THE PROBLEM OF CONSISTENT RADIOLOGICAL DIAGNOSIS IN COALMINERS' PNEUMOCONIOSIS

AN EXPERIMENTAL STUDY

BY

C. M. FLETCHER and P. D. OLDHAM

From the Pneumoconiosis Research Unit of the Medical Research Council

(RECEIVED FOR PUBLICATION, DECEMBER 31, 1948)

INDEX	Page
INTRODUCTION	168
DESCRIPTION OF EXPERIMENT	169
RESULTS	172
I. Lack of Agreement	172
Discrepancies of opinion	172
Inconsistency of observers	173
Average opinions	173
Effect of radiographic technique on categorization	173
Certification	174
II. Further analysis	176
Arithmetical methods	176
Bias of observers	177
Changes in mental standard	178
Skill and experience	178
Relative difficulty of films	180
Effect of practice and fatigue	181
DISCUSSION	181
SUMMARY	182
ILLUSTRATIONS	170

Introduction

The classification of cases of pneumoconiosis according to their severity is often necessary. In the practical field it is essential in the assessment of cases applying for compensation for industrial lung diseases, and in deciding whether to recommend suspension from employment in dusty environments. In research it is necessary for such purposes as the study of the progression of the disease and of its varying incidence in different environmental conditions. Since a chest radiograph provides the only sure method of diagnosing pneumoconiosis in life, and since no simple method for the objective assessment of respiratory disability is available, most schemes of classification have been based primarily upon the radiological picture. (For an account of the literature see Hart and Aslett, 1942.)

In view of the great practical importance of the problem, it is surprising that little attention has been given to the difficulties inherent in the accurate classification of chest radiographs in pneumoconiosis. It has been shown that in the radiological diagnosis of the presence of pulmonary fibrosis (Herbert, 1939) and of pulmonary tuberculosis (Birkelo and others, 1947) great disparities exist between the opinions of different trained observers and that their opinions are not always self-consistent, and it has also been shown that in the clinical assessment of malnutrition (Derryberry, 1938 ; Huws Jones, 1938) and of cyanosis (Comroe and Botelho, 1947), and even in the instrumental assessment of hæmoglobin concentrations (Macfarlane, 1945), similar disparities of opinion arise. It seemed desirable to investigate the reliability and precision which it is possible for experienced observers to achieve in the classification of chest radiographs in pneumoconiosis.

This information was particularly needed for an investigation undertaken by the Pneumoconiosis Research Unit of the Medical Research Council into the progression of coalminers who had been radiographed by Hart and Aslett (1942) between 1938 and 1940 and who had continued to work underground for some years after. Some of these men were radiographed again in 1946 and the radiological progression of their pneumoconiosis was assessed in relation to the dust exposure that they had undergone in the interval. This radiological assessment demanded the classification both of the original Hart and Aslett films and of the more recent films taken by the Pneumoconiosis Research Unit. It was important to find out how accurately this was being done and whether the classification employed corresponded to that employed by the Silicosis Medical Boards.

Description of Experiment

An experiment in x-ray classification was designed with the intention of answering these questions. A set of films was chosen from those used by Hart and Aslett, the Pneumoconiosis Research Unit, and the Silicosis Medical Boards, representing the early stages of pneumoconiosis (in which the greatest difficulties of accurate diagnosis arise). These films were classified into five categories of increasing abnormality by ten observers of varying experience of pneumoconiosis, so that the effect of experience on what should be a purely objective diagnosis would be apparent. To assess the consistency of the classification given to the films, each observer read them a second time in a different order without being told what his previous classifications were.

The films were chosen so that certain other questions could be answered by the experiment. The level of the radiographic abnormality necessary before a Silicosis Board can certify a case is not defined in the Coalminer's Pneumoconiosis Scheme of 1943, so the Silicosis Medical Boards have had to determine for themselves the level of abnormality which should be considered "certifiable." This level of abnormality was within the range covered by the chosen films, and thus evidence of the way in which the definition of "certifiability" was interpreted in practice, and the consistency of the interpretation, could be obtained from the results. A group of films was also included which would show the effect of different radiographic techniques on classification. The experiment was also intended to suggest the answers to certain other questions.

The Observers.—Throughout this paper the ten observers will be referred to by the capital letters A to K (omitting I). Their differing experience of films of pneumoconiosis is indicated by the following brief notes.

A and B.—Members of the medical staff of the Pneumoconiosis Research Unit who were engaged on the classification of the films of the Hart and Aslett Survey and its follow up.

C and D.—Members of the medical staff of the Pneumoconiosis Research Unit with scientific training but with less experience in the interpretation of radiographs of early pneumoconiosis.

E and F.—Senior members of the Silicosis Medical Boards. They had long experience in the classification of films as "certifiable" and "not certifiable," but less experience in the finer classification of films that were "not certifiable."

G.—A junior member of the Silicosis Medical Boards who had spent several months working with the Pneumoconiosis Research Unit in the classification of the Hart and Aslett films referred to above. He combined, for

shorter periods, the experience of observers A and B, and E and F.

H.—A tuberculosis physician who had taken part in the classification of the films of the Hart and Aslett Survey in 1940. He had continued to work in a South Wales mining area and so had continued to see many cases of coalminers' pneumoconiosis.

J.—A tuberculosis physician who had taken part in the original classifications of Hart and Aslett's films in 1940 but who had, for the past five years, seen little of coalminers' pneumoconiosis.

K.—A tuberculosis physician who had worked for several years as a Mines Medical Officer and was engaged upon a study of early pneumoconiosis in working miners.

It was suggested that the lack of agreement and consistency that were discovered might be attributable to the fact that none of the observers was a trained radiologist. The two consultant radiologists on the staff of London teaching hospitals, who had special knowledge of chest diseases in general were, therefore, asked to read the films, and they agreed to do this. They are referred to respectively as observers L and M. Since the main investigation was intended to explore opinions of doctors concerned primarily with coalminers' pneumoconiosis, and since the opinions of observers L and M were not included when the standard categories referred to below were determined, their opinions are considered separately, and are included in the tables in this paper in brackets and in the figures in dotted lines. Observer M would only use categories 1, 2, 4, and 5, so that it is not possible to calculate some of the indices from his readings.

Reading of Films.—The films were arranged at random (by drawing their reference numbers out of a hat) and each was given a serial number. Each observer read them in order 1 to 102 on the first occasion and in the reverse order (102 to 1) on the second. He was given his own choice of viewing screen.* He was asked to classify each film into one of the following categories: (1) within normal limits; (2) very early pneumoconiosis; (3) definite pneumoconiosis, but not certifiable for purposes of compensation; (4) definite pneumoconiosis of certifiable degree but showing only "reticulation"; (5) more than "reticulation," that is, "nodulation" or "coalescent shadows."

These categories were later used as a basis for the formal classification of simple pneumoconiosis published by Davies and Mann (1948). The numbers used by Davies and Mann, however, are different. They use the figure 0 to indicate normal films, so that their category 1 is roughly equivalent to our category 2, their category 2 to our category 3, and so on.

Before the films were read, the categories were discussed and, if necessary, elaborated or explained

* All observers except E and J used a General Electric X-ray Corporation Fluoroline illuminator. E and J used a Kodak illuminator.

FIG. 1.—Stage 2.
Average opinion 1·9. Standard opinion 2.

Category	..	..	1	2	3
No. of opinions	..	..	7	8	5

Davies and Mann (1948) category 1.

FIG. 2.—Stage 3.
Average opinion 3·0. Standard opinion 3.

Category	..	..	2	3	4
No. of opinions	..	..	4	12	4

Davies and Mann (1948) category 2.

FIG. 3.—Stage 4.
Average opinion 4·0. Standard opinion 4.

Category	..	..	3	4	5
No. of opinions	..	..	2	16	2

Davies and Mann (1948) category 3.

FIG. 4.—Stage 5 (nodulation)
Average opinion 4·25. Standard opinion 4.

Category	..	..	4	5
No. of opinions	..	..	15	5

Davies and Mann (1948) category 4.

FIGS. 1 to 4.—Photographs of the right mid-zone of typical examples of stages 2, 3, 4, and 5. It should be noted that these are respectively examples of categories 1, 2, 3, and 4 according to the classification of simple pneumoconiosis published by Davies and Mann (1948). Under each photograph is given the average opinion of all observers on the film, the standard opinion, and the number of opinions of each category given to the film in the course of the experiment. These are all Victor Medium Films.

by the authors. After the reading of the films had started there was no further discussion or comment on any of the diagnoses, except that the observer was invited to comment on the radiographic technique of any film if it disturbed his judgment. He was also told the age of each case whose film was under consideration if he asked for it. Names and identification numbers on the films were obscured by a mask.

Observers A, B, C, D, G, H, J, K, and M read the films alone. Observers E, F, and L read them with colleagues who occasionally disagreed with their opinion, but the opinion of the observer under test was taken in every such case.

The interval between the two readings of the films by each observer was as follows: observers A, B, C, F, G, 1 week; observers D, E, H, K, 2 weeks; observer J, 5 weeks.

Selection of Films.—Films chosen from three sources were given a preliminary classification by the authors in order to ensure fair representation of the particular degrees of abnormality whose diagnosis we wished to investigate. This preliminary classification was used for reference purposes only; later it will be shown how a "standard category" was determined for each film from the experimental results. To make the distinction clearer the word "stage" will be used whenever the preliminary classification is referred to, the word "category" when a classification made during the experiment is described. For example, three stage 2 films might subsequently be given standard categories of 2, 2, and 3 when the experimental results were examined.

Photographs of the right mid-zone of typical examples of stages 2, 3, 4, and 5 are reproduced (figs. 1 to 4).

SOURCE (A) VICTOR FILMS

Set used: Victor 4-valve Photo-Röntgen Unit with Coolidge Rotating Anode Tube.

Anode-Film distance: 6 feet.

Focal Spot: 2 mm.

Anode Current: 400 mA.

Peak Voltage: 55–58 kV (according to chest thickness).

Film: Kodak code 6, with High Definition Intensifying Screens.

Exposure Time: (average) 0.1 sec.

Processing: Kodak Blue Label Developer, for five minutes at 65° F. Intermediate rinsing and fixing for at least twenty minutes, followed by prolonged washing in running water.

Comment: Films were taken in 1946. They were a little more penetrated than is usual for chest work, the spine but not the intervertebral discs being just visible through the cardiac shadow.

Films were chosen from cases that had been examined by Hart and Aslett in 1938 and were considered to have shown no radiological progression between then and their recent examination in 1946, when the two films from each case were compared side by side, and whose

recent (Victor) films appeared to fall into the stages 1 to 4 defined above. Ten examples were taken in each stage. From these forty "Victor" films, examples of each stage were chosen which appeared on juxtaposition to form a series of equal graduations from 1 to 4. These were taken as reference films. In each stage those films which most closely resembled the reference films were selected in order to form groups of films within the four stages. The other films were rejected. In this way, four stage 1, six stage 2, six stage 3, and six stage 4 films were selected. Four cases of nodulation and four with early coalescent or early "ambiguous" (Davies and Mann, 1948) shadows on a background of reticulation were also selected. This set of films will be referred to as "Victor Medium" films.

There existed for nine of the cases included among the Victor Medium films (three each of stages 2, 3, and 4) additional films of greater and less penetration, taken on the same day. The over-penetrated films will be called "Victor Hard" films and the under-penetrated films "Victor Soft" films. These nine triple-film cases provided an opportunity for estimating the effect of technique on the categories of the films, as will be seen later.

SOURCE (B) "H AND A" FILMS

Set used: Coolidge 10 kW.

Anode-film distance: 4 feet 6 inches.

Focal spot: 3.8 mm.

Anode current: 100 mA.

Peak voltage: 60–70 kV (according to chest thickness).

Film: Ilford, with Fluorazene screens.

Exposure time: 0.1 sec.

Comment: The films were taken between 1938 and 1940 by Messrs. Portable X-Rays.

The films were selected from the same cases as the thirty Victor Medium films. Since they were all cases in which no progression was apparent in the interval between the films, they were assumed to be in the same disease-stages, although it was found later that this assumption was not confirmed by the observer's opinions in the course of the experiment (see Table 11).

SOURCE (C) "SI.B." FILMS

Dr. N. Keating of the Cardiff Silicosis Medical Board allowed us to select twenty-four films from the patients examined by his Panel during 1946, half of whom had been rejected and half of whom had been certified. From these, twelve certified and twelve rejected cases were chosen which were considered after comparison with the standard reference films to fall into stages 3 and 4. Unfortunately, it was later found that by a clerical error thirteen rejected cases and eleven certified cases had been chosen. These films are referred to respectively as "Si.B.R." and "Si.B.C." films.

Set used: Watson 4-valve.

Anode-film distance: 5 feet.

Focal spot: 1 mm.

Anode Current: 300 mA.

Peak Voltage : 45-70 kV (according to chest thickness).

Film : Kodak Blue Brand with Kodak fast screens.

Exposure time : 0.08 sec.

Comment : The films were taken in 1946 by a private radiologist.

Thus the final set of 102 films selected for the experiment was made up as follows :

Victor Medium : 4 stage 1, 6 stage 2, 6 stage 3, 6 stage 4, 8 stage 5, Total : 30.

Victor Soft : 3 stage 2, 3 stage 3, 3 stage 4, Total : 9.

Victor Hard : 3 stage 2, 3 stage 3, 3 stage 4, Total : 9.

H and A : 4 stage 1, 6 stage 2, 6 stage 3, 6 stage 4, 8 stage 5, Total : 30.

Si.B.R. : 9 stage 3, 4 stage 4, Total : 13.

Si.B.C. : 3 stage 3, 8 stage 4, Total : 11.

All Groups : 8 stage 1, 18 stage 2, 30 stage 3, 30 stage 4, 16 stage 5, Total : 102.

The distribution of standard categories (based on the results of our experiment) within the various groups of films will be found in Table 11.

During the course of the trial, some observers objected to certain films on technical grounds, considering them to be unreadable. In certain of the calculations these films were excluded, but in no case did such exclusion have any significant effect on the results.

Results

PART I. LACK OF AGREEMENT

Discrepancies of Opinion.—With the exception of C and D the ten observers were all experienced in the radiological diagnosis of early coalminers' pneumoconiosis, and with the exception of J they were engaged in the scrutiny of such films every day in South Wales. Their opinions might thus be expected to be fairly uniform. In fact there was a remarkable disparity between the twenty opinions given on each film. This disparity may be expressed as in Table 1 by stating the range of extreme opinions expressed. There were three films on which the opinions ex-

pressed covered the whole range of four categories, being considered normal by one observer, while another considered them as examples of "more than reticulation." (In each of these films the diagnosis of category 5 was based on the supposed presence of an area of coalescent shadowing in some part of the film; they were not diagnosed as examples of "nodulation." In each of them, however, there were one or more diagnoses of category 4.) There were in addition thirty films, just under one third of the total, in which there was a range of three categories in extreme opinion—from "normal" to "certifiable reticulation" or from "very early pneumoconiosis" to "more than reticulation."

When the films which any observer had considered unreadable were removed from the group, the results were not much altered. Only one of the films with a range of opinion of 4 categories and one with a range of 3 categories fell in the "unreadable" class.

When the opinions on all the films of observers L and M are considered, we find there were fourteen films diagnosed by L as category 4, but diagnosed by M as category 1; ten films which L diagnosed as category 5 but which M diagnosed as category 2, giving twenty-four films in which the opinions of these two observers differed by three categories. There were in addition, four films which observer L diagnosed as category 5 and M diagnosed as category 1. It is thus apparent that the disparity of opinion between these two consultant radiologists is at least as great as that between the ten other observers.

Perhaps the most remarkable feature of the disparity of opinion shown by this trial is the disagreement revealed as to the limits of normality. Including the opinions of L and M together with the other ten observers, we find that of the 56 films considered "normal" on at least one occasion :

5 were considered on at least one other occasion to be in category 2

17 were considered on at least one other occasion to be in category 3

21 were considered on at least one other occasion to be in category 4

11 were considered on at least one other occasion to be in category 5

The question of determining the limits of normality is of great importance in investigations of the incidence of pneumoconiosis in various industries or in different factories or mines within a given industry. In Table 2 are given the numbers of films placed by each observer in category 1 (normal) on either test, the first test, the second test, and on both tests. It will be seen that if these observers had undertaken independent surveys to determine the

TABLE 1
RANGE OF EXTREME OPINIONS OF OBSERVERS A-K

Range of one category		Range of two categories		Range of three categories		Range of four categories	
Ex- treme opin- ions	No. of films	Ex- treme opin- ions	No. of films	Ex- treme opin- ions	No. of films	Ex- treme opin- ions	No. of films
1-2	5	1-3	18	1-4	16	1-5	3
2-3	1	2-4	21	2-5	14	—	—
3-4	4	3-5	15	—	—	—	—
4-5	5	—	—	—	—	—	—
Total	15	Total	54	Total	30	Total	3

incidence of pneumoconiosis among some group of workmen, their conclusions on the incidence of disease would be very different.

Inconsistency of Observers.—It will have already been noted in Table 2 that each observer did not give the same opinion on the films on the two occasions on which he saw them. One would not expect to find that the inconsistencies between the two readings of the film given by each observer would be as great as the disparities that were found in the twenty readings of the films given by all ten observers. Table 3 shows for each observer the number of films which he placed in the same category on the two readings, and the number of films in which his opinion differed by 1, 2, or 3 categories on the two occasions. It will be seen that all the observers placed at least half the films in the same category on two occasions, and a change of

TABLE 2

THE NUMBER OF NORMAL FILMS ACCORDING TO EACH OBSERVER

Observer	Number of films called normal on :			
	First test	Second test	Either test	Both tests
A	5	7	7	5
B	5	8	9	4
C	5	6	8	5
D	3	4	4	2
E	14	13	18	9
F	11	8	12	7
G	10	5	10	5
H	17	11	18	9
J	21	20	25	14
K	29	21	30	20
(L)	(10)	(10)	(13)	(7)
(M)	(44)	(42)	(52)	(34)

three categories (categories 2 to 5) occurred only once. In this table the changes of opinion in either direction are considered together. The question of general trends in the differences of opinion between the observers and in the two opinions of observers on two occasions will be considered later.

Average Opinion.—In the course of analysing our results, we shall have occasion to consider groups of film-classifications, the characteristics of which can most easily be described by the use of averages. The different averages we shall use are defined as follows :

(a) *The average opinion of any observer on a particular group of films* is the arithmetic mean of the category-numbers which he allocated to each film in the group.

(b) *The average opinion of a group of observers on any given film* is the arithmetic mean of all the category-numbers allocated by the observers to this film.

(c) *The average opinion of an observer on any one film* is the arithmetic mean of the two category-numbers allocated by him to the film on the two occasions on which he saw it.

Effect of Radiographic Technique on Categorization.—To provide evidence on the effect of radiographic technique on the classification of films we have available the nine cases of each of which three films were taken on the same day with medium penetration, over-penetration and under-penetra-

FIG. 5.—Diagram showing effect of radiographic technique on the average opinion of all observers on the twenty-seven films from the nine cases, each of whom had three films taken on the same day with low penetration ("soft" films), medium penetration, and high penetration ("hard" films). The mean opinion of all observers on the three films of each case are plotted on the same vertical line. The vertical axis is marked in category numbers.

- × Mean opinion on "soft" films (low penetration).
- Mean opinion on "medium" films.
- Mean opinion on "hard" films (high penetration).

tion. The effect of technique on opinion may be analysed in two ways.

First we can consider the effect of the "average opinion of all observers" on each of these twenty-seven films. In fig. 5 the average opinion of all the observers on each of the three films of each case is plotted against an axis marked in category-numbers. The average opinions on the three films taken of each case are plotted on the same vertical line. It will be seen that, except for Case 9, the soft (under-exposed) film is in each case rated higher than the medium and hard (over-exposed) films, and

except for Cases 6, 8, and 9 the medium film is rated higher than the hard film.

It will also be noted that the effect of technique on the average opinion is not the same in each of the cases. This may be because unfortunately the degree of over- and under-penetration was not the same in the corresponding films in this group, so that the differences between hard, medium, and soft were not always of the same order. To provide some absolute measure of differences in technique within each set of three films, densitometer readings were taken in the third right anterior rib space of each film, and the average change of opinion with change of technique was plotted against the corresponding differences of optical density, but only a very rough linear relation between the two was apparent. In particular there were six points which diverged widely from the others. Those derived from Case 9 may be explained by the fact that the films of this case were of a characteristic generalized pin-head type (Davies and Mann, 1948) in which the diagnosis of certifiable reticulation was reached easily by nearly all the observers and the differences in technique were insufficient to obscure the pattern. No explanation could be suggested for the other divergent points. Thus, while optical density is one factor which results in changes of opinion with changes in radiographic technique, other factors are probably involved.

The effect of technique on the opinions of the individual observers can also be examined. In fig. 6, the "average opinion of each observer" on all the nine hard, nine medium, and nine soft films are plotted on the same vertical lines. It will be seen that some observers (for example C, E, F, and J) are seriously affected by technical differences in the expected direction,

but others (for example A, B, and G) are only slightly affected. Observers B, D, and G either compensate completely for over-penetration or even over-compensate so that a lower mean opinion is given by them for the medium than

FIG. 6.—Diagram showing the effect of radiographic technique on the average opinion of each individual observer on the twenty-seven films from the nine cases each of whom had three films taken on the same day with low penetration ("soft" films), medium penetration, and high penetration ("hard" films). The average opinion of each observer on all nine "hard" films (over-penetrated) of all nine medium films and all nine "soft" films (under-penetrated) is given on the same vertical line. The vertical axis is marked in category numbers.

- × Mean opinion of each observer on "soft" film (low penetration).
- Mean opinion of each observer on "medium" film.
- Mean opinion of each observer on "hard" film (high penetration).

for the hard film. Observers H and M over-compensated for under-penetration.

Certification.—For the purposes of reaching decisions concerning compensation for pneumoconiosis, the Silicosis Medical Boards are not required at present (early 1948) to classify films into several stages. They are only required to place the films above or below a certain standard of abnormality. This standard has never been defined but it is based on the description of reticulation given by Hart and Aslett (1942, p. 48). "The lung fields show a fine network, sometimes sharp and lace-like in pattern, but much more often blurred in appearance. The network occupies from half of one lung field to (more frequently) the whole of both fields, though the apices are relatively free. It is as marked at the periphery of the fields as it is nearer the hila." This definition is only partly quantitative and it would not be surprising, in view of the disparities and inconsistencies of opinion which we have already described, to find different observers interpreting it in different ways. Since categories 4 and 5 in our experiment were defined respectively as "certifiable reticulation" and "more than certifiable reticula-

TABLE 3
CHANGES OF OPINION BETWEEN FIRST AND SECOND READINGS

Observer	Number of films in which difference between first and second opinions was :			
	Nil	One category	Two categories	Three categories
A	75	27	—	—
B	84	18	—	—
C	55	46	1	—
D	66	32	4	—
E	64	37	1	—
F	70	30	2	—
G	62	33	7	—
H	59	40	3	—
J	54	37	10	1
K	58	40	4	—
(L)	(59)	(35)	(8)	(—)

TABLE 4
NUMBER OF FILMS CONSIDERED "CERTIFIABLE" BY
EACH OBSERVER

Observer	Number of films "certified" on :				Consistency of certification (%)
	First test	Second test	Either test	Both tests	
A	41	32	44	29	66
B	29	27	30	26	87
C	45	40	55	30	55
D	52	53	63	42	67
E	34	34	36	31	86
F	50	53	58	45	78
G	41	51	55	37	67
H	38	47	49	36	73
J	32	35	41	26	63
K	12	9	16	5	31
(L)	(54)	(58)	(65)	(47)	(72)
(M)	(36)	(34)	(40)	(30)	(75)

tion," it is possible to use our results to examine the question of errors in certification; for all films placed in categories 4 and 5 by any observer must have been regarded by him as showing sufficient abnormality for certification. Table 4 tabulates the results from this point of view. It shows the number of films "certified" by each observer on the first, second, or both tests. It will be seen that there are great differences between the observers. Observer D, for instance, "certified" 63 films on at least one occasion whilst observer K "certified" only 16, and observer F, of the Silicosis Medical Board, "certified" 50 per cent. more cases than his colleague E. There are also considerable inconsistencies in this matter of certification between the two opinions given by each observer, although here too, the inconsistencies between the observer's two opinions are less marked than the disparities between the various observers.

There is another factor that will greatly affect an applicant's chance of certification. This is the radiological technique of the film with which he applies. The effect of technique on certification is

shown in Table 5, which gives the number of certifying opinions (out of twenty given) on each of the twenty-seven films of the nine cases who had a hard, medium, and soft film taken on the same day. The effect of technique is striking, especially in the cases in standard category 3 near the border-line of certification. Had Case IV, for instance, applied to a panel made up of the observers in the experiment, he would have had a fourfold greater chance of certification if he had applied with his soft film than if he had applied with his hard film.

These results show how greatly the chance of certification of any applicant to the Silicosis Medical Board may depend upon the particular doctor who assesses his film and even on changes in this particular doctor's opinion from day to day, and they reveal the need for some degree of assessment which will lead to more consistent results.

The films in this trial, however, were deliberately chosen to include a large number in the earlier stages of the disease. There would clearly be very much less or even no inconsistency in the certification of cases with completely normal films or with films showing massive shadows. It is, therefore, of importance to examine the relevance of our results to the present practice (1948) of the Silicosis Medical Board. A review was made of four hundred films of men applying to the Silicosis Medical Boards at Cardiff and Swansea during the first six months of 1948, for the purpose of estimating in what proportion of them conflict of opinion concerning certification might arise.

Two films were selected from the experimental group of 102, which (from the average opinion of the ten observers), appeared to mark the boundaries of the range of abnormality in which "certifiability" could be disputed. (One film had the lowest average opinion of those on which there was only one *certifying* opinion given by observers E and F, members of the Silicosis Boards; the other had the highest average opinion of those on which only one *non-certifying* opinion was given by E or F). The four hundred films were then

TABLE 5
NUMBERS OF "CERTIFYING" OPINIONS (CATEGORIES 4 OR 5) ON NINE CASES WITH HARD, MEDIUM, AND SOFT FILMS TAKEN
ON THE SAME DAY

	Standard category	2			3				4		Whole group
	Case No.	I	II	III	IV	V	VI	VII	VIII	IX	
No. of "certifying" opinions when tech- nique was :	Hard	—	—	1	4	4	8	5	17	15	54
	Medium	—	—	1	9	9	6	7	18	16	66
	Soft	—	5	1	16	11	14	11	19	14	91

divided into those that lay below, between, or above these two standards in degree of abnormality, assessed by side-by-side comparison. It will be suggested later that this comparative technique is inherently more accurate than the assessment *in vacuo* used during the trial and in general radiological practice.

The results of this review are shown in Table 6. During this period, 20 per cent. of cases applying to the Cardiff Panel and 24 per cent. applying to the Swansea Panel fell into the disputable class that might be certified by E or F, of the Silicosis Medical Boards, on one day, but rejected on another, or certified by one but rejected by the other on the same day. This review suggests that variability of opinion on certification is a matter of considerable importance in the present work of the Silicosis Medical Boards.

TABLE 6
THE PROPORTION OF "DIFFICULT" FILMS FOUND IN PRACTICE

Sample of 200 films from Silicosis Boards at :	Level of abnormality in relation to standard films		
	No. below	No. between	No. above
Cardiff : first 95 films ..	61	19	15
second 105 films	52	21	32
Total	113 (57%)	40 (20%)	47 (24%)
Swansea : first 100 films..	43	25	32
second 100 films	39	23	38
Total	82 (41%)	48 (24%)	70 (35%)

The question, "Is this film certifiable?" is often asked. It is evident from our finding that, with the present methods of assessment of films, this question can only be answered confidently in cases with obviously normal or grossly abnormal films. In films with intermediate degrees of abnormality, such as those which we selected for our experiment from the Silicosis Medical Board, the decision can only be given with limited confidence. A measure of this confidence for each film is provided by the proportion of our ten observers who gave a certifying opinion upon it. Table 7 shows, for the Silicosis Board films, the number and percentage of the twenty opinions given on each film by our ten observers that were "certifying" opinions (that is, opinions of category 4 or 5). The average percentage (44 per cent.) of those films that had, in fact, been certified is higher than that (22 per cent.) of those that had not been certified, but there is considerable overlap between the groups, some of those actually certified being less often "certified" in this experi-

ment than some of those actually rejected. When the opinions of observers E and F (of the Silicosis Board) are considered alone, the overlap still exists, but the actually certified films were "certified" on the average three times more often than the actually rejected films, instead of, for the whole group of observers, only twice as often.

PART II. FURTHER ANALYSIS

Arithmetical Methods.—Hitherto we have discussed the observers' inconsistencies and the disparities between their opinions in terms only of simple differences of categories. We felt that it might be interesting to go further into these questions, and to examine any trends of opinion which might be found and to derive indices which would measure them. Factors of interest include the relative accuracy of observers in the interpretation of chest radiographs of pneumoconiosis and the relative difficulty of interpretation of the different

TABLE 7
FREQUENCY OF "CERTIFICATION" OF FILMS ACTUALLY CERTIFIED OR REJECTED BY THE SILICOSIS MEDICAL BOARD

Film No.	Observers A-K		Observers E and F	
	No. of certifying opinions	Fre- quency of certi- fication (%)	No. of certifying opinions	Fre- quency of certi- fication (%)
<i>Rejected :</i>				
1	8	40	1	25
2	7	35	1	25
3	4	20	1	25
4	1	5	0	0
5	4	20	1	25
6	10	50	2	50
7	2	10	0	0
8	1	5	0	0
9	0	0	0	0
10	7	35	3	75
11	2	10	1	25
12	5	25	1	25
13	7	35	1	25
All films	58	22	12	23
<i>Certified :</i>				
1	3	15	2	50
2	3	15	2	50
3	5	25	1	25
4	12	60	4	100
5	13	65	4	100
6	14	70	4	100
7	9	45	3	75
8	16	80	3	75
9	11	55	4	100
10	10	50	3	75
11	1	5	1	25
All films	97	44	31	70

groups of films and stages of disease included in our material. In order to do this we have treated the category numbers employed in our experiment as if they were actual numbers and have used them for statistical processes, such as the averaging to which we have already referred.

To do this is to make the assumption that the stages of pneumoconiosis represented by the five category numbers lie on a linear scale of increasing radiographic abnormality. How justifiable this assumption is it is not easy to decide, but two possible ways in which it might be considered false can be considered. First, the range of abnormality covered by each category might not be the same, one category perhaps extending over a wider range than the others. It would then follow that errors or inconsistencies of diagnosis would have different significance at different places in the range.* Second, there is a difference of character in category 5 which makes it rather distinct from the others. Films were placed in this category if in addition to "reticulation" the observer considered there was definite nodulation or an area of coalescent shadowing. Thus, while the distinction between the first four categories was chiefly quantitative, a qualitative element entered into the differentiation of category 5.

We have proceeded to use the category-numbers

* This point has been investigated, and the results suggest that category 1 covered a slightly wider range than the others. The method used was to find positions in the range of abnormality where the upper boundaries of categories 1, 2, 3, and 4 should fall in order that the average amount of disagreement over each film remaining after the bias of each observer was allowed for should be least. These positions were found to be such that categories 1, 2, 3, and 4 occupied respectively 32, 21, 26, and 21 per cent. of the range from absolute normality to the lower boundary of category 5, instead of occupying 25 per cent. each.

for calculation without making any allowance for the effects of these or other factors. It seems reasonable to suppose that the validity of our results will not be seriously affected, in particular if the results are regarded as indicating general trends rather than precise measurements.

Bias of Observers.—We now proceed to see if the discrepancies of opinion that have been discovered show any consistent trends, such as persistent bias of certain observers towards over- or under-classification of the films. The existence of such trends would be demonstrated by differences between the average opinion of all observers and the average opinion of a particular observer in the different groups of films.

These differences are given in Table 8, and are presented diagrammatically for the whole group of films in fig. 7. It will be noted from the table that with a few exceptions each observer does show a consistent bias from the general average in each group of films. Observer D, for example, places the films in categories higher than the average and observer K in categories lower than the average.

In attempting to assess the degree of abnormality in a radiograph of pneumoconiosis the observer must compare it with mental pictures—or standards—built up from his past experience of such radiographs. These mental standards may be entirely subjective, but they may include partly objective criteria such as the extent of the lung field in which abnormality appears or the degree of obscuration of the lung markings. The fact that the bias of the individual observers' average opinions is consistent in its direction throughout the various groups of

TABLE 8

DIFFERENCES BETWEEN THE AVERAGE OPINION OF EACH INDIVIDUAL OBSERVER AND THE AVERAGE OPINIONS OF ALL OBSERVERS

Observer	Victor Medium	Victor Soft	Victor Hard	H. and A films	Si.B. films	All films
A	+0.073	+0.020	+0.234	+0.200	+0.185	+0.146
B	−0.027	−0.202	+0.067	−0.116	−0.315	−0.129
C	−0.027	+0.202	−0.044	+0.400	+0.310	+0.214
D	+0.399	+0.572	+0.652	+0.268	+0.456	+0.409
E	−0.227	−0.147	−0.321	−0.366	−0.106	−0.242
F	+0.306	+0.520	−0.155	+0.150	+0.435	+0.268
G	+0.323	+0.353	+0.790	+0.050	+0.144	+0.244
H	+0.057	−0.313	+0.290	−0.050	+0.081	+0.018
J	−0.227	−0.480	−0.599	+0.250	−0.461	−0.124
K	−0.644	−0.702	−0.877	−0.783	−0.731	−0.732

FIG. 7.—Differences between average opinion of each observer and average opinion of all observers. The average opinion of all observers is represented by the horizontal dotted line and the average opinion of each individual observer by solid circles. The vertical axis is marked in category numbers.

films suggests that during the experiment each observer derived his own individual set of mental standards from the written definition of the categories, and from his own experience. The smallness of the average deviations shows that, with the exception perhaps of observers D and K, the mental standards of all the observers were very similar.

Inconsistency of Observers.—The inconsistency found between the first and second opinion given by each observer might be due either to the difficulty he found in matching the films with his mental standards or to a change in his mental standards between the two occasions, or to both.

A change in an observer's mental standards between the two occasions would be suggested by

finding a difference between the average of his first and second readings which should be consistent in sign for each group of films.

The differences between the average of the first and of the second opinion of each observer are given in Table 9 and are plotted diagrammatically in fig. 8. It will be seen that while some observers (for example B, C, D, and E) on the average maintain their mental standards practically unchanged on the two occasions, others (for example G, H, J, and K) noticeably lowered their mental standards between the two readings, thus giving a higher average opinion on the second occasion, while observer A raised his slightly, giving a lower average opinion on the second occasion.

Skill and Experience.—The observers in this trial all had different experience of the radiographs of coalminers' pneumoconiosis, and it is interesting to see if the skill each observer showed can be related in some way to his experience.

The particular kind of skill that was necessary in this experiment can be separated into two components. The observer must first be able to assess the degree of abnormality of the radiograph in terms of the categories that were adopted, and, secondly, he must be able to perform this classification consistently on two occasions. For the consideration of the first type of ability we are faced with the difficulty, discussed by Yerushalmy (1947) of establishing the correct category of each film with which the various opinions of the observers may be compared.

TABLE 9
MEAN DIFFERENCES BETWEEN SECOND AND FIRST READINGS OF EACH OBSERVER

Observer	Victor Medium	Victor Soft	Victor Hard	H. and A. films	Si. B. films	All films
A	-0.133	-0.111	-0.444	-0.133	-0.250	-0.186
B	-0.067	-0.111	-0.333	-0.033	-0.083	-0.088
C	-0.100	0	-0.333	-0.067	0	-0.078
D	0	+0.222	-0.111	-0.067	-0.125	-0.039
E	+0.233	0	0	-0.067	+0.042	+0.059
F	+0.133	0	+0.111	+0.033	+0.250	+0.118
G	+0.300	+0.111	+0.444	+0.300	+0.417	+0.324
H	+0.233	-0.111	+0.111	+0.233	+0.458	+0.245
J	+0.067	+0.333	+0.111	+0.167	+0.458	+0.216
K	+0.300	+0.444	+0.667	+0.167	+0.250	+0.294
(L)	(+0.133)	(+0.667)	(0)	(+0.300)	(-0.167)	(+0.088)

FIG. 8.—Changes in the average opinion of each observer between the first and second readings. The change in the average opinion of each observer on the second reading is represented by the distance, positive or negative, between the appropriate solid circle and the horizontal dotted line. The vertical axis is marked in category numbers.

We have approached the question by trying to decide the category into which each film would be placed if the ten observers were discussing it in consultation. Presumably this would be, in most cases, the category into which the largest number of observers placed the film. In some cases, however, this "modal" opinion was unsatisfactory. For example, one of the H and A films was assessed during the trials as follows :

9 opinions of category 2
5 opinions of category 3
4 opinions of category 4
2 opinions of category 5

The modal opinion here is category 2, but the majority of the observers put the film into a higher category. Consultation would probably lead to acceptance of the category nearest to the average as the final opinion (in this case category 3), and this is the rule we have adopted. Each film was given a *standard category* which was that category nearest to the average opinion of all observers. The category nearest the average, and not the average itself, must be used as the standard, since by the terms of the experiment the observer had to allocate each film to a particular category and could not use fractions of categories for border-line films. The standard category of the corresponding medium film was always taken as the standard category of Victor Soft and Hard films.

The lack of bias of an observer at classifying the films can now be measured by the smallness of the deviation of his opinion, on the average, from the standard categories. In the same way, his consistency of classification can be measured by the average difference between his first and second opinions on the same film. In each case the difficulty of dealing with positive and negative deviations can be met by taking the average of the squares of the deviations.

In this way we obtain two indices, the Index of Disagreement (I_D) and the Index of Inconsistency (I_C), which measure the two components of the skill shown by an observer. I_D will be zero when the observer agrees with the standard category of each film, and it will increase with increasing amount of disagreement with the standards up to a maximum which depends on the distribution of standard categories in the group of films but which cannot exceed 16. In fact, the indices of the observers seldom exceed unity in any group of films. When the numbers one to five were allocated at random as assessments of the films, the index of disagreement was found to be 4.8, so we might expect the I_D of a completely unskilled observer to be of this order of magnitude. I_C will be zero for a perfectly consistent reader, and will increase with the degree of inconsistency of the reader to a maximum of 8 attainable only if extreme and opposite opinions are expressed on every film.

Symbolically, the two indices are given by :

$$I_D = \sum_{i=1}^n (x_i + y_i - z_i)^2 / 4n$$

$$I_C = \sum_{i=1}^n (x_i - y_i)^2 / 2n$$

where x_i and y_i are the first and second opinions of an observer on the i^{th} film of a group of n , and z_i is the standard category of the film.

Neither the index of disagreement nor the index of inconsistency alone gives a fair index of the "skill" of each observer in the sense in which we are defining it. A good observer might consistently differ from the standard opinion (perhaps correctly if some absolute standard could be achieved) so that I_D alone is not sufficient as a measure of skill. Similarly I_C on its own is inadequate, for an observer intent on consistency might formulate rules for the assessment of difficult films which would ensure his two opinions being the same on two occasions. He might, for example, decide to call all such films—or indeed every film—category 3, in which case his I_C would certainly be zero. But inevitably in such a case his I_D would be large. In order to establish a satisfactory index for the individual observer's skill we need to consider both indices in conjunction. This is most simply done by taking the average of the two, thus obtaining an index, I_s , equal to $\frac{1}{2}(I_D + I_C)$. The three indices for each observer are given in Table 10, and fig. 9 shows them graphically.

It is interesting to note the very-good performance

TABLE 10
INDICES OF DISAGREEMENT AND INCONSISTENCY,
AND COMBINED INDEX

Observer	Index of disagreement	Index of inconsistency	Combined index	Ranking
A	0.245	0.128	0.187	2
B	0.223	0.093	0.158	1
C	0.453	0.240	0.347	6
D	0.539	0.236	0.388	7
E	0.277	0.196	0.237	3
F	0.466	0.187	0.327	5
G	0.473	0.304	0.389	8
H	0.341	0.231	0.286	4
J	0.534	0.383	0.459	9
K	0.725	0.275	0.500	10
(L)	(0.738)	(0.332)	(0.535)	(11)

of observers A and B on all indices. It will be remembered that these observers had had the greatest experience of all in the accurate classification of radiographs of coalminers' pneumoconiosis. A and B are followed by the Silicosis Board observers and the tuberculosis physician who had continued work amongst miners in South Wales. The relatively inexperienced medical scientists come next, followed by G who had worked with the Pneumoconiosis Research Unit and the Silicosis Medical Board; his combined index is lowered by a high index of inconsistency due to a large systematic shift in his opinions on the second reading (fig. 8). The final group of observers with least skill includes the tuberculosis physician who had not worked among cases of pneumoconiosis for some years, the London radiologist with little current experience of pneumoconiosis, and observer K who has a low index of skill owing to his relatively consistent under-estimation of the films (see fig. 3). It appears that the

TABLE 11
AVERAGE VALUES OF THE DIFFICULTY-OF-AGREEMENT INDEX
(No. of films in each group given in brackets)

Group	Standard category				
	1	2	3	4	5
Victor Medium ..	0.50 (3)	0.61 (7)	0.49 (9)	0.36 (6)	0.36 (5)
Victor Soft ..	—	1.00 (3)	1.15 (4)	0.35 (2)	—
Victor Hard ..	—	0.57 (3)	0.58 (4)	0.38 (2)	—
H. and A. films ..	0.25 (4)	0.54 (10)	0.64 (7)	0.40 (8)	0.35 (1)
Si. B. films ..	—	0.60 (1)	0.69 (18)	0.59 (5)	—

combined index does give the observers rankings which might be expected from their varying experience.

Relative Difficulty of Films.—It has been shown in Table 1 that the range of opinion expressed on certain films was greater than on others. It seemed worth investigating this point further, although the experiment was not designed with this end in view and the material is not entirely satisfactory for the purpose.

FIG. 9.—Diagrammatic representation of indices of disagreement, consistency, and the combined index of "skill" of observers A to L. The indices of disagreement of each observer are plotted on the left hand vertical line and the indices of consistency on the right hand vertical line. The index of "skill" derived from these by averaging is given on the central vertical line. The values for the index of disagreement are much greater than the values of consistency because of the much greater divergence found between the opinions of the different observers than that found between the two opinions of each observer.

As a measure of this relative difficulty of classification of a film the mean square of the difference between each expressed opinion and the standard category of the film may be used, that is :

$$\sum_{i=1}^{20} (x_i - z)^2 / 20$$

where x_i is an individual opinion of the film and z is its standard category. (There are, of course, twenty individual opinions on each film.)

The size of this index increases with the amount of disagreement found in the classification of each film, both between the observers and within an observer's own two opinions. Fig. 10 is a histogram of the number of films falling within certain ranges of value of this "difficulty-of-agreement" index. It will be seen that the distribution is roughly of Normal form between the limits 0 and 1.1, and that there are four films above this range. Three of these were Victor Soft films, and the remaining one a "Si-B" film. This suggests that technique may influence the size of the index of film-difficulty, and that it might be of interest to examine the relative difficulty of the different groups of films used in this experiment.

FIG. 10—Histogram of the distribution of films according to the size of their difficulty of agreement index.

The average values of the difficulty-of agreement index for the different standard categories are given in Table 11. It appears that the difficulty of the different standard categories as measured by this index agrees with what would be expected. Categories 1, 4, and 5, which represent the extremes of the range of abnormality with which we are concerned, were found easier to agree about than categories 2 and 3, which subdivide the intermediate range.

There is some disagreement among various workers concerned with pneumoconiosis as to the best radiographic technique to use for the diagnosis of the disease. It might be thought that the difficulty-of-agreement index could be used to obtain an objective measure of the value of different techniques. The results of our experiment do not encourage this supposition, for we find that the Victor Medium and "H and A" films have much the same indices within corresponding categories, although it would be generally agreed that the Victor Medium films were superior in technique. Further, although the Victor Soft films have very high indices, the Victor Hard films have low indices despite gross over-penetration. However, this latter conflict with expectation may be due to the small number of films in each group. Further work is necessary to establish the value of the index.

Effect of Practice and Fatigue.—It would be of interest to know whether the performance of an observer classifying radiographs gets better or worse towards the end of a long series; in other words whether his skill improves temporarily as a result of practice, or whether fatigue has a dominating effect and the last films are read with less skill than the first. In this experiment the films were read first in order 1 to 102 and second in the reverse order 102 to 1. In consequence, the first and second readings of each film occurred at the same interval from the beginning and end respectively of each observer's trial and there is no way of detecting changes of an individual's consistency with time. But by treating all first and second readings separately we can detect trends in the magnitude of the "difficulty-of-agreement" index by calculating its linear regression against film serial-number. The resulting equations are :

$$\text{1st Trial : } V = 0.75 - 0.0022n.$$

$$\text{2nd Trial : } V = 0.55 + 0.00014n.$$

where V stands for the value of the "difficulty-of-agreement" index and n is the serial-number of a film. In each trial therefore, V is slightly smaller at the end than at the beginning, so that there is a suggestion that the observers' assessments agreed better after they had been reading films for some time than when they first begin the trials. Further investigation would obviously be necessary before it could be said definitely that practice had more effect in increasing the skill with which films were read than fatigue has in decreasing it.

Discussion

A certain amount of care is necessary in drawing general conclusions from the results of this experiment. The set of films used was specially made up with certain problems in mind, and was not necessarily typical of any set that might be encountered in practice. The comparison of the skill of the different observers has been made in relation only to the problem of classification in pneumoconiosis, and obviously would not hold for work in other fields. The assumptions that have been made in treating the five categories as if they were numbers have already been pointed out; these assumptions may perhaps be more easily appreciated if it is considered that the categories might equally well have been called A, B, C, D, and E, and the meaning of the "average opinion" of a film, for example, would then have been much less obvious. The purpose of the investigation was to establish the existence of inconsistency in x-ray interpretation and to provide measures of it, in order that the value of other methods of interpretation might subsequently be estimated. We believe that for these purposes the methods we have employed are valid. If the reservations made above are borne in mind, certain general conclusions may with confidence be drawn from the results.

The assessment of the degree of abnormality due to pneumoconiosis in any film by even an experienced observer is not an absolute one. It will frequently differ to an extreme degree from the opinions of other observers and will often differ to a lesser degree from the same observer's opinion on the same film on a subsequent occasion. We have suggested that these varying opinions are due to the observer matching the film under consideration with mental pictures derived from his experience. These mental pictures vary from observer to observer and within one observer from time to time.

An analogy with another sort of clinical assessment may be used to illustrate this point. A patient's hæmoglobin level may be measured by an assessment of the redness of a diluted specimen of blood. If the method used by the radiologist were employed, the pathologist would just hold the specimen to the light and classify it in comparison with a mental standard of redness. In fact, the pathologists have found this method unsatisfactory and have come to use standard colour tubes for direct comparison and have thus achieved a far greater degree of accuracy. But even so, it has been shown that standards may vary in their precise tints and that the technique of comparison may greatly influence the opinion given by the pathologist. For greater consistency it has been found necessary to lay down very precise instructions for such comparative estimations (Macfarlane, 1945).

It might be suggested that radiologists should follow this method in the diagnosis of pneumoconiosis, using standard films for comparison. The problem in this case is not quite so straightforward, however. Radiographs of pneumoconiosis differ not only in the degree of abnormality but also in type of abnormality. For this reason, standards of each type would be required. This difficulty, however, should not be an insuperable one. Further, the present experiment has shown the great influence of radiographic technique on the opinion of most observers. This difficulty could be lessened by strict regulations of the radiographic technique to be employed.

The question of the accurate radiological diagnosis of pneumoconiosis is one of considerable importance, in view of the weighty decisions which rest upon it. At present, the chest radiograph forms the main basis on which decisions are reached concerning the occupational future and economic compensation of men working in industries with a silicosis risk. For such decisions, the inaccuracies of diagnosis revealed in the experiment here described are far too great. The inaccuracies are also of great importance in research. For example, they will seriously affect the results of radiological surveys

undertaken to estimate the prevalence of the disease under different environmental conditions. They will also affect the results of studies of progression, of methods of prevention and treatment, and of the value of special radiological methods, such as miniature films, in the diagnosis of the disease.

Further study is necessary to establish means of reducing and compensating for these inaccuracies. We have already carried out a preliminary trial with standard films which leaves little doubt that, with their assistance, it is possible for observers to achieve a greater degree of agreement and consistency in classifying radiographs of pneumoconiosis than that shown in the experiment we have reported. We are now engaged upon an experiment intended to provide a more precise assessment of the uses and limitations of such standards. The results of this experiment will be reported later.

Summary

1. Very little is known of the errors inherent in the detection of pneumoconiosis by means of radiographs and in the classification of radiographs of pneumoconiosis according to the severity of the disease, although accurate detection and classification is essential in research, in industrial health surveys, and especially in reaching decisions concerning compensation for industrial diseases of the lung.

2. Ten doctors, with varying experience in the radiological diagnosis of coalminers' pneumoconiosis in South Wales, were asked to classify 102 radiographs of early coalminers' pneumoconiosis into five categories. Each doctor classified all the films on two separate occasions. In addition two consultant radiologists, working in London, subsequently undertook the same repeated classification of the radiographs.

3. The opinions of these observers were found to differ to a remarkable degree, both amongst themselves, and, to a lesser extent, from the one occasion to the other. The divergence of opinion between the two consultant radiologists was as great as between the ten other observers. The variation of opinion was greatest in films that were neither normal nor grossly abnormal.

4. There was serious disagreement concerning the limits within which a film may be regarded as normal. Thus, there were 32 films which were considered to be within normal limits by one or other of the observers, but to show "certifiable" pneumoconiosis (that is, sufficient to merit compensation under the Workmen's Compensation Acts) by at least one other of the observers.

5. The number of films considered "certifiable" on both occasions by the various observers ranged from 5 to 45. On one of their readings, two colleagues from the Silicosis Medical Boards "certified," respectively, 34 and 53 of the films. A review of cases applying to the Silicosis Medical Boards during the first half of 1948 showed that 22 per cent. of them fell into the class on which the opinions of these doctors concerning certification might differ.

6. Radiographic technique was found to have a seriously disturbing effect on classification. Under-exposed films were generally considered to show more disease than over-exposed films of the same case taken on the same day. Some observers were less affected than others by this factor.

7. A statistical analysis of the results was made in order to show the general trends in the differences of opinion between the observers and between the opinions of the same observers on the different occasions, and to derive indices of the relative skill of the observers, and of the relative difficulty of diagnosis of the various films. The validity of the methods used is discussed and is accepted for these comparative purposes.

8. It is suggested that the amount of inconsistency of opinion that was found is intolerable in view of

the important consequences of the radiological diagnosis of pneumoconiosis, and that the present method by which films are diagnosed by simple scrutiny should be superseded by a method of comparison of unknown films with accepted standard reference films. A further experiment is being planned to investigate the possibilities of this method.

We wish to pay tribute to the courage of the observers who allowed us to examine their competence by such stringent standards as those that we have employed. We wish also to thank Dr. P. Hugh-Jones for his advice in the design of the experiment and for his help in some of the analyses.

REFERENCES

- Birkelo, C. C., and others (1947). *J. Amer. med. Ass.*, **133**, 359.
 Coalminers' Pneumokoniosis Scheme (1943). H.M.S.O. 885 and 886.
 Comroe, J.H., Jr., and Botelho, S. (1947). *Amer. J. med. Sci.*, **214**, 1.
 Derryberry, M. (1938). *Publ. Hlth Reps., Wash.*, **53**, 263.
 Davies, I., and Mann, K. J. (1948). "Proceedings of Ninth International Congress on Industrial Medicine."
 Hart, P. d'A., and Aslett, E. A. (1942). *Spec. Rep. Series, Med. Res. Counc. London*. No. 243.
 Hebert, G. T. (1939). *Tubercle*, **20**, 145.
 Huws Jones, R. (1938). *J. roy. Statist. Soc.*, **101**, 1.
 Macfarlane, R. G. (1945). *Spec. Rep. Ser. Med. Res. Counc. London*, **252**, 59.
 Yerushalmy, J. (1947). *Publ. Hlth Rep., Wash.*, **62**, 1432.