THE CLASSIFICATION OF CHEST RADIOGRAPHS FOR COALWORKERS' PNEUMOCONIOSIS

A STUDY OF THE PERFORMANCE OF TWO READERS OVER A PERIOD OF SIX YEARS

BY

J. R. ASHFORD

From the Scientific Department, National Coal Board

(RECEIVED FOR PUBLICATION NOVEMBER 26, 1959)

The paper is concerned with the performance over the past six years of the two doctors responsible for the classification of chest radiographs obtained by the National Coal Board's Pneumoconiosis Field Research. A brief description is given of the categories of pneumoconiosis used in the Research and of the procedures adopted to decide the final or "definitive" classification. The readings obtained are then considered in detail and three main points emerge from the analysis. First, it is shown that there have been marked systematic fluctuations in the reading levels of these two very experienced observers. In spite of these individual variations it is found that the standards applied on the definitive readings, which are based on the joint discussion by the two doctors of films on which their individual readings are inconsistent, remain remarkably constant over a period of several years. The third main feature of the analysis is the considerable differences in the magnitude of the reading errors between the various categories of pneumoconiosis. It is concluded that, although the existing procedures provide a satisfactory way of estimating the prevalence of pneumoconiosis, it would be desirable to increase the number of film readers. The beneficial effect of the system whereby a running check is kept on film reading standards is also established.

As part of the National Coal Board's Pneumoconiosis Field Research (Fay, 1957), periodic medical surveys are carried out at the selected collieries at intervals of about four years. These surveys have been in progress since late 1953. Up to the present time the initial medical examinations have been completed at the 26 collieries* associated with the Research, and follow-up examinations have been completed at five collieries. One of the most important components of the medical surveys is the taking and subsequent classification of a 14 in. × 14 in. postero-anterior chest radiograph for each man presenting himself for examination. There are two mobile chest radiography units, which respectively cover the collieries in the north and south of the country, and each is in the charge of an experienced doctor. Every film is classified at least once by each of the two doctors and these readings, of which more than 100,000 have been made during the past six years, provide what is probably a unique

body of information about the performance of two experienced film readers over a long period of time.

Classification of Films

The categories of pneumoconiosis used in the medical surveys are those currently accepted by international agreement (International Labour Office, 1953), which recognizes the distinction between simple and complicated pneumoconiosis or progressive massive fibrosis (P.M.F.).† Simple pneumoconiosis is subdivided on a quantitative basis into three categories, increasing in degree of abnormality, and numbered 1-3, category 0 being reserved for films which are within normal limits. Films showing evidence of complicated pneumoconiosis are separately classified into four categories —A, B, C and D.

In view of the considerable element of doubt associated with a single reading of pneumoconiosis,

* In 1958 one of the original 25 selected collieries was closed and it was subsequently replaced by another colliery in the same coalfield.

† The classification of P.M.F. by the Pneumoconiosis Field Research has subsequently been revised, in the light of recent experience.

even by an experienced radiologist (Fletcher and Oldham, 1949), the final classification of the radiographs is based on a series of readings by both doctors (Observers A and B). During the period under consideration three different procedures have been applied for this purpose as follows.

(a) In the first place each film was read independently on three occasions, twice by the doctor responsible for the survey and once by the other doctor. When these three "routine" readings did not differ by more than one category* and where there was no suggestion of other disease which would complicate the pneumoconiosis readings, the final or "definitive" classification was assigned on a majority basis. For all other films the definitive classification was assigned as a result of joint discussion between the two doctors. This procedure was applied at the first two collieries examined (C_1 and C_2) but was subsequently modified to guard against the possibility that the definitive readings might be disproportionately affected by the performance of the doctor responsible for the survey.

(b) On the first medical surveys at collieries C_3 - C_{24} the method of final classification by joint discussion was extended to all films on which the three routine readings were not consistent.

(c) The second change in film reading procedures took place as a result of a special study of the accuracy of the joint reading procedure (Fay and Ashford, 1960). On the basis of this investigation it was decided to discontinue the second reading by the doctor responsible for the survey and to base the definitive classification on a single reading by each observer, together with the joint discussion of disagreements. To obtain a check on the consistency of reading, a repeat classification by the doctor responsible for the survey was carried out on a random one in 10 sample of films. This revised procedure was applied on the first surveys at collieries C_{25} and C_{26} and on the second round of surveys at all collieries so far examined.

The second survey films are at present classified without reference to any film or films of the same man obtained on the first survey.† To obtain a comparison of the reading levels on successive surveys at the same colliery, a random one in 10 sample of first survey films, selected mainly from those on which there were disagreements on the first survey,‡ is intermixed with the second survey films and classified again, according to the standard procedures.

During the first round of medical surveys the various sets of readings at a colliery extended over a

period of as long as a year and it was impossible to provide the doctors with up-to-date information about their current reading levels. On the second round of surveys, however, a running check was kept on the doctors' performance and as far as possible each survey was completed and the results were analysed and discussed before the next survey was undertaken.

Distribution of Abnormality at the Various Collieries

It is apparent on general grounds that the level of agreement between repeat readings by the same observer or between single sets of readings by different observers in terms of the standard categories of pneumoconiosis will depend on the amount of abnormality in the batch of films being classified, as well as on the standard of performance of the observers. For example, a batch containing a higher proportion of "normal" films (most of which will be assigned to category 0 on each set of readings) will be read more consistently than a batch containing a lower proportion of such films. Thus, in any realistic study of reading errors it is essential to take account of the distribution of abnormality in the films on which the results are based. A more detailed quantitative treatment of the problem (Ashford, 1959) lends further support to this conclusion.

There is a wide variation in the prevalence of pneumoconiosis at the selected collieries on the first medical surveys. The percentages of men assigned to category 0 lie mainly between 72.5 and 95.0, although figures of less than 70% were recorded at four collieries. The proportions classed as category 1 vary from 2.5 to 17.5%, with the majority lying between 5.0 and 15%, and the proportions classed as category 2 tend to be slightly lower. At all except five collieries less than 2.5% of the population were classed as category 3 and at all except three, less than 5% were classed as P.M.F. On average, approximately 82% of films were classed as category 0, 10% as category 1, 5% as category 2, 1% as category 3 and 2% as P.M.F. on the definitive readings. At collieries where the second medical surveys have been carried out the results obtained are similar to those recorded on the corresponding first surveys.*

Consistency of the Duplicate Readings by the Same Observer

A measure of the error associated with the film reading process may be obtained by comparing

* Readings of simple and complicated pneumoconiosis on the same films were considered to differ by more than one category.

† Studies of the relative merits of side-by-side and independent classification of the films of the same man on successive surveys are currently in progress.

‡ At Collieries C_7 - C_{13} the sample consisted entirely of the disagreement films. At Collieries C_4 onwards a quarter of the sample consisted of films on which the first survey readings were consistent.

* It should be borne in mind that, due to labour turnover, there may be appreciable differences in the populations examined at a colliery on successive surveys.

TABLE 1
FIRST SURVEY AT COLLIERY C₆; DUPLICATE READINGS BY OBSERVER B
No. OF FILMS

SECOND READING	FIRST READING							
	Category	0	1	2	3	A	B, C, D	All
	0	441	30					471
	1	41	65	12		1		119
	2	1	29	57	3	1		91
	3			2	16	1		19
	A					21	1	22
	B, C, D						5	5
	All	483	124	71	19	24	6	727
Percentage of consistent readings								83.2

duplicate readings carried out by the doctor responsible for the survey. At any given colliery these results may conveniently be expressed in terms of a "two-way" table and a typical example is given in Table 1, which refers to the readings by Observer B at colliery C₆. The differences between the two sets of readings are revealed in tables of this kind by the number of films in the "off-diagonal" positions. For example, a total of 30 films was read as category 1 on Observer B's first reading and as category 0 on his second, whereas 41 of the films read as category 1 on his second reading had been read as category 0 on his first reading. The differences in successive readings may be attributed to two distinct sources of error. In the first place there may be systematic differences in the placing of the various category boundaries on the two sets of readings, which will be indicated by corresponding differences in the marginal totals. On this particular batch of films Observer B tended to read higher (*i.e.* to read more abnormality) on the second set of readings than on the first, particularly on the boundary between categories 1 and 2, and statistical tests* show that this difference is significant. The second source of error is the variability of the reading process itself, which leads to different results between two sets of readings even when there is no change in average reading levels. The combined effect of the two sources of error on Observer B's readings on colliery C₆ was that different duplicate readings were recorded on 16.8% of the films. As is generally found, almost all the disagreements were of one category only.

* In this context the testing of the significance of the difference in average reading levels on the first and second occasions is based on the assumption that, if the same standard were applied on both sets of readings, the expected number of films in any particular position in the two-way table above the diagonal must be equal to the expected number of films in the corresponding position below the diagonal.

Details of the percentage of consistent duplicate readings on the first survey films at all collieries and on the second survey films at the first five collieries so far examined are given in Fig. 1, where these results have been plotted against the percentage of men classed as category 0 on the corresponding set of definitive readings. In this figure a distinction has been made between the two doctors and between the first and second survey readings. As might have been expected on general grounds the proportion of consistent readings shows a marked tendency to increase with increasing proportions of "normal" films in the batch, from 77% at a colliery with just less than 60% of normal films to more than 97% at a colliery with more than 95% of normal films. There are, however, considerable differences in the levels of consistency at collieries with similar distributions of abnormality, which suggest that other factors (such as systematic differences in reading levels or in the technical quality of the films) are also involved. The percentages of normal films at the collieries surveyed by Observer A tend to be higher than those at the collieries surveyed by Observer B and the consistency of his readings is correspondingly greater. On the whole, the second survey readings at a colliery are more consistent than the corresponding first survey readings.

When the average levels of classification applied by Observer A on his duplicate readings on the first survey films are compared, there is no evidence of any significant difference in his standards on seven of the 12 surveys for which he was responsible. At the remaining five collieries his second reading was significantly higher than his first. Observer B was responsible for the first surveys at 14 collieries and at four of these there was no significant difference between his first and second readings. At seven of the remaining 10 collieries his second reading was

FIG. 1.—Relationship between amount of abnormality and percentage of consistent duplicate readings by the same Observer.

significantly lower than his first and at the other three collieries his second reading was significantly higher.

On the first round of surveys the readings of the films obtained at any one colliery extended over a period of as long as a year and the observed differences between the duplicate readings thus reflect periodic fluctuations in average standards of performance. It is, however, interesting to note that at all collieries where the differences between Observer A's duplicate readings were significant he tended to read higher on the second reading than the first. This trend may well be associated with the fact that the first set of readings forms the basis for any immediate medical advice which may be given to the subject. On the second round of surveys the duplicate readings are made within a comparatively short space of time and at the five collieries for which results are so far available there have been no significant differences between the corresponding sets of readings.

Consistency of the Readings by Different Observers

As it appears from a study of the duplicate readings that the standards of classification of a single observer may vary in time, the direct comparison of the readings by different observers is subject to some uncertainty. For example, if the doctor responsible for the survey at a colliery has adopted different standards on his duplicate read-

ings, it is possible that the readings by the other doctor may differ significantly from one set but not from the other. For this reason it is more convenient to compare each set of routine readings individually with the best available estimate of the true position. In a subjective process such as radiograph reading there can be no absolute standards and the most reasonable substitute is provided by the definitive readings, which (except at collieries C_1 and C_2) represent the agreed opinion of both observers. The use of the definitive readings for this purpose is supported by the results of a detailed quantitative study of the film reading process (Fay and Ashford, 1960), which shows that the proportion of films assigned to an incorrect definitive category is small.

Each set of routine readings may be conveniently compared with the corresponding definitive readings in terms of a two-way table, and a typical example is given in Table 2, which refers to the three sets of routine readings at colliery C_{11} . At this particular colliery both sets of routine readings by Observer A were in close agreement with the definitive classification but Observer B tended to read significantly higher. The proportions of consistent readings were respectively 97.2, 97.6 and 96.9% on Observer A's first and second readings and Observer B's single reading.

When the proportions of films on which the routine and definitive readings were consistent are compared with the percentage of men classed as category 0 on the corresponding set of definitive

TABLE 2
FIRST SURVEY AT COLLIERY C₁₁; COMPARISON OF ROUTINE AND DEFINITIVE READINGS

(a) OBSERVER A'S FIRST READING

No. OF FILMS

Observer A	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	1,190	12					1,202
	1	11	24	6		1		42
	2		5	11	1			17
	3				3			3
	A					3		3
	B, C, D						2	2
	All	1,201	41	17	4	4	2	1,269
Percentage of consistent readings								97.2

(b) OBSERVER A'S SECOND READING

Observer A	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	1,195	14					1,209
	1	6	24	5		1		36
	2		3	12	1			16
	3				3			3
	A					3		3
	B, C, D						2	2
	All	1,201	41	17	4	4	2	1,269
Percentage of consistent readings								97.6

(c) OBSERVER B'S READING

Observer B	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	1,180	8					1,188
	1	20	27	2				49
	2		5	13				18
	3			1	4			5
	A	1	1	1		4		7
	B, C, D						2	2
	All	1,201	41	17	4	4	2	1,269
Percentage of consistent readings								96.9

readings it is found that the amount of agreement increases with the proportion of "normal" films. The general level of consistency is, however, considerably higher than that between the duplicate readings by the same observer. This confirms that the definitive readings are subject to appreciably less error than any single set of routine readings. There is on the whole very little difference between the

general levels of performance of Observers A and B. In general, the consistency of the second survey readings tends to be higher than that of the first survey readings at the same colliery.

The variation in time of the standards applied on the corresponding routine and definitive readings is illustrated in Tables 3 and 4, which refer respectively to the readings at the collieries surveyed by

TABLE 3
COMPARISON OF ROUTINE AND DEFINITIVE READINGS AT COLLIERIES SURVEYED BY OBSERVER A

Observer	Colliery														
	C ₁	C ₃	C ₅	C ₇	C ₉	C ₁₁	C ₁₃	C ₁₅	C ₁₇	C ₁₉	C ₂₁	C ₂₃	C ₁ *	C ₃ *	C ₅ *
Observer A's first reading	0	0	+	—	—	0	—	—	+	+	+	+	0	+	+
Observer A's second reading	0	+	+	—	—	—	—	0	+	+	The second routine reading was discontinued at this point				
Observer B's reading	+	+	0	+	—	+	—	0	—	0	0	0	0	—	—
Date of survey	12/53	3/54	5/54	9/54	12/54	2/55	4/55	6/55	12/55	3/56	5/56	6/56	11/57	4/58	9/58

Notes + = Significantly higher than definitive reading.
0 = No significant difference between routine and definitive readings.
— = Significantly lower than definitive reading.
* Second survey.

TABLE 4
COMPARISON OF ROUTINE AND DEFINITIVE READINGS AT COLLIERIES SURVEYED BY OBSERVER B

Observer	Colliery															
	C ₂	C ₄	C ₆	C ₈	C ₁₀	C ₁₂	C ₁₄	C ₁₆	C ₁₈	C ₂₀	C ₂₂	C ₂₄	C ₂₆	C ₂ *	C ₄ *	C ₆ *
Observer B's first reading	0	+	0	+	0	0	+	—	+	+	0	—	—	—	+	0
Observer B's second reading	+	+	0	0	0	0	—	—	+	0	0	—	—	The second routine reading was discontinued at this point		
Observer A's reading	0	—	—	—	+	+	+	+	—	—	0	+	+	+	—	0
Date of survey	11/53	2/54	5/54	6/54	9/54	11/54	2/55	5/55	6/55	11/55	1/56	4/56	6/56	11/57	4/58	6/58

Notes—As for Table 3. * Second survey.

Observers A and B. These results are based on a series of two-way tables of the same form as Table 2, and the routine readings have been classed as either (i) not significantly different from the definitive readings, (ii) significantly higher than the definitive readings or (iii) significantly lower than the definitive readings. This comparatively simple index of comparison was adopted in the absence of any more convenient method of presentation, although a detailed mathematical analysis of the results (Fay and Ashford, 1960) does provide a quantitative measure of the differences in reading levels. When considering these results it should be borne in mind that on the first surveys the routine and definitive readings at a colliery may cover a period of as long as a year and that there may be an interval of several months between any two successive sets of readings.

The main point which emerges from Tables 3 and 4 is that the reading levels of both doctors show definite and systematic fluctuations in time, in comparison with the definitive readings. In general, it is found that when one observer is reading high the other is reading low, and vice versa, although this is not always the case. The definitive readings depend largely on the joint discussion sessions, which may

take place many months after the routine readings, and they appear to represent the average of the current reading levels of the two observers.

The average reading levels of the two doctors on the first round of medical surveys are compared with the definitive readings in Tables 5 (Observer A) and 6 (Observer B). Reference to the results given in Table 5 (a) shows that at the collieries surveyed by Observer A his routine readings were, on average, in close agreement with the definitive readings, the only appreciable difference being on the boundary between categories 2 and 3, where he tended to read somewhat higher than the definitive level. At the collieries surveyed by Observer B on the other hand (Table 5 (b)) Observer A tended to read lower than the definitive readings on the boundary between categories 0 and 1, and higher on the boundaries between categories 1 and 2, and 2 and 3. When all collieries are considered (Table 5 (c)) a similar trend is apparent. Reference to Tables 6 (a), 6 (b), and 6(c) shows that Observer B's reading levels were higher than the definitive levels at the collieries surveyed by both himself and Observer A. When all collieries are considered Observer A's routine readings were consistent with the definitive readings on 92.1% of films and Observer B's routine readings

TABLE 5
COMPARISON OF OBSERVER A'S ROUTINE READINGS WITH THE DEFINITIVE READINGS
FIRST MEDICAL SURVEYS

(a) COLLIERIES SURVEYED BY OBSERVER A

Observer A	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	25,578	630	31		7		26,246
	1	603	1,675	225	1	27		2,531
	2	14	248	725	42	33		1,062
	3		5	90	236	21		352
	A	3	28	59	12	264	8	374
	B, C, D	3	2	2		37	150	194
	All	26,201	2,588	1,132	291	389	158	30,759
Percentage of consistent readings								93.1

(b) COLLIERIES SURVEYED BY OBSERVER B

Observer A	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	12,612	555	23	1	10	2	13,203
	1	339	942	172	3	9		1,465
	2	11	210	632	42	20		915
	3			77	186	12		275
	A	1	25	27	14	322	17	406
	B, C, D		2	2	1	15	137	157
	All	12,963	1,734	933	247	388	156	16,421
Percentage of consistent readings								90.3

(c) ALL COLLIERIES

Observer A	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	38,190	1,185	54	1	17	2	39,449
	1	942	2,617	397	4	36		3,996
	2	25	458	1,357	84	53		1,977
	3		5	167	422	33		627
	A	4	53	86	26	586	25	780
	B, C, D	3	4	4	1	52	287	351
	All	39,164	4,322	2,065	538	777	314	47,180
Percentage of consistent readings								92.1

were consistent with the definitive readings on 91.4% of films.

The Sequences of Routine Readings

The method of classification by joint discussion was applied to all films on which the routine readings did not agree. Thus, the consistency of the routine

readings, which is a reflection of both the inter- and intra-observer variation, is of considerable interest from the point of view of the overall accuracy of the reading procedure (Fay and Ashford, 1960).

There is a marked tendency for the percentage of consistent routine readings to increase with increasing proportions of "normal" films, and the

TABLE 6
COMPARISON OF OBSERVER B'S ROUTINE READINGS WITH THE DEFINITIVE READINGS
FIRST MEDICAL SURVEYS

(a) COLLIERIES SURVEYED BY OBSERVER B

Observer B	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	23,995	749	34		24		24,802
	1	973	2,201	301	2	37	2	3,516
	2	17	368	1,295	88	45	3	1,816
	3		2	129	376	16		523
	A	8	21	34	9	594	31	697
	B, C, D		3	1		40	271	315
	All	24,993	3,344	1,794	475	756	307	31,669
Percentage of consistent readings								90.7

(b) COLLIERIES SURVEYED BY OBSERVER A

Observer B	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	13,538	331	16		6	2	13,893
	1	372	881	91		14		1,358
	2	15	170	443	31	20		679
	3		4	54	127	5		190
	A	4	8	15	2	159	9	197
	B, C, D		1	1		6	71	79
	All	13,929	1,395	620	160	210	82	16,396
Percentage of consistent readings								92.8

(c) ALL COLLIERIES

Observer B	Definitive							
	Category	0	1	2	3	A	B, C, D	All
	0	37,533	1,080	50		30	2	38,695
	1	1,345	3,082	392	2	51	2	4,874
	2	32	538	1,738	119	65	3	2,495
	3		6	183	503	21		713
	A	12	28	49	11	753	40	894
	B, C, D		4	2		46	342	394
	All	38,922	4,739	2,414	635	966	389	48,065
Percentage of consistent readings								91.4

figures reported at the collieries where three such readings were made vary from about 60 to 95%. When the results obtained on the films belonging to the various definitive categories are considered separately there are wide variations in the consistency of the routine readings from category to category. For the films assigned to definitive category 0 the proportion of consistent routine

readings increases very slightly with increasing proportions of normal films, with an average of about 92%. For films assigned to the other definitive categories, however, the proportion of consistent readings is independent of the distribution of abnormality and average figures of about 25%, 37%, 55% and 54% were respectively reported for categories 1, 2, 3, and P.M.F.

Consistency of the Definitive Readings

The reading procedures on the first round of medical surveys were such that it is not possible to obtain any estimate of the consistency of the definitive readings. During the second surveys, however, a random sample of the first survey films is classified again and a comparison of the corresponding definitive readings provides a measure of their repeatability over a long period of time. At collieries C₁-C₃ the whole of the random sample

consisted of films on which the routine readings on the first survey were inconsistent, but at collieries C₄ and C₅ a quarter of the sample was selected from films on which the original routine readings were consistent.

Details of the definitive readings recorded at the various collieries on the films on which there was disagreement on the original routine readings are given in Table 7. The main point of interest is the fact that at four of the five collieries (C₂-C₅) there

TABLE 7
COMPARISON OF DEFINITIVE READINGS ON SAMPLE OF FIRST SURVEY FILMS ON WHICH THE ORIGINAL ROUTINE READINGS WERE INCONSISTENT

(a) COLLIERY C₁

		First Survey							Second Survey
		Category	0	1	2	3	A	B, C, D	
		0	40	14	2				
		1	4	20			1		
		2			4				
		3							
		A					1		
		B, C, D						1	
		All	44	34	6		2	1	
Percentage of consistent readings									75.9

(b) COLLIERY C₂

		First Survey							Second Survey
		Category	0	1	2	3	A	B, C, D	
		0	12	5	1				
		1	7	13	6		2		
		2		9	9	1			
		3			2	2			
		A			1		4	1	
		B, C, D						1	
		All	19	27	19	3	6	2	
Percentage of consistent readings									53.9

(c) COLLIERY C₃

		First Survey							Second Survey
		Category	0	1	2	3	A	B, C, D	
		0	62	23	1				
		1	15	31	8				
		2		3	4	2			
		3				1			
		A		1	2		3		
		B, C, D					1		
		All	77	58	15	3	4		
Percentage of consistent readings									64.3

TABLE 7—*continued*(d) COLLIERY C₄

Second Survey	First Survey							
	Category	0	1	2	3	A	B, C, D	All
	0	47	22					69
	1	19	35	11		1		66
	2		7	9	2	1		19
	3			1	4	1		6
	A			1		7	3	11
	B, C, D					2	2	4
	All	66	64	22	6	12	5	175
Percentage of consistent readings								59.4

(e) COLLIERY C₅

Second Survey	First Survey							
	Category	0	1	2	3	A	B, C, D	All
	0	41	8	1				50
	1	10	38	5		1		54
	2		3	15	1	2		21
	3			2		1		3
	A		2	4		7	2	15
	B, C, D		1	1			3	5
	All	51	52	28	1	11	5	148
Percentage of consistent readings								70.3

was no evidence of any significant change in reading levels. At one colliery (C₁) the second survey readings were significantly lower than the first. Further investigation of this anomalous result showed that it was due to a temporary variation in the reading levels of one of the doctors on his second survey readings of the check films, and arrangements were made to reclassify the films concerned. The corresponding first and second survey definitive readings on the sample of first survey films on which the original routine readings were consistent are compared in Table 8. The two sets of readings were consistent on all except one of the 83 films concerned and this single difference was of one category only.

The proportions of consistent definitive readings on the films on which there was disagreement between the original routine readings vary from 53.9% to 75.9% with an average of about 65%. When considering these results it should be borne in mind that the films concerned are deliberately selected to cover the most difficult cases. They represent, on average, only about one-fifth of the total number and almost 100% of the definitive readings on the remaining four-fifths are consistent.

Discussion

The main feature of these results is the fact that the reading levels of two very experienced doctors exhibit marked fluctuations over a long period of time, as revealed both by the comparison of the duplicate readings by the same observer and of the corresponding readings by different observers. In spite of these variations the standards applied on the definitive readings appear to remain remarkably constant, even over a period of as long as five years. This indicates that the method of joint discussion, which applies to the majority of films lying on the borderline between two categories (Fay and Ashford, 1960) does lead to a reasonably stable reading standard and confirms that the film reading procedures described provide a satisfactory way of measuring the prevalence of radiological pneumoconiosis.

The presence of systematic variations in individual reading levels suggests that it would probably be desirable, in a programme of film reading such as that being carried out by the Pneumoconiosis Field Research, to increase the number of film readers. In this way any "averaging" procedure used to obtain the definitive readings should provide more

TABLE 8
COMPARISON OF DEFINITIVE READINGS ON SAMPLE OF FIRST SURVEY FILMS ON WHICH THE
ORIGINAL READINGS WERE CONSISTENT

(a) COLLIERY C₄

Second Survey	First Survey								
	Category	0	1	2	3	A	B, C, D	All	
	0	42						42	
	1		1					1	
	2			1				1	
	3				1			1	
	A								
	B, C, D								
	All	42	1	1	1			45	
Percentage of consistent readings									100.0

(b) COLLIERY C₅

Second Survey	First Survey								
	Category	0	1	2	3	A	B, C, D	All	
	0	34	1					35	
	1		2					2	
	2			1				1	
	3								
	A								
	B, C, D								
	All	34	3	1				38	
Percentage of consistent readings									97.4

stable results than can be obtained with two readers. Even amongst a large group of film readers, the use of frequent joint reading sessions to classify the films on which the routine readings do not agree should ensure that each reader becomes aware of any possible deviations of his own standards from those of the team as a whole. The various sets of routine and definitive readings carried out on the first medical surveys at each colliery extended over a long period of time and it was not therefore possible to provide the doctors with up-to-date information about their current performance. On the second surveys, however, the readings at each colliery are normally completed within a period of at most three months, and are analysed and discussed with the doctors before the next survey is undertaken. This close monitoring of standards has already led to an increase in the consistency of the routine readings and has had the effect of reducing periodic fluctuations in the doctors' average standards of classification. Our present experience suggests that some system of this kind must be followed if the best use is to be made of the effort applied by the film readers.

Another major feature of the analysis is the tendency for the magnitude of the reading errors on a batch of films to increase with increasing amounts of abnormality in the batch. When the various categories of pneumoconiosis (other than category 0) are considered separately, however, this trend is not apparent, although there are marked differences in the average error from category to category.

This paper is published by permission of the National Coal Board but the views expressed are not necessarily those of the Board. The work described forms part of the Board's Pneumoconiosis Field Research. My thanks are due to Drs. Stewart Rae and P. J. Chapman (who read the films) and to Dr. J. W. J. Fay for helpful discussions and advice.

REFERENCES

- Ashford, J. R. (1959). A Problem of Subjective Classification in Industrial Medicine. *Applied Statistics*, VIII, 168-185.
Fay, J. W. J. (1957). *Nature*, 180, 309.
—, and Ashford, J. R. (1960). *Brit. J. industr. Med.*, 17, 279.
Fletcher, C. M., and Oldham, P. D. (1949). *Ibid.*, 6, 168.
International Labour Office (1953). Third International Conference of Experts on Pneumoconiosis, Sydney, 1950. Record of Proceedings (publ. Int. Lab. Office, Geneva, 1953.)