THE USE OF STANDARD FILMS IN THE RADIOLOGICAL DIAGNOSIS OF COALWORKERS' PNEUMOCONIOSIS

BY

C. M. FLETCHER and P. D. OLDHAM

From the M.R.C. Pneumoconiosis Research Unit, Llandough Hospital, near Cardiff

(RECEIVED FOR PUBLICATION APRIL 4, 1951)

In an earlier paper (Fletcher and Oldham, 1949), we reported the results of an experiment in which ten doctors, experienced in the radiological diagnosis of coalworkers' pneumoconiosis, were asked to classify 102 films of coal miners, by degree of abnormality, into five categories. Serious disagreement was found between the readings of different observers (inter-observer error), and between the first and second readings of the same observers (intra-observer error), and we suggested that such inaccuracies of diagnosis might be lessened by the use of standard films, exemplifying the limits of the categories, which could be used for direct side-by-side comparison with the films to be classified. We now report the results of an experiment designed to test the value of standard films for this purpose.

In analysing the results of our previous experiment, we found that although the various readers differed in the quantitative classification of simple pneumoconiosis, it could be shown that on the average their use of the four categories which we had suggested did in fact sub-divide the range of abnormality of simple pneumoconiosis, from the earliest detectable change to the most advanced abnormality, into four roughly equal divisions. Moreover, sequential radiographs of the same cases have suggested that progression of simple pneumoconiosis is much more closely related to quantitative increase of radiographic abnormality than to qualitative changes. (Davies, Fletcher, Mann, and Stewart, 1949). At the Pneumoconiosis Research Unit it was, therefore, decided to use a quantitative classification of simple pneumoconiosis and abandon the use of qualitative terms such as "reticulation", "stippling", "nodulation", and so on. The precise basis for this quantitative classification was worked out by Davies and Mann (1948), and has been more fully discussed by Fletcher, Mann, Davies, Cochrane, Gilson, and Hugh-Jones (1949). Qualitative differences between various types of films are still recognized, but the main basis for classification is a subdivision of simple pneumoconiosis of coalworkers into four categories, increasing in degree of abnormality, numbered 1 to 4, the figure 0 being reserved for films which are within normal limits. Films showing evidence of progressive massive fibrosis (P.M.F.) are separately classified in four categories A, B, C, and D. This classification has been used by the Pneumoconiosis Research Unit for three years and has also been adopted by the pneumoconiosis medical panels of the Ministry of National Insurance. A closely similar system has recently been recommended by the International Conference of Experts on Pneumoconiosis (Fletcher, 1950; I.L.O., 1950).

From the "Victor Medium" group of films used in our first experiment, standard films were selected according to the classification given during that experiment. A film on which opinion was equally divided between two categories was taken as defining the lower limit of the higher of these two categories. In this way, four films were selected exemplifying the lower limit of categories 1, 2, 3, and 4, and these were used in the routine work of radiographic assessment in the Pneumoconiosis Research Unit.

It was clearly important to assess experimentally the value of such standard films in increasing the accuracy of diagnosis. To this end a group of films was selected and four observers on the medical staff of the Pneumoconiosis Research Unit read them with and without standard films, classifying them according to a number of characteristics. This test is Part I of our experiment. It was in certain respects unsatisfactory and too complicated, so a selection was made from the same group of films, and six observers of the pneumoconiosis medical panels of the Ministry of National Insurance read this shorter series with and without standards using only the simple quantitative classification. This set of readings is Part II of the experiment.
A further short trial of the use of standard films by observers wholly inexperienced in the classification used by the Pneumoconiosis Research Unit was carried out later.

The results of the experiments were less conclusive than had been hoped, but they do suggest that standard films, when properly used, may result in increased accuracy of radiological diagnosis and may assist the inexperienced to use a classification with reasonable accuracy.

Methods: Part I

Observers.—Observers W, X, Y, and Z were all members of the medical staff of the Pneumoconiosis Research Unit (P.R.U. Observers). W, X, and Y had had considerable experience in the use of the Pneumoconiosis Research Unit’s system of classification of films. Z was well acquainted with the system but had had less experience in its use.

Selection of Films.—We decided not to use the same set of films as in our previous experiment, because these included many films of unsatisfactory technical standard. Moreover, the standard films which we were using had been extracted from the original set, which was thus incomplete. We hoped originally to use the experiment to investigate not only the diagnosis of simple pneumoconiosis but also to test the accuracy of diagnosis of abnormalities of lung markings, emphysema, and the earlier stages of progressive massive fibrosis. For these reasons, we decided to select 120 films falling into the following groups:

(1) Sixteen normal films from men who had never been exposed to a dust risk and had no respiratory symptoms; (2) 16 films of men attending an asthma clinic, who had had no coal dust exposure; (3) 16 films of coal miners who had been exposed to dust inhalation for at least 10 years but whose radiographs still appeared to be within normal limits; (4) 16 films of coal miners with category 1 simple pneumoconiosis; (5) 16 films of coal miners with category 2 simple pneumoconiosis; (6) 16 films of coal miners with category 3 simple pneumoconiosis; (7) eight films of coal miners with category 4 simple pneumoconiosis; (8) 16 films of coal miners with category 3 or 4 simple pneumoconiosis, with, in addition, early progressive massive fibrosis of category A.

The selection was made by one of us (C. M. F.) with Mr. W. G. Clarke (chief radiographer of the Pneumoconiosis Research Unit), from films already in the Unit files. Care was taken to ensure that the age distribution of the cases and the radiographic technique was similar in each group. Only those films were accepted in which the vertebral column was visible through the heart shadow but in which the intravertebral spaces could not be distinguished. Even within these limits there may be a considerable difference in penetration, but care was taken to ensure that within each group there was the same proportion of slightly under-penetrated or slightly over-penetrated films. All the films were taken with the technique described in our earlier paper for the “Victor Medium” films.

Reading of Films.—The films were arranged in an order derived from a table of random sampling numbers and each was given a serial number from 1 to 120. The set was read four times by each observer, (1) using standard films, (2) without standard films, (3) without standards, and finally (4) with standards again. This order was chosen because we intended at this stage to include inexperienced readers in the trial who, we assumed, would have to use the standards for the first reading in order to appreciate the degree of abnormality represented by each category.

During all the readings, the observers had the following written definitions of the categories before them.

Categories of Simple Pneumoconiosis

There are four categories of increasing abnormality, in addition to films that fall within normal limits (Category 0). These categories are:

Category 1 (Minimal Simple Pneumoconiosis).—In these films a small number of typical opacities can be seen, usually in the second, third, or fourth anterior rib spaces in the middle longitudinal zone of the lung, more commonly on the right than on the left. The lung markings are clearly visible.

Category 2 (Moderate Simple Pneumoconiosis).—The opacities are more numerous and are distributed in all zones with the exception of the outer third of the lung fields where they are sparse or absent. The lung markings are still visible although not so clearly as in the previous category.

Category 3 (Marked Simple Pneumoconiosis).—With the exception of an occasional large vessel in the upper and lower zones, the lung markings are now obscured by minute, small, or medium opacities, singly or in combination, profusely scattered throughout both lung fields, including the outer third of the lung.

Category 4 (Maximal Simple Pneumoconiosis).—In these films, the lung fields are, as it were, filled “to capacity” with typical opacities.

It should be noted that the distinction of the categories is made chiefly on the following grounds: the profusion of the opacities, the presence or absence of the opacities at the periphery of the lung fields, and the degree of obscuration of lung markings.

In cases in which the severity of simple pneumoconiosis varies from one area of the lung field to another, the category is determined by the most advanced disease which is present over at least half of one lung field.

During the readings with standards, five standard films were used, exemplifying the lower limits of categories 1, 2, 3, and 4 of simple pneumoconiosis and of the least definite localized shadow acceptable for a diagnosis of progressive massive fibrosis. As each film was placed on the viewing screen, the observer asked for the appropriate standard film which was then placed on an adjacent screen. Thus if he were debating in his mind whether to diagnose category 1 or 2, he
would ask for standard 2. If he considered the abnormality to be as great or greater than that in the standard film, he would diagnose category 2. The standard film category 3 would then be placed on another adjacent screen, so that he could be sure that the abnormality was not as great as the lower limit of category 3. Thus, in the reading of nearly every film, two standards were used for comparison. The second standard was unnecessary when the unknown film was close in degree of abnormality to the first.

In connexion with each film, the observer was asked to state the category of simple pneumoconiosis if this were present. If it was not, he was asked whether he thought that there was any evidence in the film that the man had worked as a miner. (The purpose of this question was to see whether there was any abnormality even earlier than category 1 simple pneumoconiosis by which men with dust exposure could be distinguished from those without exposure.) Finally, he was asked whether there was any evidence of tuberculosis or massive fibrosis, whether the lung markings were abnormal and whether there was any evidence of emphysema. (The findings from these questions will not be reported in the present paper.) The observers' readings were recorded by one of the authors or by an assistant. The films were read alternately in the orders 1–120, and 120–1. The intervals lapping between the various readings were two to six days.

After the individual readings had been carried out, observers W, X, Y, and Z met together and read the films in consultation using standard films, each film being discussed until agreement was reached. Where one observer insisted on disagreeing with the others, his opinion was disregarded, the majority opinion being accepted. Where there was an equal division of opinion, two observers differing from the other two, both opinions were recorded. This consultative reading was repeated after two days and, finally, W, X, Y, and Z met once more to reconsider those films on which their joint opinion on the two previous occasions had differed. A final agreed opinion was thus reached on every film.

Standard Films.—The standard films used were the "Victor Medium" films selected in the way that has already been described.

A preliminary analysis of the results of Part I showed that the observers had not all increased their agreement or consistency by using standard films. Nevertheless, their performance was much better than that found in the trial reported in our previous paper. It was thought that this result might be due to the fact that observers were so well acquainted with the classification and with the standard films that they had a clear picture of them in their minds and thus little further advantage was to be gained from using the actual films themselves.

We therefore decided to extend the trial to other observers without experience of the use of standards. At the same time, we decided to simplify the experiment, for observers W, X, Y, and Z had found it distracting to have to consider localized shadows, lung markings, and emphysema in addition to the category of simple pneumoconiosis, and they had also found 120 films rather more than could be read without fatigue in a single session. This extension of the experiment in a simpler form constitutes Part II.

Methods: Part II

Observers.—These were all on the medical staff of the South Wales pneumoconiosis medical panels of the Ministry of National Insurance. They had all been using the P.R.U. method of classification for some months. Since they were accustomed to reading films in pairs, it was agreed that each "observer" should consist of a pair of doctors working together. These observers are referred to as A, B, C, D, E, and F, or the P.P. observers.

Selection of Films.—On the basis of the final consultative reading of observers W, X, Y, and Z, the following groups of films were selected from the 120 films in Part 1: 16 normals (without dust exposure); eight "normal miners", who were cases that W, X, Y, and Z had agreed to have less abnormality than category 1, but which they considered showed evidence suggestive of dust exposure; 16 cases of simple pneumoconiosis category 1; 16 cases of simple pneumoconiosis category 2; 16 cases of simple pneumoconiosis category 3; eight cases of simple pneumoconiosis category 4; thus making a total of 80 films.

Reading of Films.—The films were placed in the order in which they had appeared in Part I of the experiment but they were re-numbered 1–80. Before the reading, the observers were given the same written definitions of the categories of simple pneumoconiosis that observers W, X, Y, and Z had used. Each observer then read the film in the light of these written definitions but without standard films. Two observers worked simultaneously, the first reading films 1–40 while the second read films 41–80. The first then read 41–80 while the second read 1–40. They were asked to state only the category of simple pneumoconiosis of each film, and the results were recorded by one of the authors or an assistant. An interval of half an hour followed, and the standard films were demonstrated and discussed. Finally, the reading of the films was repeated in identical order using standard films. On the few occasions when the doctors comprising each observer failed to agree, the opinion of the senior doctor was accepted.

It might be thought that the short interval between the two readings would have enabled the observers to remember their previous opinions, but apparently this did not happen. An observer sometimes commented: "I remember this film", but invariably added, "I wonder what I called it last time".

Later the readings without and with standards were repeated after an interval of four weeks in cases of C, D, and F and eight weeks later for A, B, and E. (On the second occasion, the junior doctor of "observer A" was absent on leave so the senior doctor read alone.)
Standard Films.—Since the films were read by two observers simultaneously, it was necessary to have two identical sets of standard films and it was, therefore, impossible to use the same set which had been used by observers W, X, Y, and Z, for, of these, a duplicate existed only of the standard lower limit category 2. The standards used were considered by observers W, X, Y, and Z to correspond very closely to the original set except that the new films of lower limit 1 showed a little less abnormality, and the new films of lower limit 3 had the lung markings a little less obscured. Observers A, B, C, D, E, and F all commented on the extremely small evidence of abnormality in the standard lower limit category 1 which they used, but they accepted the other films without criticism.

In summary then, the differences between Part I and Part II of the experiment were the following.

<table>
<thead>
<tr>
<th>Part I</th>
<th>Part II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observers . . .</td>
<td>W, X, Y, Z members of Pneumoconiosis Research Unit. Read as individuals. (P.R.U. observers.)</td>
</tr>
<tr>
<td>Films . . .</td>
<td>120 films including 80 used in Part II</td>
</tr>
<tr>
<td>Order of Reading . . .</td>
<td>1: standards</td>
</tr>
<tr>
<td></td>
<td>2: no standards</td>
</tr>
<tr>
<td></td>
<td>3: no standards</td>
</tr>
<tr>
<td></td>
<td>4: standards</td>
</tr>
<tr>
<td>Standard Films . .</td>
<td>Original &quot;P.R.U. set&quot;</td>
</tr>
<tr>
<td></td>
<td>Temporary set</td>
</tr>
</tbody>
</table>

We shall make the same assumption as in our previous paper, that the best estimate of the "true" category of a film may be taken to be the category nearest to the mean of the category numbers allocated to the film by all the observers. This will be called the standard category. Since it may be argued that the categories defined in words, and those defined by the standards may differ, a separate calculation of the standard categories was made from the readings with and without standard films. In fact, the standard categories with and without standard films were the same for all but four films.

We may now consider the performance of individual observers and groups of observers by relating the opinions they gave on a film to its standard category (using for the four films the standard category appropriate to the method of reading that was used). Since only five of the 3,200 opinions given in the first two parts of the experiment differed by more than one category from the standard category, we introduce only a small error by regarding all such differences as of equal weight and considering first of all simply the proportion of opinions which disagreed with the standard categories. These are given in Table 2. It is seen that the P.R.U. observers appear mostly to be assisted by standards in the normal films, whereas the pneumoconiosis panel observers are not. In the abnormal films the situation is reversed, the pneumoconiosis panel observers apparently finding the standards a help, the P.R.U. observers being for the most part unaffected. We shall discuss this finding later. However, we find that very few of the differences that appear can be assumed to indicate real change of performance, and not mere random fluctuations. Only those differences given in italics in Table 2 are statistically significant at the 5% level. That is, observer X improved significantly with standards in the "normal miner" films, observer D in category I films, and

Results

Agreement.—One of the most remarkable findings in our previous experiment was the wide range of opinions given on the same film. In Parts I and II of the present experiment there was less disagreement, even without the use of standard films. Unfortunately, the two experiments are not comparable since both the observers partaking and the films used were different, so we cannot say from these findings whether the more precise written definitions of the categories used in our present experiment were responsible for the smaller range of extreme opinion. The present results are given in Table 1. Even without standards only two films showed a range of three categories, and more than half the films had a range of one category or less.

When standard films were used, the number of films showing a wide range of classifications was further reduced, though one film remained with a range of three categories, being classified as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>Opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 opinion</td>
</tr>
<tr>
<td>2</td>
<td>8 opinions</td>
</tr>
<tr>
<td>3</td>
<td>10 opinions</td>
</tr>
<tr>
<td>4</td>
<td>1 opinion</td>
</tr>
</tbody>
</table>
### Table 1
RANGE OF OPINIONS BY ALL OBSERVERS ON 80 FILMS COMMON TO BOTH PARTS OF THE EXPERIMENT

<table>
<thead>
<tr>
<th>Range of Extreme Opinions of</th>
<th>No. of Films</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nill</td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
<td>No.</td>
</tr>
<tr>
<td>Without standards</td>
<td>5</td>
<td>6.3</td>
<td>41</td>
<td>51.3</td>
<td>32</td>
</tr>
<tr>
<td>With standards</td>
<td>5</td>
<td>6.3</td>
<td>54</td>
<td>67.5</td>
<td>20</td>
</tr>
</tbody>
</table>

### Table 2
PERCENTAGE OF OPINIONS DIFFERING FROM STANDARD CATEGORY (ANALYSIS BY OBSERVERS AND GROUPS OF FILMS WITH SAME STANDARD CATEGORY)

<table>
<thead>
<tr>
<th>Observers</th>
<th>Groups of Films by Standard Categories</th>
<th>Normal</th>
<th>Simple Pneumoconiosis</th>
<th>All Normals</th>
<th>All Abnormals</th>
<th>All Films</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Norms</td>
<td>Miners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cate-</td>
<td>Cate-</td>
<td>Cate-</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>gory 1</td>
<td>gory 2</td>
<td>gory 3</td>
<td>Cate-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>9</td>
<td>3</td>
<td>25</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>3</td>
<td>0</td>
<td>76</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td></td>
<td>19</td>
<td>6</td>
<td>83</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td>13</td>
<td>6</td>
<td>33</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>All P.R.U.</td>
<td></td>
<td>11</td>
<td>4</td>
<td>56</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>9</td>
<td>31</td>
<td>8</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>6</td>
<td>6</td>
<td>17</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>38</td>
<td>50</td>
<td>17</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>6</td>
<td>3</td>
<td>8</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>All P.P.</td>
<td></td>
<td>11</td>
<td>18</td>
<td>10</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>11</td>
<td>12</td>
<td>28</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

S = with standards. NS = without standards. Figures in italics show a significant change with use of standards.

### Table 3
OPINION OF OBSERVERS W, X, Y, AND Z ON 32 FILMS OF MEN WITHOUT DUST EXPOSURE

<table>
<thead>
<tr>
<th>Readings</th>
<th>Number of Opinions of</th>
<th>Total Number of Opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal</td>
<td>&quot;Evidence of Dust Exposure&quot;</td>
</tr>
<tr>
<td>Individual without standards</td>
<td>178</td>
<td>56</td>
</tr>
<tr>
<td>Individual with standards</td>
<td>189</td>
<td>57</td>
</tr>
<tr>
<td>First two consultative with standards</td>
<td>60</td>
<td>4</td>
</tr>
<tr>
<td>Final consultative with standards</td>
<td>32</td>
<td>0</td>
</tr>
</tbody>
</table>

* Including one opinion of Category 2
STANDARD FILMS IN RADIOLOGICAL DIAGNOSIS OF PNEUMOCONIOSIS

observer A in category 3 films, while observer C was worse with standards in the “normal miner” films.*

The results of the final consultative readings by observers W, X, Y, and Z using standard films, were interesting. The classification they reached only differed from the appropriate standard category in three films (4%). This provides supporting evidence for the view that the standard category, as we have defined it, adequately represents the classification that a film would get if errors of diagnosis were eliminated, that is, the “true” classification.

In Part I of the experiment there were 32 films of men (16 of them clinically quite normal, 16 of them asthmatic) who had never been exposed to an occupational dust-hazard, and can therefore be presumed to be free from pneumoconiosis. The accuracy of consultative reading is suggested again by the fact that none of these films was ever called category 1 pneumoconiosis in a consultative reading, though “evidence of dust exposure” was reported in four of them. Moreover, the final consultative reading even rejected these classifications and classed all the men as absolutely normal. These results and the readings given by observers W, X, Y, and Z on the films of non-miners with and without standards are given in Table 3. It will be seen that the use of a standard category I film reduced the number of times a wrong diagnosis of pneumoconiosis was made by the individual observer, but did not affect the frequency with which a mistaken diagnosis of “evidence of dust exposure” was made. It may also be noted that the “standard” category of the 16 of these absolutely normal films which were included in Part II of the experiment was also always 0, which gives another check on the validity of using this averaging method for determining the “true” category of a film.

According to Fletcher and others (1949) the diagnosis of the earliest stage of coalworkers’ pneumoconiosis depends on the detection of characteristic minute opacities. Similar minute opacities are occasionally seen in normal films especially when fine vascular markings run at right angles to the plane of the film. There is a tendency for individual observers to mistake such opacities for evidence of coalworkers’ pneumoconiosis. During the consultative readings, however, where one observer suggested such an abnormality in a normal film, one of the others would object, and in the course of critical discussion there was finally a rejection of all such spurious diagnoses. That experience is important in the use of standard films in the avoidance of these mistakes is shown by the later experiment with inexperienced readers.

Indices of Disagreement.—The effect on agreement between observers of using standard films may be illustrated graphically by using the index I_D.*

![INDICES OF DISAGREEMENT](image)

**Fig. 1.** Improvement is shown by a lower index.

In Fig. 1 the indices of disagreement for each observer for all films are plotted on vertical scales. The effect of using standard films is indicated by the lines joining the corresponding indices with and without standard films. It will be seen that seven observers (A, B, C, D, F, X, and Y) appear to be considerably improved, one (Z) not influenced and two

---

* Each percentage was transformed to a variable \( x = \sin^{-1}(\text{percentage}) \) and the differences in \( x \) were compared with the theoretical standard errors calculated from \( s_x = \sqrt{\frac{8207}{n}} \) where \( n \) is the number of films, or from an analysis of variance if this gave a significantly larger standard error.
(E and W) made worse with standards, but a test of significance shows that only the changes in A, D, and Y can be taken as definitely established.

Consistency.—The extent to which observers changed their opinion on a film between two readings is shown in Table 4. Only six observers changed their opinion on a film by more than one category without standards and only two changed their opinions by this amount with standards. The percentage of films on which the two opinions were identical was increased by the use of standards in the case of all but three observers (B, C, and F). The P.R.U. observers were more consistent than the P.P. observers with and without standards. The percentage of consistent opinions of the whole group of observers changed from 69% to 75% by the use of standards, but this apparent improvement is not technically significant. The highest level of consistency was reached in the two consultative readings, but these were closely approached by observers X and D, using standards.

The effect of standards on the index of inconsistency \( I_4 \) which was defined in our previous paper,* is shown in Fig. 2. Observers A, D, and E can be considered as significantly improved, while X and Y are possibly improved.

### Table 4

**CONSISTENCY OF CLASSIFICATION WITH AND WITHOUT STANDARD FILMS**

<table>
<thead>
<tr>
<th>Observer</th>
<th>Percentage of Films in which First and Second Opinions Differed by</th>
<th>One Category</th>
<th>Two Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NS</td>
<td>S</td>
<td>NS</td>
</tr>
<tr>
<td>A</td>
<td>56</td>
<td>81</td>
<td>41</td>
</tr>
<tr>
<td>B</td>
<td>67</td>
<td>60</td>
<td>32</td>
</tr>
<tr>
<td>C</td>
<td>81</td>
<td>73</td>
<td>18</td>
</tr>
<tr>
<td>D</td>
<td>66</td>
<td>83</td>
<td>33</td>
</tr>
<tr>
<td>E</td>
<td>55</td>
<td>74</td>
<td>45</td>
</tr>
<tr>
<td>F</td>
<td>71</td>
<td>70</td>
<td>28</td>
</tr>
<tr>
<td>W</td>
<td>71</td>
<td>76</td>
<td>29</td>
</tr>
<tr>
<td>X</td>
<td>75</td>
<td>84</td>
<td>25</td>
</tr>
<tr>
<td>Y</td>
<td>69</td>
<td>78</td>
<td>31</td>
</tr>
<tr>
<td>Z</td>
<td>65</td>
<td>68</td>
<td>31</td>
</tr>
<tr>
<td>P.P.</td>
<td>66·2</td>
<td>73·3</td>
<td>32·5</td>
</tr>
<tr>
<td>P.R.U.</td>
<td>67·0</td>
<td>76·3</td>
<td>29·1</td>
</tr>
<tr>
<td>All</td>
<td>69·3</td>
<td>74·5</td>
<td>28·7</td>
</tr>
<tr>
<td>Consultative</td>
<td>-</td>
<td>-</td>
<td>85</td>
</tr>
</tbody>
</table>

**INDICES OF CONSISTENCY**

Fig. 2.—Improvement is shown by a lower index.

Skill.—The average of the indices of disagreement and inconsistency for each observer with and without standards is shown graphically in Fig. 3. There is a tendency towards improvement with the use of standard films in the case of all observers except B and C but this improvement is only significantly established in the case of observers A, D, and Y.

Readings by Observers with no Previous Knowledge of the Classification.—In view of the rather inconclusive results which we have described, we think it appropriate to refer here to a further small extension of our experiment from which useful conclusions may be drawn.

The International Labour Office conference of experts on pneumoconiosis at Sydney adopted in February, 1950, a classification of radiographic appearances in pneumoconiosis very similar to that put forward by Fletcher and others (1949). “Simple pneumoconiosis” was there described as “pneumoconiosis with discrete opacities”, and it

* See footnote on previous page.
two other observers in South Wales, one a chest physician and the other a consultant radiologist, with experience of pneumoconiosis but with no experience in classifying films into categories, were asked to do the same. These four "inexperienced" observers (called P, Q, R, and S) read the films once according to the above definitions alone, and once with the additional help of standard films. Both readings took place on the same afternoon. The standard films were those used as a routine by the Pneumoconiosis Research Unit and now (February, 1951) also used by the pneumoconiosis medical panels of the Ministry of National Insurance.

The results of this short test showed first, that reading with the written definitions but without standards these inexperienced observers all made more mistakes than any of the other observers except B (Table 5 compare with Table 2), and this despite the fact that they had fewer opportunities for making mistakes since they were only using three categories of abnormality, whereas observers A-F and W-Z were using four categories. On the other hand, with the help of standard films they reached a level of accuracy similar to that of the more experienced observers. Secondly, standard films affected the accuracy of these observers differently in the group of normal films and the group of abnormal films. Standard films reduced their average percentage of mistakes from 52% to 23% in the abnormal films but increased their mistakes from 26% to 43% in the normal films. In this respect they behaved in a similar fashion to observers A to F but the contrast between the effect of standards in normal and abnormal films was even more striking. This trend was most marked in the films of "normal miners". Thirdly, indices of disagreement both with and without the use of standards may be calculated for these observers and for observers A, W, X, and Z, who took part in the experiment described in our previous paper, making the latter comparable by amalgamating their diagnoses of categories 3 and 4. A very rough comparison may be made with a similarly modified index of disagreement calculated from the "Victor Medium" group of films in our previous paper,* excluding for this purpose the normal non-miner films from our present experiment.

This comparison is made in Fig. 4, in which dotted lines connect the indices of those observers who partook in both experiments. There is no evidence that the more accurate verbal definitions used in the present experiment resulted in increased agreement on classification, the lower indices of observers A, E, Z, W, and X (E, F, C, A, and B respectively in the previous experiment) being possibly due to the fact that they were all to some extent conversant with the method of classification before the experiment took place. The chief benefit of standard films is clearly shown in the reduction of the gross disagreement of the inexperienced observers to levels similar to those of the more experienced.

Discussion and Conclusions

As we pointed out in the discussion in our previous paper, it is dangerous to argue too closely from a particular test on a small number of selected films.

* Category 1.—In these radiographs a small number of opacities may be seen in at least two anterior rib spaces extending over not more than half of the medial two-thirds of the lung field.

* Category 2.—In these radiographs opacities extend over more than half of the medial two-thirds of the lung field, but are sparse or absent in the lateral third.

* Category 3.—In these radiographs profuse opacities extend over the whole of both lung fields including the lateral third although they may be sparse or absent above the clavicles.

* The "Victor Medium" group of films was of similar radiographic technique and covered a similar range of abnormality.
films and observers to the general problem of accurate radiological diagnosis in pneumoconiosis. We believe nevertheless that certain conclusions may reasonably be drawn from our results and further lines of investigation suggested.

We find no evidence that increased care in the verbal definition of the radiographic appearances in the categories assists observers to agree in their classification of films. The use of standard films enabled most of the observers to improve their accuracy of diagnosis, but this increased accuracy was not uniform. The most experienced observers (W to Z) were only assisted in the classification of normal films. These observers achieved reasonable accuracy in the abnormal films without standards and gained no additional accuracy by the use of standards. With the less experienced observers and especially with the least experienced observers of all, the position was reversed. Considerable improvement in accuracy was achieved by the use of standards in the case of abnormal films, but these less experienced readers appeared to be misled by the standard film of category 1 into diagnosing abnormality where none existed.

It might be suggested that this result was found because the standard film category 1 pneumoconiosis was really within normal limits. If this were so, normal films might often show a similar appearance when compared with the standard, and would be diagnosed as having category 1 simple pneumo-

---

**Table 5**

<table>
<thead>
<tr>
<th>Observer</th>
<th>Normal</th>
<th>Simple Pneumoconiosis</th>
<th>All Normals</th>
<th>All Abnormals</th>
<th>All Films</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-miners</td>
<td>Miners</td>
<td>Category 1</td>
<td>Category 2</td>
<td>Categories 3 and 4</td>
</tr>
<tr>
<td>P</td>
<td>NS 38 38</td>
<td>NS 33 67</td>
<td>NS 94 29</td>
<td>NS 81 12</td>
<td>NS 4 8</td>
</tr>
<tr>
<td>Q</td>
<td>NS 0 6</td>
<td>NS 0 17</td>
<td>NS 61 59</td>
<td>NS 56 12</td>
<td>NS 42 21</td>
</tr>
<tr>
<td>R</td>
<td>NS 19 38</td>
<td>NS 17 83</td>
<td>NS 67 41</td>
<td>NS 94 0</td>
<td>NS 67 33</td>
</tr>
<tr>
<td>S</td>
<td>NS 56 56</td>
<td>NS 33 100</td>
<td>NS 61 47</td>
<td>NS 31 12</td>
<td>NS 17 8</td>
</tr>
<tr>
<td>All</td>
<td>NS 28 34</td>
<td>NS 21 67</td>
<td>NS 71 44</td>
<td>NS 66 9</td>
<td>NS 32 18</td>
</tr>
</tbody>
</table>

NS = without standards  S = with standards

---

FIG. 4.—Comparison of previous experiment and present results. Observers common to both experiments joined by dotted lines.
normal limits. Our reason for rejecting this explanation is chiefly based on the performance of observers W to Z, who, both in their individual reading and especially in consultation, were apparently able to see some particular appearance in the standard film which was not present in the films of non-miners. Very careful scrutiny was necessary to make this distinction which was only made with complete accuracy when they discussed the films in consultation.

It appears that considerable experience is necessary for the identification of the earliest characteristic radiological abnormalities of pneumoconiosis and that only observers W, X (observers A and B of our first paper), A, D, and F showed themselves capable of detecting this appearance with reasonable accuracy. The other observers, when using the standard appear to have been misled by some incidental appearance not characteristic of pneumoconiosis. We are engaged upon a further experiment to test this point, using a larger number of normal films. Preliminary results suggest that experienced readers can consistently distinguish films of miners with no more abnormality than the lower limit of our category I from films of men who have never been exposed to dust. Methods of assisting the inexperienced to use the standard film of category I effectively are being explored.

There is probably considerable individual variation in ability to use standard films. Certain readers may never use them easily. For instance, observer W, whose accuracy was the best of the whole group of observers when working without standards, was a quick competent reader and appeared unwilling to change his immediate impression of a film when the standard was presented to him. He appeared to prefer his own mental standard to the objective standard films.

We have observed the high consistency of a consultative reading in which four observers read the films using the standards and discussed each film until they reached agreement. When this consultative reading was repeated on two occasions, and the small number of films on which the two separate readings were different were further discussed until agreement was reached, the final result agreed with the average opinion of all the observers when they were using standards in all but three of the 80 films. Moreover, both the consultative opinion and the standard category gave the correct diagnosis when films which were known to be normal were in question. If it is accepted that the average opinion of a set of observers gives the true category of a film, consultative readings will also tend to give this true category.

We have at present, therefore, adopted the consultative technique in the routine reading of films taken on radiological surveys by the Pneumoconiosis Research Unit. The films are read on two occasions by two or three observers using standard films, and those films on which the two readings are different are subjected to a third scrutiny to reach a final opinion. This technique is similar to that suggested by Yerushalmy, Harkness, Cope, and Kennedy (1950) for the most accurate radiographic diagnosis of tuberculosis.

Certain drawbacks and difficulties in the use of standards became apparent in the course of our experiment. First, there is considerable variety in the type of appearance in radiographs of coal-workers' pneumoconiosis (Fletcher and others, 1949). The standards used in Parts I and II of our experiment were all of the "mixed" type of pneumoconiosis. A number of films in the series were of a much finer type (pinhead) than the standards and, in such cases, the observers complained of inability to use the standard films for comparison. They were being asked, they said, to compare dissimilar appearances. Some confirmation of this dissatisfaction was found by considering those films which were predominantly of the pinhead type with those films which were predominantly of the mixed or nodular type. One of us (C. M. F.) and observer Y sorted all the films in category 2, 3, and 4 into those predominantly of the pinhead type and those predominantly mixed or nodular. It was found that in Parts I and II of the experiment, the proportion of correct opinions in the pinhead group was 65% working without standards, and 66% working with standards, whereas in the nodular and mixed group, 72% of the opinions were correct without standards and 79% correct with standards. The lack of improvement with standards in the pinhead group confirms the observers' expressed difficulty in assessing the pinhead films with standards. For this reason we have decided that, at least in the case of category 3, it is necessary to have two standard films, one of the mixed nodular type and the other of the pinhead type, so that like can be matched with like. (The set of standards used by the "inexperienced" observers P, Q, R, and S included a "mixed" and a "pinhead" category 3 film.)

Secondly, there is the labour involved in reading with standards. More than one viewing screen is required and for each film read at least one standard has to be selected and placed on the screen for comparison. This means that reading with standards slows the speed at which films can be scrutinized. It is, however, possible that by a specially designed viewing screen, the use of standards could be made less laborious.
Even when standard films are used, our results suggest that the level of accuracy achieved by the observers who took part in our experiments is still far from satisfactory.

Radiographs are generally used as the chief means of diagnosis of pneumoconiosis in life for purposes of diagnosis and compensation. It has been suggested that men with less than category 2 simple pneumoconiosis are unlikely to be disabled by this disease and may be rejected on scrutiny on their films if they apply for disablement benefit on account of pneumoconiosis. The cases in our experiment with films of agreed category 2 might then be considered for benefit. Of this group of films, 20% of the opinions were of category 1 when read with standards, so that if observers A to F and W to Z were responsible for scrutiny of those cases, we could expect 20% of them to be incorrectly deprived of assessment for disablement benefit, even if standards were used.

Again, it has been suggested that progressive massive fibrosis, which is responsible for the most serious disablement in coal miners' pneumoconiosis, seldom arises in cases with less than Category 3 simple pneumoconiosis. If this were confirmed, and men were removed from dust exposure before they reached this level, they would then be protected from developing the serious disability associated with massive fibrosis (Fletcher, 1950). To do this it would be necessary to diagnose Category 2 with accuracy. We find that, working with standards, 10% of the opinions given on films of Category 2 simple pneumoconiosis were of Category 3 and 19% of the Category 3 films were called Category 2. Thus our observers would make some 15% mistakes in diagnosing cases at this "critical stage". On the other hand, observers W, X, Y, and Z working together in consultation only mis-diagnosed one of the 17 cases of Category 2 and none of 16 cases of Category 3. We suggest that for important decisions based on radiographical diagnosis of pneumoconiosis the technique of consultative readings on at least two separate occasions should be used.

Lastly, we must consider briefly the influence of radiographic technique on diagnosis which, in our previous paper, we showed was considerable. We did not explore this question in our present experiment but we have considerable further experience in our routine work of the serious influence of changes in radiographic technique. Two films of the same man on the same day, taken with different techniques, can lead to opinions differing by one category or occasionally by two categories even when consultative reading with standards is used. We can only emphasize that for accurate radiographic diagnosis it is essential that films should only be used which are of first-class standardized technique, such as that recently recommended by the Medical Research Council (1950).

Summary

An experiment was designed to test the value of "standard films" in the qualitative classification of radiographs of simple pneumoconiosis of coal-workers, the standard films being examples of the limits of the categories and used in side-by-side comparison with the films to be classified.

The experiment fell into two main parts. In the first, experienced readers from the Pneumoconiosis Research Unit took part; in the second, members of the pneumoconiosis medical panels, and in a subsequent extension, observers who had no previous experience of the method of classification.

We have no evidence that increased precision in the verbal descriptions of the categories enabled the observers to classify the films with any greater agreement than had been found in an earlier experiment with less precise definitions.

The use of standards enabled some, but not all of the observers to achieve a greater accuracy and consistency in classification. The least experienced observers benefited most, but they tended to make false diagnoses of abnormality in normal films when working with standards. On the other hand, the more experienced readers received greater help in the discrimination of normal films from those with slight abnormality, than in the classification of the more abnormal films. This finding is discussed.

The greatest consistency was achieved by four experienced observers reading in consultation with standard films and their classification agreed very closely with the standard categories derived from averaging the opinions of all the observers.

The chief practical use of standard films appears to be to enable wholly inexperienced observers to use a radiographic classification with an accuracy approaching that of the more experienced observers.

It is suggested that in routine work films should be read by at least two observers on two independent occasions, the films on which the two opinions differ being given a third and deciding reading. Only films of first-class radiographic technique can be used for accurate diagnosis.

We should like to thank all the observers who helped us in this experiment and also Dr. C. L. Sutherland, Chief Medical Officer of the Pneumoconiosis Medical Board of the Ministry of National Insurance, for agreeing to the collaboration of observers from the pneumoconiosis medical panels.
REFERENCES


——, and Mann, K. J. (1948). Ibid., p. 769.


——, and Oldham, P. D. (1949). British Journal of Industrial Medicine, 6, 168.

