

## Simple method of estimating severity of pulmonary fibrosis on a numerical scale

T ASHCROFT,\* JUDY M SIMPSON,† V TIMBRELL‡ *From the \*Department of Pathology, Freeman Hospital, Newcastle upon Tyne, †Department of Medical Statistics, University of Newcastle upon Tyne, and the ‡MRC Epidemiology Unit, Cardiff, South Glamorgan, Wales*

**SUMMARY** A continuous numerical scale for determining the degree of fibrosis in lung specimens was devised for correlation with other pulmonary variables such as lung function tests or mineral burden. Grading was scored on a scale from 0 to 8, using the average of microscope field scores. The system allows fibrosis to be measured in small samples of tissue (1 cm) which can provide a detailed description of the changes in a lung, currently not possible with most existing methods.

Quantitative assessment of interstitial pulmonary fibrosis is usually carried out in simple broad categories. Beattie and Knox used a system of grading asbestosis into minimal, moderate, and severe categories but their criteria were related specifically to asbestosis and not generally applicable.<sup>1</sup> The UICC Working Group on Asbestos and Cancer recommended a system of grading asbestosis into minimal, slight, moderate and severe categories which took account of both the extent and severity of pulmonary fibrosis.<sup>2</sup> This system was extended and rationalised by Hinson *et al*<sup>3</sup> so that it could be applied to cases of pulmonary fibrosis other than asbestosis, but it still consisted simply of four broad categories for severity of fibrosis. More recently, a further extension of the protocol of Hinson *et al* has been proposed by Craighead *et al*.<sup>4</sup> For purposes of correlation with other pulmonary variables, such as lung function tests or mineral burden, a continuous numerical scale of fibrosis in lung samples would be highly desirable, and we devised the following scheme.

### Material and methods

A paraffin section of lung, stained either by haematoxylin and eosin or by a trichrome method, was systematically scanned in a microscope using a × 10 objective. Each successive field was individually assessed for severity of interstitial fibrosis and allotted a score between 0 and 8 using a predetermined scale of severity (table 1). After examining the whole section the mean score of all the fields was taken as the fibrosis score for the section and was expressed correct to two decimal places.

In every field the predominant degree of fibrosis was recorded—that occupying more than half of the field area. The whole area of the circular microscope image was considered and the observer first decided whether the parenchyma in the field was normal or fibrotic. If normal tissue predominated the field was allotted a score of 0. If fibrotic tissue predominated the observer decided on the predominant level of fibrosis in the field, trying to place the field in one of the odd numbered categories (table 1). If there was any difficulty in deciding between two odd numbered categories the field would be given the intervening even numbered score. Fields predominantly occupied by portions of large bronchi or vessels, or by malignant tumour deposits, were not counted, and fields partly occupied by such structures were assessed on the predominant fibrosis in the remaining field area. Fibrosis alone was considered when arriving at a score and other changes which might be present, such as emphysema, were ignored. Cellular inflammatory

Table 1 *Criteria for grading lung fibrosis*

<i>Grade of fibrosis</i>	<i>Histological features</i>
0	Normal lung
1	Minimal fibrous thickening of alveolar or bronchiolar walls
2	
3	Moderate thickening of walls without obvious damage to lung architecture
4	
5	Increased fibrosis with definite damage to lung structure and formation of fibrous bands or small fibrous masses
6	
7	Severe distortion of structure and large fibrous areas; "honeycomb lung" is placed in this category
8	Total fibrous obliteration of the field

exudate in airspaces was ignored but organised exudate was treated as fibrosis. The criteria used in determining scores are outlined in table 1. Representative photomicrographs of the grades are shown in the figure. With practice, a typical lung section of about  $2 \times 1$  cm could be read in two to three minutes, yielding a total of 40 to 50 fields.

#### INTRA-OBSERVER VARIABILITY OF FIBROSIS SCORES

To assess the repeatability of fibrosis scores one of us (TA) took 20 slides of necropsy lung tissue from workers exposed to asbestos and scored each for fibrosis on five separate occasions. The slides were selected to include sections with high, medium, and low fibrosis scores.

As the variability between the five replicates increased with the mean, the square root of each measurement was taken, enabling an average repeatability to be obtained, regardless of the size of the fibrosis score. Variability from the two sources, between slides  $s_b^2$  and within slides—that is, between replicates  $s_w^2$ —were estimated by one way analysis of variance. The estimates were:  $s_b^2 = 0.33$ ;  $s_w^2 = 0.01$ ; mean = 1.18, giving a very small within slide standard deviation,  $s_w = 0.10$  (coefficient of variation 8.5%), which is a measure of repeatability.

#### VARIABILITY IN SCORES FOR INDIVIDUAL FIELDS

To test inter- and intraobserver variation in scores for individual fields, 100 fields from several sections of lung tissue were photographed and numbered. The fields were chosen at random by moving the stage controls of the microscope blindly and photographing the resulting field without further stage adjustment. Fields showing predominantly carcinoma, large bronchi, or vessels were omitted except for two examples, giving 98 scorable fields. Sections were chosen to ensure that all grades of fibrosis were represented.

Five observers, including one of us (TA), scored each photograph on two occasions. The observers were all experienced histopathologists and were provided with a typed copy of the method of scoring and a set of standard photomicrographs of the different grades, but received no other tuition. All analyses were done on the untransformed scores which ranged from 0 to 8.

#### Results

##### INTRA-OBSERVER VARIABILITY

Separate one way analyses of variance for each observer gave estimates of the variability between and within photographs (table 2). The within photo (between duplicates) standard deviation shows the

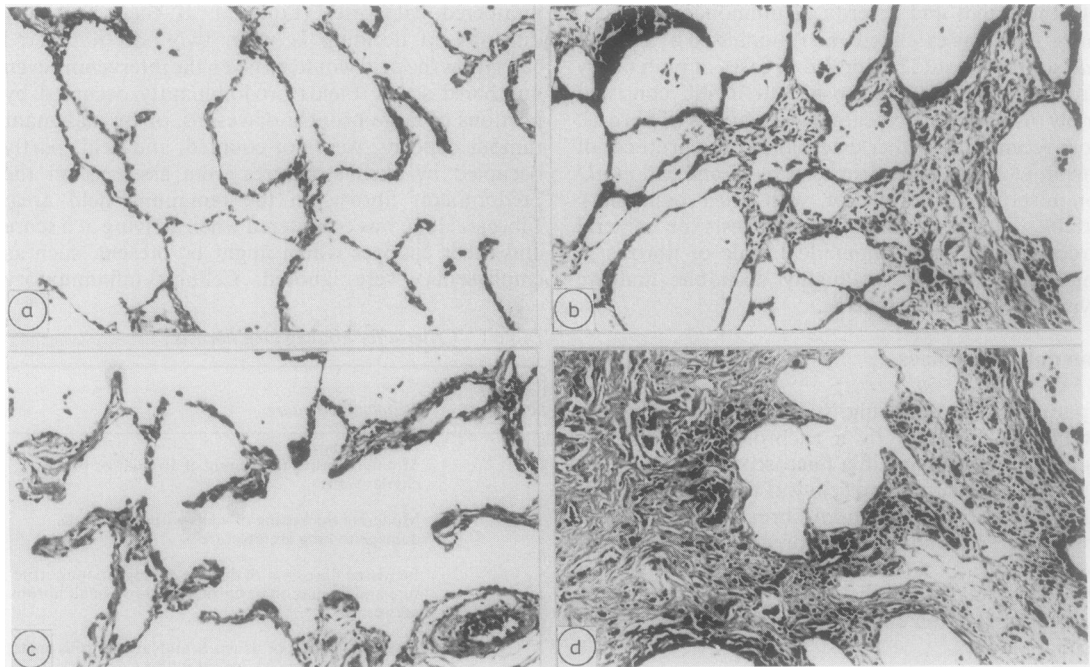


Figure (a) Grade 1 fibrosis; (b) grade 5; (c) grade 3; (d) grade 7. (Haematoxylin and eosin.)

Table 2 Intraobserver variability for scores for individual photographs

	Observer				
	1	2	3	4	5
Variance between photos $s_b^2$	4.60	4.55	6.50	3.53	4.56
Variance within photos $s_w^2$ (between duplicates)	0.52	0.78	0.61	0.56	1.28
Within photo standard deviation $s_w$	0.72	0.88	0.78	0.74	1.13
Mean score	2.56	2.94	3.04	3.82	2.93
First score mean	2.65	2.67	3.01	3.56	3.31
Second score mean	2.47	3.20	3.06	4.07	2.56
Difference	-0.18	0.53*	0.05	0.51*	-0.74*

\* $p < 0.001$  for paired  $t$  test, difference = 0.

repeatability of the scoring system for each of the observers. Table 2 also gives the mean score for each observer on each occasion, and the difference between them, which was significantly different from zero for three of the five observers.

#### INTEROBSERVER VARIABILITY

One way analysis of variance for each set of scores (first and second) separately gave estimates of the variance components, which were similar for the two sets, and so were averaged. The variability between photographs was 4.27 and between observers (within photos) was 1.41, with a mean score of 3.06. The within photo standard deviation of 1.19 measures the reproducibility of the scoring system when used by different observers. There was found to be a highly significant difference between the mean scores for the five observers ( $F_{4,388} = 26.3$   $p < 0.0001$ ).

#### Discussion

The repeat scoring of sections by a single observer shows that the method is capable of good repeatability in the hands of an experienced reader, with a coefficient of variation under 10%. When scores for individual photographs were compared, greater variability was observed—to be expected when dealing with individual scores rather than means. When averaging over  $n$  observations, the standard deviation is reduced by a factor of  $\sqrt{n}$ . Typically, the fibrosis score for a section is obtained from the mean of 20 to 50 fields. Intraobserver variability for individual scores was consistent with that for fibrosis scores when this is taken into account, except for observer 5. Only observers 1 and 3 showed no significant difference between the means of all 98 scores on the two occasions.

There was appreciable interobserver variation, with a highly significant difference between the mean scores and a variability about two to three times greater than

intraobserver variability. Observer 3 (TA) was experienced in the method but the other four people had not used the scoring procedure before these tests. Further experience and discussion of difficult fields would probably have improved their performance.

To measure reliability (repeatability and reproducibility) we just used the residual (within photo) standard deviation. Another measurement which is often used is the intraclass correlation coefficient  $s_b^2/(s_b^2 + s_w^2)$ . This is only appropriate, however, when observations are made on a random sample from the population to which the measurement will be applied.<sup>5</sup> This was not the case in this study as the photos were chosen to give a good spread across the scale from 0 to 8. Consequently, the variability between photo  $s_b^2$  was larger than it would be with a random sample and the intraclass correlation coefficient is inflated. For example, it gave values of repeatability ranging from 0.78 to 0.91 for the five observers and reproducibility of 0.75 between all observers. Previous schemes for estimation of lung fibrosis have primarily been devised for assessment of asbestosis. Even where it is claimed that the scheme is applicable to fibrosis of other aetiology,<sup>3,4</sup> the histological criteria used are closely based on the progression of the lesions of asbestosis and are not necessarily applicable to other fibrotic diseases. In contrast, the present method recognises basic histological changes in individual microscope fields, which are common to a wide variety of fibrosing pulmonary conditions, and it is therefore as appropriate in non-occupational diseases such as idiopathic fibrosing alveolitis as it is in asbestosis.

The main advantage of our method is that it permits fine gradation of fibrosis. The averaging of field scores gives a continuous numerical scale from 0 to 8 which can be correlated mathematically with other measurements expressed on a continuous scale—for example, mineral dust concentrations or lung function tests. Previous schemes have provided for the assessment of the extent of fibrosis in a lung, when whole lungs are available for study, but again only in very broad categories. The present method allows fibrosis to be measured in quite a small sample of tissue (1 cm) and is thus applicable to biopsy specimens. Great caution should be used in extrapolating the score in a biopsy specimen to the whole lung because of the small size of the sample and the wide variation possible in severity of fibrosis within the same lung. In the case of necropsy lungs, however, it would be quite feasible to take a large number of samples systematically from a lung and map the extent and severity of fibrosis for correlation with, say, results of chest radiographs or isotope scans. Such a detailed description of the changes in a lung is not possible with most existing methods.

Our thanks are due to our colleagues in the pathology department at Freeman Hospital for help in evaluating the method and to Miss Carol Skingsley for typing the manuscript.

#### References

- 1 Beattie J, Knox JF. Studies of mineral content and particle size distribution in the lungs of asbestos textile workers. In: Davies CN, ed. *Inhaled particles and vapours*. London: Pergamon Press, 1961:419-33.
- 2 Report and recommendations of the UICC working group on asbestos and cancer. *Ann NY Acad Sci* 1965;132:715-6.
- 3 Hinson KFW, Otto H, Webster I, Rossiter CE. Criteria for the diagnosis and grading of asbestosis. In Bogovski P, ed. *Biological effects of asbestos*. Lyon, France: World Health Organisation, 1973:54-7.
- 4 Craighead JE, Abraham L, Churg A, et al. Asbestos associated diseases. *Arch Pathol Lab Med* 1982;106:592-3.
- 5 Winer BJ. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971:286.

Requests for reprints to: Dr T Ashcroft, Consultant Pathologist, Freeman Hospital, Freeman Road, High Heaton, Newcastle upon Tyne NB7 7DN, England.