DISAGREEMENT BETWEEN OBSERVERS IN AN EPIDEMIOLOGICAL STUDY OF RESPIRATORY DISEASE

BY

R. S. F. SCHILLING, M.D., D.P.H., D.I.H.

J. P. W. HUGHES, M.D., D.P.H.

AND

I. DINGWALL-FORDYCE

(From the Nuffield Department of Occupational Health, University of Manchester)

The investigation of the incidence and severity of a disease in a community may depend on the clinician's skill in taking histories and observing physical signs. The less opportunity there is of using objective measurements, such as *x*-ray films, blood pressures, and blood counts, the more does the clinician have to depend on his skill and on the memory and veracity of the patient.

In a study of byssinosis in Lancashire cotton-workers the investigators had to depend almost entirely on their clinical observations and on the patients' histories, because, unlike most industrial pulmonary diseases, byssinosis shows no specific *x*-ray changes in the lungs. In its later stages the clinical picture is often that of chronic bronchitis with emphysema, which is endemic in Lancashire. The diagnosis is made on an odd but characteristic history of chest tightness and breathlessness on Mondays, gradually extending, as the disease progresses, to other working days (Schilling, 1952).

Cochrane *et al.* (1951) and Fletcher (1952) have shown that clinical observations in respiratory disease may be very unreliable. In this study, therefore, it seemed to be necessary for two investigators to examine each worker independently to see if the method of diagnosing byssinosis was reliable. The two examinations also made it possible to assess the reliability of a number of other clinical observations and of patients' histories.

This paper deals with the disagreement between observers, and shows that there was fair agreement in diagnosing byssinosis on the history, but some of the clinical data were so unreliable that they cannot be used to indicate the nature or the severity of disease.

Procedure

The subjects studied were 187 men aged between 40 and 60 years exposed to fine cotton dust in 27 mills and 88 men (the controls) of similar ages who worked in two factories in which there was no known hazard of industrial pulmonary disease and who had never been exposed to cotton dust. The survey was made between March and December, 1953. Each observer examined about half of the subjects in the first part of the survey and saw the other half in the second part. There was an interval of approximately four months between the two examinations of each worker. Practically all the 275 men were examined at their place of work, but a few were seen at their trade union office. All were seen privately and usually in a reasonably quiet room. Neither observer knew of the other's findings until the field work was completed.

An occupational history and a record of previous illness were taken at the first examination. Both observers asked detailed questions about chest complaints and examined each man for the following signs: clubbing of fingers; cyanosis; movement of chest *en bloc*—in which the chest wall moves up and down as a whole, without showing the normal progressive expansion (Hart and Aslett, 1942); kyphosis—graded as normal, moderate, severe; Hoover's sign—in the early stages of emphysema the subcostal angle remains constant during inspiration; as the disease progresses the subcostal angle is narrowed on inspiration; the observer's thumbs are placed symmetrically along the costal borders as indicators (Hoover, 1913); absence of apical impulse; abnormal chest sounds. Chest expansions and systolic and diastolic blood pressures were also recorded.

Results

Diagnosis of Byssinosis in the Cotton-workers

Both observers graded the men for byssinosis as follows: Normal—no chest symptoms. Grade I—chest tightness and/or breathlessness on Mondays only. Grade II—chest tightness and/or breathlessness on Mondays and other days.

TABLE I.—*Comparison of Byssinosis Grading in* 183* *Cotton-workers*

	Grade	Observer A		
		Normal	I	II
Observer B	Normal ..	**72**	6	—
	I	6	**47**	17
	II	1	14	**20**

Complete agreement in 139 men (76%). Agreement on the presence or absence of byssinosis (139+14+17=170 men (93%).
* Four men could not be compared because their working conditions had changed completely when they were seen on the second occasion.

The observers disagreed in their findings in 24% of the subjects* (see Table I). On the actual presence or absence of byssinosis there was only a 7% disagreement, which was not due to one observer diagnosing byssinosis more readily than the other.

Clinical Signs

For signs recorded as either present or absent the observer error was considerable, and ranged from 30% in absent apical impulse to 4% in cyanosis (Table II). These figures, however, tend to exaggerate the reliability of the observers in detecting those abnormalities which were found infrequently. For example, Observer A found 10 men with cyanosis and B none. Comparing the disagreement on the positive cases, A found significantly more men with kyphosis and cyanosis, and B significantly more men with an absent apical impulse and movement of chest *en bloc*.

TABLE II.—*Observer Error in Clinical Signs in* 275 *Men*

Clinical Sign	Agreement		Disagreement			
	Found by		Found by		Observer Error	Significance
	Neither	Both	A Only	B Only		
Absent apical impulse	106	85	20	64	84 (30%)	B>A P<0·0001
Abnormal chest sounds	188	22	32	33	65 (24%)	—
Movement of chest *en bloc*	205	10	10	50	60 (22%)	B>A P<0·0001
Kyphosis	213	18	35	9	44 (16%)	A>B P<0·0001
Hoover's sign ..	229	6	16	24	40 (15%)	—
Clubbing of fingers	248	6	6	15	21 (8%)	—
Cyanosis	265	0	10	0	10 (4%)	A>B P<0·001

The different findings of each observer for the cotton-workers and controls are compared in Table III. Observer A found similar incidences of abnormalities in both groups, and Observer B found consistently a higher incidence among the cotton-workers. A likely explanation of the different standards of diagnosis of the two observers is that either Observer A was biased against or Observer B was biased

*These percentage disagreements will from now on be referred to as observer error.

TABLE III.—*Physical Signs Observed in Cotton Workers and Controls*

Clinical Sign	Observer A			Observer B		
	Cotton-workers (187)	Controls (88)	Significance	Cotton-workers (187)	Controls (88)	Significance
Absent apical impulse ..	71 (38%)	34 (39%)	—	109 (58%)	40 (45%)	P<0·05
Abnormal chest sounds ..	40 (21%)	14 (16%)	—	46 (25%)	9 (10%)	P<0·001
Movement of chest *en bloc*	14 (7%)	6 (7%)	—	48 (26%)	12 (14%)	P<0·01
Kyphosis ..	35 (19%)	18 (20%)	—	20 (11%)	7 (8%)	—
Hoover's sign	15 (8%)	7 (8%)	—	22 (12%)	8 (9%)	—
Clubbing of fingers ..	8 (4%)	4 (5%)	—	15 (8%)	6 (7%)	—
Cyanosis ..	8 (4%)	2 (2%)	—	—	—	—
Total ..	191	85		260	82	

towards finding abnormalities in the cotton-workers. Both types of bias were probably operating. Kyphosis was assessed independently on lateral *x*-ray films by two other observers. While Observer A graded kyphosis more severely than the *x*-ray observers and tended to be more unfavourable to the controls than to the cotton-workers, Observer B graded less severely than the *x*-ray observers and tended to be more favourable to the controls. This suggests that both observers were biased in assessing kyphosis—A against finding abnormality in the cotton-workers and B against finding it in the controls.

The standards of the observers also appeared to change as the investigation proceeded. Fewer abnormal physical signs were found at the second examinations than at the first; and the observers found an apical impulse less frequently at the second examination than at the first. It seems that they got less perceptive or imaginative as the work went on, probably through fatigue.

Chest Expansion

Chest expansion was measured at the nipple line with a spring-loaded tape measure (Cotterill, 1951). The observers disagreed by ½ in. (1.3 cm.) or more in 42% of all patients. Observer A recorded bigger chest expansions than B—significantly so for the cotton-workers (Table IV).

TABLE IV.—*Means of Chest Expansions*

Observer	Cotton-workers	Controls
A	2·198	2·256
B	2·055	2·174
Difference	0·143	0·082
S.E. of means	0·06	0·10
	P=0·015	

These results suggest that there was a similar bias to that already found—namely, Observer A tended to record high and B low chest expansions in the cotton-workers in comparison with the controls.

Blood Pressures

For blood pressures the observers disagreed by more than 10 mm. of mercury in 43% of the 275 patients for the systolic and 22% for the diastolic pressure.

There was no significant difference between the means of the observers' readings for the cotton-workers or for the controls. On these figures there was no evidence of any obvious observer bias, but blood pressures will be considered in more detail in another paper.

Symptoms*

Except for abnormal chest sounds and cyanosis, which change from time to time, differences in the findings for clinical signs in this study are very likely to be due to the observers alone. For symptoms, however, differences in the observers' records may be due to inconsistencies in the patient's replies to questions, but, as will be seen from the following results, the observer may influence the replies of the patients or possibly interpret the same replies in a different way.

Dyspnoea was graded by using a modified form of the Pneumoconiosis Research Unit's standard questionary.

	Grade
Is your breathing as good as that of normal men of your own age and build on climbing hills or stairs?	1
Are you able to keep up with normal men on the level but unable to keep up on hills and stairs?	2
Are you unable to keep up with normal men on the level but able to walk a mile or more at own speed without stopping?	3
Are you unable to walk about half a mile but able to walk about a hundred yards without stopping?	4

The observers agreed exactly in their gradings in 65% of the cotton-workers and in 87% of the controls (see Table V).

TABLE V.—*Comparison of Observers' Dyspnoea Gradings*

	Dyspnoea Grades	Observer A						
		Cotton-workers				Controls		
		1	2	3	4	1	2	3
Observer B	1	**100**	6			**74**	3	
	2	34	**10**	1		7	**3**	
	3	10	10	**8**		1		**0**
	4		1	2	**1**			
		Agreement in 119 cases (65%)				Agreement in 77 cases (87%)		

A graded more severely than B by 7 grades
B " " " " A " 68 "

A graded more severely than B by 3 grades
B graded more severely than A by 9 grades

Observer B graded more severely than Observer A, probably because A started at the top and B at the bottom of the questionary. However, the difference in severity of grading was more marked for the cotton-workers than for the controls, and there is again evidence from the figures in Table V that either A was so biased against, and/or B towards, finding abnormality in the cotton-workers that they actually influenced the men in their replies.

A similar influence is apparent in the men's replies to the questions about the effect of winter and fog on their breathing (Table VI).

TABLE VI

Breathing Worse in	Observer A		Observer B	
	Cotton-workers	Controls	Cotton-workers	Controls
Winter	23 (13%)	10 (12%)	45 (24%)	10 (12%)
Fog	81 (44%)	20 (23%)	95 (52%)	18 (21%)

Progress of Disease

The cotton-workers were asked what was hoped to be a non-leading question about the progress of their symptoms —" How does your breathing compare with what it was a year ago? " The answers were recorded as better, same, worse, and unknown. The results in the 98 men considered by both observers to have byssinosis show that different answers were given by 27 of the men, and that Observer B recorded more men than A as being worse (see Table VII).

TABLE VII.—*Differences in the Recorded Progress of 98 Men With Byssinosis*

		Observer A			
	Progress	Better	Same	Worse
Observer B	Better	1	1	1
	Same	1	**40**	5
	Worse	1	18	**30**

*Four cotton-workers have been omitted from the following comparisons because they had changed their work between the examinations.

Similarly in the cotton-workers' replies to questions about the presence or absence of cough in summer and winter, Observer B recorded significantly more men as complaining of cough in the winter. The observer error in the answers was 26% in respect of summer cough and 30% for winter cough. Observer B also recorded significantly more men as having sputum.

The cotton-workers were also asked how many years they had complained of their symptoms of chest tightness on Mondays. From their answers and their occupational histories it was possible to calculate the length of exposure to dust before the onset of symptoms (induction times). If reliable it was hoped to use these induction times as an indication of the severity of the risk in particular mills. Of the 98 men who were considered by both observers to have byssinosis, 58 gave answers that agreed to within five years, 19 within 5 to 10 years, and 18 within 10 to 20 years. In three men there was actually a discrepancy of more than 20 years. Possibly some of the men did not understand the question or their answers were wrongly recorded.

The discrepancies in the patient's description of their symptoms did not occur mainly in a small group of men who gave unreliable answers to all types of questions. Inconsistency in replies was distributed among many patients.

Symptoms may have changed in the interval between the examinations and particularly may have been influenced by the weather. Although the field work extended from early spring to mid-winter, the weather was mostly mild throughout and there was no obvious tendency for men to deteriorate between the first and second examinations.

Discussion

Although both observers were aware of the danger, bias occurred both in eliciting physical signs and in taking histories. Observer A consistently found less evidence than Observer B of abnormality among the cotton-workers. In therapeutic trials such bias can be prevented by keeping the clinician uninformed of the treatment given. In theory it could have been avoided in this study if the examiners had not known the occupations of the men; but this was not practicable, since the cotton-workers and controls were examined at their own factories; nor would it have been possible under any conditions, as in this type of field study most patients disclose to the doctor the nature of their work. The observer bias was not obvious or marked in the blood pressures and in the chest expansions, where actual measurements could be made. There was also no bias and the error was low in the diagnosis of byssinosis, where much care was taken to ensure that detailed and accurate histories of symptoms of dust exposure were taken. If the same care had been taken in defining and eliciting some of the physical signs the observer error might well have been reduced. Kyphosis was graded severe, moderate, or normal, but as no standards were laid down for moderate and severe kyphosis the grading did not give much help. Grading and a clear definition of standards are most likely to reduce observer error, as they have done in reading *x*-ray films of pneumoconiosis (Fletcher and Oldham, 1951).

The observer error in recording physical signs could be explained by the fact that the observers were not experienced chest physicians. It is possible to compare their mean observer error for adventitious chest sounds, movement of the chest *en bloc*, kyphosis, and clubbing of fingers, with the mean observer error of four pairs of chest physicians (all of whom were Members or Fellows of the Royal College of Physicians) in eliciting the same physical signs in Fletcher's (1952) investigation. While, because of the different conditions of the two studies, comparison should not be drawn too closely, it does show that the errors of the observers in the present study and of the chest physicians in Fletcher's study were of the same order, and that the more experienced Fellows, who were paired together, did not, on the whole, do any better than their less experienced colleagues (see Table VIII). Pyke (1954) found that finger-clubbing was often an unreliable sign and that reliability was apparently unaffected by the clinical experience of the observer.

TABLE VIII.—*Mean Observer Error in Adventitious Chest Sounds, Movement of Chest En Bloc, Kyphosis, and Clubbing of Fingers*

Four pairs of chest physicians* (examined 20 men)	14% M.R.C.P.s 27% ,, 29% F.R.C.P.s 35% M.R.C.P.s
Observers A and B (examined 275 men)	17%

* The eight physicians were described as A, B, C, D, E, F, G, and H. A and E, who were F.R.C.P.s, were paired together, and the remainder were paired in alphabetical sequence—i.e., BC, DF, and GH. By selecting the pairs the error could have been made much more or less.

The observers who had previous experience of surveys tried to design questions to get reliable answers. But, clearly, on questions such as those about cough, the effects of fog on breathing, and the progress of the disease, the patients could, when in doubt, be influenced by the questioner or perhaps the questioners interpreted equivocal replies differently. Cochrane and his colleagues (1951) found similar evidence of doctors getting different replies to the same questions. In many instances the answers to such questions are unreliable. The error probably could be reduced if observers noted equivocal answers, and made no attempt to get more definite results.

Surveys should also be kept within reasonable limits to avoid changes in standards of observation from fatigue or boredom.

Observer error is of considerable importance in epidemiological studies where patients are seen once and many of them are normal or have only early manifestations of disease. Much depends on the reliability of single observations of signs and symptoms which are sometimes indefinite. If there had been only one observer in this study, unwarranted claims might have been made about the signs and symptoms of byssinosis. Observer B would certainly have reported a higher incidence of abnormalities than Observer A. In hospital and general practice, however, it would be wrong to conclude that observer error has the same significance, since diagnosis and treatment seldom depend on single observations (*Lancet*, 1954). A good physician, by his experience, is able to assess the reliability and significance of signs and symptoms, and he may unconsciously discount much that is unreliable. But it would also be wrong to suggest that observer error is of no importance in ordinary clinical work, as the extent to which it affects clinical judgment, particularly in the less experienced doctors, is not known. Studies such as the one described in this paper help to overcome the unwillingness of doctors to appreciate their own unreliability (Newell *et al.*, 1954) and may encourage higher standards of clinical observation.

Summary

Two series of persons—187 male cotton-workers and 88 men not exposed to cotton dust—were examined independently by two observers to assess the reliability of diagnosing byssinosis, eliciting certain clinical signs of respiratory disease, measuring chest expansions, recording blood pressures, and taking histories of symptoms such as cough and breathlessness.

There was disagreement in diagnosing byssinosis in 24% of the cotton-workers. For clinical signs there was a significant disagreement between the observers in their records of absent apical impulse, movement of the chest *en bloc*, kyphosis, and cyanosis.

In measuring chest expansions the observers disagreed by ½ in. (1.3 cm.) or more in 42% of the patients.

For blood pressures the observers disagreed by more than 10 mm. of mercury in 43% of the patients for systolic and 22% for diastolic pressures.

One observer appeared to be biased against and/or the other towards finding abnormalities in the cotton-workers. The same observer bias occurred in taking histories of symptoms.

The importance of observer error in epidemiological studies and possible methods of reducing it are briefly discussed.

We are grateful to the Medical Research Council for a grant to study byssinosis. We have had much helpful advice and criticism from our colleagues inside the Department and others, particularly from Dr. J. C. Gilson and his colleagues in the Pneumoconiosis Research Unit, and from Dr. C. M. Fletcher. The lateral *x*-ray films were read for kyphosis by Dr. J. C. Gilson and Dr. W. E. Miall. To all the above we gratefully acknowledge our indebtedness and to the many workers and employers whose co-operation has made this work possible.

REFERENCES

Cochrane, A. L., Chapman, P. J., and Oldham, P. D. (1951). *Lancet*, 1, 1007.
Cotterill, M. S. (1951). *Physiotherapy*, 37, 49.
Fletcher, C. M. (1952). *Proc. roy. Soc. Med.*, 45, 577.
—— and Oldham, P. D. (1951). *Brit. J. industr. Med.*, 8, 138.
Hart, P. D'A., and Aslett, E. A. (1942). *Spec. Rep. Ser. med. Res. Coun. Lond.*, No. 243.
Hoover, C. F. (1913). *Arch. intern. Med.*, 12, 214.
Lancet, 1954, 1, 87.
Newell, R. R., Chamberlain, W. E., and Rigler, Leo (1954). *Amer. Rev. Tuberc.*, 69, 566.
Pyke, D. A. (1954). *Lancet*, 2, 352.
Schilling, R. S. F. (1952). *Proc. roy. Soc. Med.*, 45, 601.

CLINICAL EXPERIENCE OF THE INSULIN ZINC SUSPENSIONS*

BY

J. M. STOWERS, M.D., M.R.C.P.

Department of Medicine, University of St. Andrews, Dundee

AND

J. D. N. NABARRO, M.D., M.R.C.P.

Assistant Physician, The Middlesex Hospital, London, formerly of the Medical Unit, University College Hospital Medical School, London

In introducing the insulin zinc suspensions, Hallas-Møller *et al.* (1952) claimed that they enabled the majority of diabetics requiring insulin to be controlled with a single daily injection. By mixing in different proportions the relatively quick-acting amorphous and the longer-acting crystalline forms, insulins with a wide range of prolonged action could be obtained. These mixtures were consistent in their effect, unlike, for example, those of protamine zinc (P.Z.I.) and soluble insulin (S.I.). The Danish workers found that a mixture of seven parts crystalline with three parts amorphous was suitable for most patients; this was placed on the market as "lente insulin" and is known in this country as insulin zinc suspension (I.Z.S.). The amorphous and crystalline forms are also available separately as insulin zinc suspension (amorphous)—I.Z.S.(A)—and insulin zinc suspension (crystalline)—I.Z.S.(C). It was claimed, too, that the absence of any foreign protein and the use of thrice-recrystallized insulin in the preparation of the insulin zinc suspensions reduced the incidence of sensitivity reactions. Preliminary studies in this country (Lawrence and Oakley, 1953; Oakley, 1953; Murray and Wilson, 1953; Nabarro and Stowers, 1953a, 1953b) confirmed these findings, and the new insulins were made generally available in Great Britain in October, 1953. We have now used them in the routine management of 240 diabetic patients in the clinics at University College Hospital and in the Dundee area, and this report is a critical review of the results we have obtained.

*The results in this paper formed the basis of a communication given at the Annual Scientific Meeting of the British Diabetic Association on July 16, 1954.

Results

Patients already on insulin were selected for trial of the new preparations only if their previous regime was proving unsatisfactory. They therefore do not represent a cross-section of all diabetics requiring insulin, but rather some of the more difficult ones. A series of 205 patients already taking insulin have been tried on the insulin zinc suspensions, and the indications for transfer and the results obtained are shown in Table I. The term "poor control—hyperglycaemia" is used for patients with diabetic symptoms (thirst, polyuria, or unintentional loss of weight) or in whom the blood sugar was found to be over 250 mg. per 100 ml. on an ordinary day. Of the patients transferred because their previous regime involved multiple injections, 81% were satisfactorily controlled with a single injection of the new preparations. About 80% of those who had hyperglycaemia, diabetic symptoms, or repeated hypoglycaemic attacks were improved. Of those in whom the results of transfer have so far proved unsatisfactory, the trial was discontinued in 10 for reasons that are given in a subsequent paragraph. In the remainder further adjustments of the proportions of I.Z.S.(A.) and I.Z.S.(C.), as well as alterations of the diet, are being tried to improve the control of the diabetes.

TABLE I.—*Indications for Transfer* (205 *Cases*)

	No.	Trial Discontinued	Satisfactory Result
Multiple injections	120	5	98 (81%)
Poor control—hyperglycaemia ..	93	4	70 (76%)
,, ,, hypoglycaemia ..	47	—	39 (83%)
Allergy	9	1	6

The sum of the number of indications exceeds the total number of patients because in a proportion multiple indications were present.

The alterations in dose that were required when the patients were transferred from different insulin regimes to the insulin zinc suspensions are shown in Table II. Patients who had been taking two or three doses of S.I. usually needed little if any change in dosage, whereas those previously on a single dose of P.Z.I. required on an average 40% more. The difference between these two figures is highly significant, being five times the standard error. Patients previously using globin insulin (G.I.) or P.Z.I. plus S.I. required usually a slight increase in dose, but the differences do not attain the level of statistical significance. The alterations of insulin requirements that follow transfer from P.Z.I. or S.I. to the insulin zinc suspensions can to some extent be explained by the pharmacological properties of these insulins. P.Z.I., if given in a dose large enough to control postprandial hyperglycaemia in the day, is virtually certain to produce hypoglycaemia in the night or early morning. Therefore patients on P.Z.I. alone usually have to take a

TABLE II.—*Alterations of Dose in Transfer to I.Z.S.*

	No.	Mean Previous Dose	Mean Alteration	% Change
2 or 3 doses S.I. ..	19	47 (20–116)	+1·7 (−24 to +26)	+3·6
2 doses G.I. ..	11	53 (44–88)	+12 (−12 to +24)	+23
1 dose G.I. ..	19	40·5 (20–124)	+5·7 (−16 to +36)	+14
1 ,, P.Z.I. ..	36	32·2 (10–60)	+12·4 (0 to +40)	+38·5
P.Z.I.+S.I. a.m.	82	53·9 (10–104)	+5·9 (−28 to +58)	+10·9
P.Z.I.+S.I. ,, and S.I. p.m.	21	78·5 (36–128)	+10·2 (−16 to +78)	+12·9
Miscellaneous ..	7			