSON

From the National Coal Board Pneumoconiosis Field Research

(RECEIVED FOR PUBLICATION JANUARY 12, 1963)

This paper is concerned with the problem of reading individual chest radiographs for coal-workers' pneumoconiosis in terms of the I.L.O. (1953) classifications, using a group of four readers. It represents the first of a series of investigations by the National Coal Board's Pneumoconiosis Field Research into various aspects of the general subject of the classification of chest radiographs. Later papers in this series will cover related problems, including the effect of radiographic "technique" on reading standards, the classification of a series of radiographs of the same individual, and the use of the readings in the correlation of radiographic abnormality with environmental exposure.

The effect of the circumstances under which the readings must be made in the Pneumoconiosis Field Research, including the geographical separation of the four readers into two groups of two, is described, and it is shown that any acceptable procedure must involve two distinct stages. On the "first stage" every radiograph must be read by some or all of the doctors individually, and on the "second stage" a proportion of films (those on which the individual readings are "ambiguous") must be classified by the four doctors reading together. The general conduct of the joint reading sessions is then considered, and a description is given of a trial which was held to compare some alternative procedures. On the basis of these results an optimum reading procedure is derived. This is as follows:

(i) Every radiograph to be read independently by one or other of the doctors from each group, in such a way that each reads half of the total number of films. The doctors from the group producing the radiograph to read alternately consecutive batches of about 50 as they are taken during the survey; the doctors from the other group to read the first and second halves of the survey, respectively, the films being arranged in serial order.

(ii) If these two independent readings are consistent, the common reading to be taken as definitive. All other films to be classified by joint consultation between the four doctors reading together, their assessment being arrived at by general discussion with each of the readers in turn giving the first opinion of the category of the film.

One of the most important aspects of the National Coal Board's Pneumoconiosis Field Research (Fay, 1957; Fay and Rae, 1959) is the classification of the chest radiographs obtained on the successive medical surveys. A total of some 30,000 men is examined at intervals of about four years, and the way in which the films are read is of fundamental importance.

At the start of the Research in 1953 studies of observer error in classifying chest radiographs had already been published (Fletcher and Oldham, 1949; Fletcher and Oldham, 1951; Cochrane and Garland, 1952; Garland, 1949). These showed how readers could differ both with themselves and with each other in repeat readings of the same films.

Detailed descriptions have since been published of the film reading procedures adopted in the Pneumoconiosis Field Research (P.F.R.) and of the performance of the two film readers during the first six years of the Research (Fay and Ashford, 1960; Ashford, 1960). The analysis of their readings
revealed the presence of systematic variations in individual standards. When one of the original readers left in 1959, it was decided to share the responsibility for the classification of the films amongst a larger group, and the appointment of three new doctors brought the number of readers available to four.

An appraisal was then made of the possible ways of classifying the chest radiographs using the four readers, with the object of obtaining the best possible estimate of the category of the individual films, using the I.L.O. (1953) system of classification. The choice of the procedure to be followed was restricted by a number of considerations. In the first place, because two of the doctors work in Edinburgh and two in Cardiff, the classification of every radiograph by the four doctors reading together is deemed impracticable. On the other hand, the possibility of systematic differences in reading levels between the doctors working in Cardiff and those in Edinburgh, which had been noted previously (Ashford, 1960), means that regular joint reading sessions by the two groups are essential in order to achieve a uniform standard of classification. Thus any effective reading procedure must consist essentially of two stages, on the first stage every radiograph being read by the doctors in the two groups separately, and on the second stage a proportion of the radiographs (i.e. those on which the first-stage readings are “ambiguous”) being classified by the four doctors reading together.

As far as the first-stage readings by the two doctors in each group are concerned, there are two possible methods of approach; either the two doctors can read together as a pair or, alternatively, they can read independently. Following a preliminary period of training, a reading trial was held to examine the standards of the three newly-appointed doctors as compared with each other and with the survivor from the original pair of doctors (Fay and Brown, 1960). In each group there was close agreement between the standards of the two readers, and it was concluded that a joint reading by both the doctors working together would provide little additional information in comparison with a reading by either of the doctors independently. For this reason, and in order to monitor the performance of the individual readers, the method of independent readings was preferred to that of paired readings.

At this point three main questions remained to be settled: (1) the optimum number of independent first-stage readings by the various doctors, (2) the method of selecting a proportion of the radiographs for second-stage reading by the four doctors jointly, and (3) the best way of conducting these joint reading sessions. In order to resolve these three problems it was decided first to compare, by means of a special film reading trial, various alternative methods of joint reading with independent readings by the four doctors. Having determined the best method of conducting the joint reading sessions, the remaining uncertainties about the film reading procedures could then be settled.

This paper is concerned solely with the classification of single radiographs in terms of the I.L.O. (1953) classification. It represents the first of a series of investigations by the Pneumoconiosis Field Research into the general subject of the classification of chest radiographs for coal-workers’ pneumoconiosis. Later papers in this series will cover related problems, including the effect of radiographic “technique” on reading standards, the classification of a series of radiographs of the same individual for progression studies, and the use of the readings in the correlation of radiographic abnormality with environmental exposure.

The Design of the Trial

The object of the trial was to determine which of a number of alternative procedures for conducting the joint reading sessions leads to the best estimate of the category of a single radiograph. Of the four doctors concerned, two (readers W and X) work in Edinburgh and two (readers Y and Z) in Cardiff. Reader W has a wide experience of the classification of radiographs for coal-workers’ pneumoconiosis and has been actively concerned with the subject since the beginning of the Research in 1953. Readers X, Y, and Z are less experienced and were introduced to the system of classification by reader W. The categories of pneumoconiosis used in the trial (I.L.O., 1953) recognize a distinction between simple and complicated pneumoconiosis. Simple pneumoconiosis is subdivided into four categories increasing in degree of abnormality and labelled 0, 1, 2, and 3. Films showing evidence of complicated pneumoconiosis are separately classified in terms of three categories, A, B, and C.

The following three possible procedures to be followed at the joint reading sessions were put forward. In no case were the previous readings to be made available to the participants.

(a) The “Informal” Method.—This involves a discussion of the appearance of each film (including the categorization) without following any formal procedure, the final reading being arrived at by general agreement.

(b) The “Taking Turns” Method.—This involves the discussion of the appearance of each film, but without reference to the categorization. This discussion is followed by each of the four doctors taking
Classification of Chest Radiographs for Pneumoconiosis

It was decided that it would be desirable to have two sets of readings on each film, the first set being carried out by the observers independently, and the second set as a joint discussion by all the observers. Thus no time would be wasted in the event of disagreement as to which classification was correct.

The trial involved the classification of a sample of films first by each of the doctors independently (the "first independent" readings), then by joint discussion by the three procedures described above, and finally again by each of the doctors independently (the "second independent" readings). In order to obtain a measure of the reproducibility of the various methods, each of these five sets of readings was carried out in duplicate. To guard against the possibility of the results being affected by any tendency for the readers to remember their previous readings, and at the same time to keep the number of films involved within reasonable limits, it was decided to carry out the trial in two stages, each involving a separate sample of 250 films. On the first stage the first method of joint discussion to be investigated was the "secret" method, followed by the "taking turns" method and the "informal" method. On the second stage this order was reversed.

The selection of the sample films for the trial was governed by the need to obtain the greatest amount of information about any possible differences in reading standards between the various procedures with the minimum number of films. This meant that, as far as possible, the sample should consist of "difficult" films which are on the boundary between two categories, so that even a small change in standards would lead to a difference in the category assigned to the film. At the same time, it was considered necessary to include at least a proportion of "easy" films, so that the material used in the trial should not differ too widely from that encountered on more representative series of radiographs. In the event, it was decided that a fair balance would be achieved by selecting half of the sample at random from the films obtained on the P.F.R. medical surveys on which the original independent readings were inconsistent and the remainder from the films on which the original readings were consistent. In this way the composition of the sample was deliberately biased towards the more difficult type of film, whilst also including a reasonable proportion of more typical films.

Each set of duplicate readings by individual doctors and each duplicate joint discussion session was carried out within the space of two successive days. On each stage of the trial the full series of joint discussion readings took place during a single week, and all the readings were completed within a period of one month. There was an interval of one month between the end of the first stage of the trial and the beginning of the second. On each set of readings the films were arranged in a random order, and the classifications were made without reference to previous readings on the same film.

Results

The following sets of duplicate classifications are available on each film:

1. the "first independent" readings by readers W, X, Y, and Z (4 sets)
2. the "secret" discussion readings by readers W, X, Y, and Z (4 sets)
3. the "majority" readings based on (2) giving each of readers W, X, Y, and Z the casting vote (4 sets)
4. the "taking turns" discussion readings (1 set)
5. the "informal" discussion readings (1 set)
6. the "second independent" readings by readers W, X, Y, and Z (4 sets).

In the first place the numbers of films assigned to the various pneumoconiosis categories on corresponding duplicate readings were compared. In no instance was there any evidence of a significant change in standards between the original and repeat classifications. This was not surprising as the duplicate readings by each of the procedures were completed within a space of two days. The mean level on the two sets of readings has therefore been taken as a measure of the average standard of classification associated with each of the procedures under investigation, and the proportion of films on which the first and second readings were consistent has been adopted as the most convenient measure of reproducibility.

The average standards applied on the various sets of readings on the first and second stages of the trial are summarized in Tables 1(a) and 1(b), respectively. These tables are set out in the order in which the readings were made, those in the upper part of the table having taken place at an earlier stage of the trial than those in the lower part.

*At least two such readings are available on each film.
On the "first independent" readings on the first stage of the check there was a marked difference in standards between readers W and Z (Edinburgh unit) on the one hand and readers Y and Z (Cardiff unit) on the other; the corresponding average numbers of films classed as category O were 191-196 and 136-5-139. On the "secret" discussion readings, which were the first to be carried out by the four doctors in consultation, there is evidence of a considerable reduction in the amount of disagreement, although there is still a definite distinction between readers W and X and readers Y and Z. It is interesting to note that the smallest change in standards in comparison with the "first independent" readings was recorded by reader W, who is the most experienced member of the group. The "majority" readings derived from the "secret" discussions showed slightly smaller differences than the basic readings. On the "taking turns" and "informal" discussion readings only one opinion was obtained from the whole group, and the standards applied on these two sets of readings were very similar. The "second independent" readings were carried out immediately after the joint discussion sessions, but the results recorded by the individual readers reverted to the pattern of the earlier readings, with readers W and X reading considerably less abnormality than readers Y and Z. The reading levels of readers W and X returned to their original position, whereas those of readers Y and Z remained rather closer to
CLASSIFICATION OF CHEST RADIOGRAPHS FOR PNEUMOCONIOSIS

297

TABLE 2

COMPARISON OF THE CONSISTENCY OF THE VARIOUS SETS OF DUPLICATE READINGS

(a) Stage 1

<table>
<thead>
<tr>
<th>Method of Classification</th>
<th>First Independent</th>
<th>Joint Discussion</th>
<th>Second Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual Readings</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>W, X, Y, Z</td>
<td>W, X, Y, Z</td>
<td>W, X, Y, Z</td>
</tr>
<tr>
<td>Percentage of consistent readings</td>
<td>88, 94, 84, 84</td>
<td>91, 88, 88, 86</td>
<td>90, 89, 91, 90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>W, X, Y, Z</td>
<td>W, X, Y, Z</td>
</tr>
<tr>
<td></td>
<td></td>
<td>All, All</td>
<td>W, X, Y, Z</td>
</tr>
</tbody>
</table>

(b) Stage 2

<table>
<thead>
<tr>
<th>Method of Classification</th>
<th>First Independent</th>
<th>Joint Discussion</th>
<th>Second Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual Readings</td>
<td>Informal</td>
<td>Taking Turns</td>
</tr>
<tr>
<td>Percentage of consistent readings</td>
<td>93, 94, 85, 85</td>
<td>86, 94</td>
<td>93, 95, 92</td>
</tr>
</tbody>
</table>

the levels recorded on the “taking turns” and “informal” joint discussion readings.

The results of the second stage of the trial (Table 1(b)) followed a very similar pattern to those of the first stage, with the doctors from Cardiff tending to read more abnormality than those from Edinburgh. As before, the differences were least marked on the “secret” discussion readings.

The consistency of the results obtained on the various sets of duplicate readings on the first and second stages of the trial is summarized in Tables 2(a) and 2(b), respectively. On the first stage of the trial the percentage of consistent readings by individual doctors on the “first independent” readings varied from 84% to 94% with a mean of 87.5%. In comparison with the “first independent” readings, readers W, Y, and Z were slightly more consistent in the “secret” discussion. The “secret” discussions were preceded by the “taking turns” and “informal” discussions, on which the proportions of consistent readings were respectively 88% and 91%. On the “second independent” readings there was a marked difference in the reproducibility of the opinions of the four doctors, and the proportions of consistent readings varied from 84% to 97% with a mean of just over 91%.

On the second stage of the trial (Table 2(b)) the general trends were similar. The “majority” based on “secret” discussion readings were more consistent than the individual readings from which they are derived. Amongst the various sets of readings by the individual doctors, readers W and X were, on the whole, more consistent than readers Y and Z. This difference between the two units may be due at least in part to the fact that, on average, readers Y and Z read more abnormality than readers W and X. On each stage the “second independent” readings were rather more consistent than the “first independent” readings, which confirms the observations made by the doctors during the course of the trial that they became progressively more familiar with the test films. As far as the various methods of joint discussion are concerned, there were no consistent differences between the procedures, possibly as a result of a slight “memory effect”. It is, however, interesting to note that on both sets of films the differences in reproducibility between the various procedures were relatively small.

Discussion

A number of conclusions emerge very clearly from the trial. First, there is a small but systematic difference in standards between the doctors belonging to the two groups, although within each group the reading levels of the two doctors are in very close agreement. Even during the “secret” discussion sessions, when there was a completely free discussion of all aspects of the film except the actual category, the difference between the two units was still present, although its magnitude was greatly reduced. Frequent discussions, such as occur between the doctors working in the same unit, appear, however, to lead to very close agreement between the standards of the readers concerned.

Secondly, it was found that the results achieved on the “taking turns” and “informal” joint discussion
sessions were roughly midway between the reading levels of the two pairs of readers on the original "independent" readings. On the other hand, on the final set of "independent" readings, readers W and X tended to return to their original standards, whereas readers Y and Z remained rather closer to the levels applied in the joint discussion sessions. This confirms that the joint discussions do result in a fair compromise between the conflicting standards of the two pairs of readers.

In spite of the evidence of a slight "memory effect" during the course of the trial, the differences in the consistency of the various independent and joint discussion procedures were surprisingly small, so much so that reproducibility does not appear to be an important factor in distinguishing between them. This means that the main criterion must be the absence of bias in the average reading levels. On this basis the "informal" and "taking turns" discussions appear to be superior. Tables 1 and 2 show that there is little to choose between these alternatives, but the readers found the "taking turns" method to be rather more congenial in practice.

As far as the independent first-stage readings are concerned, the presence of a systematic difference in reading levels between the Edinburgh and Cardiff groups makes it necessary to restrict the choice to procedures based on an equal number of readings by each group. On the other hand, the good agreement between doctors belonging to the same group means that within each group the readers may be regarded as interchangeable. The total amount of film reading which can be carried out is limited by the number of doctors available, and the maximum number of first-stage readings which could possibly be undertaken is four, one by each reader. The only alternative system of "first-stage" reading is one based on two readings, one from each group.

With only two first-stage readings the only reasonable basis for selecting films for second-stage reading by the four doctors in consultation would be that of disagreement between them. This statement depends on the assumption that any possible differences between the reading levels of the two groups are not as great as a whole category on a high proportion of the readings. Our experience confirms that such differences as do exist are much smaller than this (Ashford, 1960). In the case of four first-stage readings, however, the proportion of disagreements would be considerably higher, but the amount of joint reading could be reduced to acceptable limits by restricting second-stage reading to those films on which the four readings are either equally divided between two categories or which cover a range of more than two categories. For these reasons two possible methods were studied further, using the techniques described by Fay and Ashford (1960). These are as follows:

(a) Method P.—Each film to be read independently by one or other doctor from each group. Where these two first-stage readings are consistent, the common reading to be taken as the final or definitive classification. All other films to be selected for "second-stage" reading, and the "definitive" classification to be assigned on the basis of the joint reading by the four doctors.

(b) Method Q.—Each film to be read independently once by each of the four doctors. Where the four readings are consistent, the common reading to be taken as definitive. Where three of the readings are consistent and the fourth differs by one category, the majority reading to be taken as definitive. All other films to be classified by joint discussion.

The accuracy of the two procedures may be compared in Table 3, which shows the proportion of films assigned to an incorrect definitive category. To take account of the differences in reading errors it is necessary to consider each category separately, although because of the small numbers of films involved categories B, C, and D have been taken together. On the whole, the errors associated with method P are greater than those associated with method Q.

The proportion of films selected for definitive classification by joint discussion under the two procedures is shown in Table 4. The number of films,

<table>
<thead>
<tr>
<th>Table 3</th>
<th>PERCENTAGE OF FILMS ASSIGNED TO AN INCORRECT DEFINITIVE CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>No. of Independent Readings</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
</tr>
<tr>
<td>Q</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>PERCENTAGE OF FILMS CLASSIFIED BY JOINT DISCUSSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>No. of Independent Readings</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
</tr>
<tr>
<td>Q</td>
<td>4</td>
</tr>
</tbody>
</table>
readings associated with method Q balance the slightly smaller amount of joint reading.

When the total reading time on the part of the four doctors is considered, method P demands only 55% of the effort required by method Q for the radiographs of category O, 70% for categories 1, 2, and 3, and 60% for categories A, B, and C. At a typical colliery that means that method P would involve less than 60% of the reading effort required for method Q. Bearing in mind that the majority of radiographs will be category O, the increase in accuracy associated with the latter procedure is relatively small and is outweighed by the greatly increased amount of reading involved. Method P was therefore preferred. This decision is based solely on a consideration of the accuracy of the definitive classification. The independent first-stage readings provide further information about the position of the film on the continuum of radiological abnormality. Fay and Ashford (1961) have pointed out that account must be taken of the individual reading errors if use is to be made of these readings in correlation studies, and suitable methods of doing this are being developed within the Research.

Recommendation

On the basis of the results of this trial it is recommended that the following procedure should be applied for the classification of single films obtained on the medical surveys:

(i) Every radiograph to be read independently by one or other of the doctors from each group, in such a way that each reads half of the total number of films. The doctors from the group producing the radiograph to read alternately consecutive batches of about 50 as they are taken during the survey; the doctors from the other group to read the first and second halves of the survey, respectively, the films being arranged in serial order.

(ii) If these two independent readings are consistent, the common reading to be taken as definitive. All other films to be classified by joint consultation between the four doctors reading together, their assessment being arrived at by general discussion with each of the readers in turn giving the first opinion of the category of the film.

REFERENCES

A Comparison of Some Alternative Procedures in the Classification of Chest Radiographs for Coal-workers' Pneumoconiosis

S. Rae, J. R. Ashford, D. C. Morgan, R. S. H. Pasqual and N. G. Pearson

Br J Ind Med 1963 20: 293-299
doi: 10.1136/oem.20.4.293

Updated information and services can be found at:
http://oem.bmj.com/content/20/4/293

These include:

Email alerting service
Receive free email alerts when new articles cite this article.
Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/