J. clin. Path. (1968), **21**, 67

Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri

J. COCKER, H. FOX, AND F. A. LANGLEY

From the Department of Pathology, St. Mary's Hospitals, Manchester, and the Departments of Pathology and Obstetrics and Gynaecology, Manchester University

SYNOPSIS A group of pathologists, all working in the same laboratory and all applying the same diagnostic criteria to the diagnosis of epithelial abnormalities in the uterine cervix, have studied the consistency with which they have applied these criteria. Epithelial abnormalities were ranked, and a series of sections were diagnosed separately by each pathologist at various times over a number of years. Both consistency and trend were studied by a graphed statistical method and it was shown that not only were there serious inconsistencies in diagnosis between the various pathologists but also between the diagnoses made by individual pathologists studying the same section at various times.

It is suggested that this inconsistency in the application of agreed diagnostic criteria is of importance when considering discrepancies between reported series of cervical epithelial abnormalities and that the type of study described is of value in assessing both variations in diagnostic criteria between different laboratories and the consistency of pathologists in training. Any slight change in the application of diagnostic criteria for any individual pathologist with the passage of time may also be detected by this technique.

Ashley (1966) has recently stated that any competent pathologist can accurately diagnose carcinoma *in situ* of the cervix. Such a proposition implies, first, that all pathologists are agreed on the histological criteria on which a diagnosis of carcinoma *in situ* is made and, second, that pathologists apply these criteria consistently. It is the second of these assumptions which is examined here.

MATERIAL AND METHODS

The first part of this investigation was carried out some four years ago when an independent member of the staff selected 28 cases of cervical epithelial abnormality from the laboratory files, the diagnosis varying from invasive squamous cell carcinoma to squamous metaplasia and reserve cell hyperplasia. Half these cases had been diagnosed originally as carcinoma *in situ* and most of the others as dysplasia or as border-line between dysplasia and carcinoma *in situ*. This constitutes series 1. Some of the specimens were from simple cervical biopsies and others from full-cone biopsies with 12 to 14 blocks, some cut serially. These were examined by three pathologists A, B, and C, who were asked to grade them according to the most serious lesion found in each specimen, using the diagnostic code of Table I. The three histopathologists had special experience in gynaecological pathology, they had worked together in the same laboratory for at least two years, and it was supposed that they used the same diagnostic criteria. For the most part the material had been seen previously by one or other of the three but, for the purpose of this study, the previous diagnoses and clinical information were not available at this time. The inconsistencies in diagnosis were surprising (Table II), hence a simpler test based on a second series of slides was made two years later. For this (series 2) 30 cases were selected by another independent member of the staff but only one section from each case was made available for examination. The examination was made by the two pathologists A and C, since B was now working elsewhere. This series differed somewhat from the first in that

TABLE I

DIAGNOSTIC CODE

Code	*Diagnosis*
1	Invasive squamous carcinoma
2	Borderline invasion
3	Carcinoma *in situ*
4	Borderline carcinoma *in situ*
5	Dysplasia
6	Epithelial changes not amounting to dysplasia *e.g.*, reserve cell hyperplasia, metaplasia

Received for publication 16 May 1967.

the cases were more evenly divided between the diagnostic categories. To test the effect of time on diagnostic consistency these were examined a second time by the same pathologists some 18 months later. In the interval, photographs of a variety of cervical lesions had been prepared by A, and by agreement, these were available as a standard of reference.

It must not be supposed that we regard the diagnostic categories listed in Table I as a sequence of changes which take place in the development of carcinoma of the cervix but rather that 1 is a more serious lesion than 2, and 2 more serious than 3 and so on. Analytically we treated these numbers as ranks, as in rank correlation. The histological definitions are essentially those proposed by the Committee for Histological Definitions (1962). Since more than one lesion may be found in a section the diagnosis used was always that of the most serious lesion.

For the purpose of comparison two parameters are needed, one of consistency and one of trend. Fletcher and Oldham (1949, 1951) were faced with a similar problem when comparing the radiographic assessment of pneumoconiosis by various workers. They used five diagnostic categories corresponding to increasing severity of the disease. For analysis they treated the categories as continuous rather than discrete variables. They calculated two indices, one of inconsistency and another of disagreement. We have also treated our categories as continuous variables but found it necessary to modify the indices of Fletcher and Oldham to make them suitable for statistical analysis. Comparisons may be made by means of correlation diagrams. In Fig. 1a the diagnoses of O and A are compared. Each column, moving from left to right, represents the diagnostic category originally given to the case by O. Similarly each row represents the category assigned to the case by observer A. Thus case 1 was given a place by O in the column headed 3 and by A in the row 3. The case was therefore scored in the square where the row and column intersect. The numbers in the squares correspond to the numbers of cases scored in this way. If there were complete agreement between the diagnoses of the two observers all cases would be scored along the diagonal. However, there was not complete agreement and there is therefore scatter about the diagonal. This scatter represents the inconsistency between the two sets of observations. This is also illustrated in Figure 1b. There were five cases in which O and A agreed this is represented by the column zero; 12 cases in which A was one category less severe than O this is represented by column −1; six cases in which A was more severe than O by one category, it is represented by column +1 and so on. It can be seen that a typical cocked-hat histogram is formed which is shifted to the left of zero. This shift indicates the trend of A's diagnoses compared with those of O, and is measured by the mean of the histogram. A narrow histogram indicates that A fairly consistently differed from O by an amount equal to the mean; a wider histogram indicates that A was not so consistent. The standard deviation is a measure of the width of the histogram and thus a measure of consistency; the smaller the standard deviation the greater the consistency and the greater the standard deviation the less the consistency. Thus we may measure trend by the mean of the differences

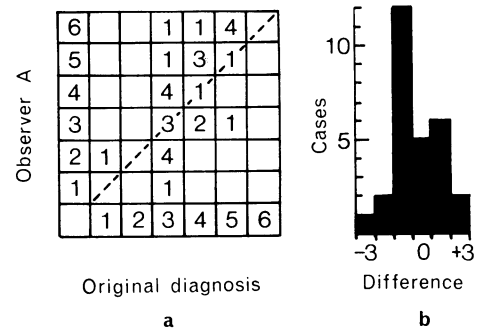

FIG. 1a. *The numbers along the bottom are the diagnostic codes of the original diagnosis, the numbers up the left hand side are the diagnostic codes assigned by observer A. b. The differences along the bottom indicate the difference between the diagnostic code originally assigned and that given by observer A.*

of categories assigned to each case by the two observers and consistency by the standard deviation of these differences or, more simply, by their variances. The significance of the trend may then be determined using the t test (Mather, 1943).

RESULTS

The results of series 1 are tabulated in Table II and analysed together with the results of series 2 in

TABLE II

COMPARISON OF THE DIAGNOSES MADE BY THREE OBSERVERS, A, B, AND C, WITH THE ORIGINAL DIAGNOSES

Case No.	*Diagnosis Original (O)*	*A*	*B*	*C*
1	3	3	5	4
2	4	3	6	1
3	3	4	5	3
4	3	2	5	3
5	3	3	5	5
6	3	4	4	4
7	5	3	5	3
8	4	4	4	5
9	5	6	6	6
10	3	2	5	3
11	4	6	4	5
12	3	2	4	5
13	4	5	5	5
14	5	6	5	5
15	3	1	4	4
16	3	4	5	6
17	3	5	4	3
18	4	5	5	5
19	4	5	5	5
20	4	3	3	3
21	3	4	5	4
22	3	2	1	2
23	5	6	6	5
24	5	6	5	5
25	1	2	2	3
26	5	5	5	5
27	3	6	5	3
28	3	3	4	1

TABLE III

STATISTICAL ANALYSIS OF THE DIAGNOSTIC COMPARISONS

Series	*Comparison*	*Difference of Means*	*Variance*	*t*	P
1	O, A	−0·322	1·48	1·36	0·2>1·1
1	O, B	−0·930	1·03	4·82	<0·001
1	O, C	−0·360	1·72	1·45	0·2>0·1
1	A, B	−0·608	1·95	2·31	0·05>0·025
1	B, C	+0·572	2·03	2·19	0·05>0·025
1	C, A	+0·036	2·08	0·13	0·9>0·8
2	A_1/C_1	−0·167	1·21	0·84	0·5>0·4
2	A_2/C_2	+0·233	1·15	1·19	0·3>0·2
2	A_1/A_2	−0·267	0·34	2·51	0·02>0·01

The subscripts 1 and 2 in series 2 indicate the first and second set of observations by A and C, separated by a year and a half.

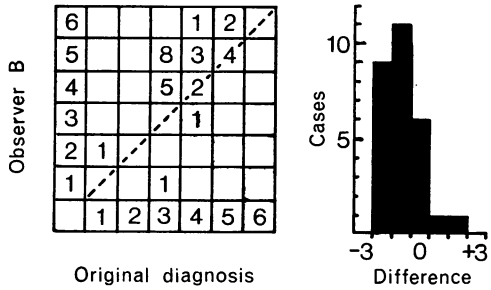

FIG. 2*a*. *Correlation diagram comparing diagnoses assigned by observer B with those originally made b. Histogram making a similar comparison.*

Table III. It is readily seen that each of the observers, by and large, gave the lesions a less serious diagnostic category than O. This is indicated by the negative sign of the mean difference. This is as might be expected since in the test no clinical information was available and there was no bias to regard the lesion more seriously than was warranted objectively. B was less severe in his diagnosis than O, A, or C, and the difference was statistically significant (Fig. 2 and Table III). However, B was more consistent in his diagnoses than A or C; this is indicated by the relatively low variance in the O, B comparison and by the slightly lower variance in the comparisons A, B and B, C than in C, A. There is close overall agreement between C and A but some fairly wide scatter (inconsistency) in their diagnoses (Fig. 3). These results suggest that whereas A and C interpreted the diagnostic criteria in a very similar manner B interpreted them somewhat differently but fairly consistently.

In the second series the differences between A and C are quite small. On the first occasion C was less severe than A and on the second he was more severe and on both occasions the difference was greater

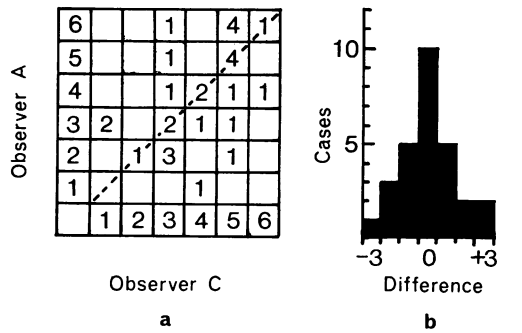

FIG. 3*a*. *Correlation diagram comparing diagnoses assigned by observer A with those of observer C (series* 1). *b. Histogram making a similar comparison.*

TABLE IV

COMPARISON OF THE DIAGNOSES OF A ON TWO OCCASIONS IN THE SECOND TRIAL SEPARATED BY AN INTERVAL OF A YEAR AND A HALF

Case No.	*First Diagnosis* (A_1)	*Second Diagnosis* (A_2)	*Difference*
1	5	5	0
2	5	5	0
3	6	6	0
4	4	5	−1
5	3	3	0
6	4	3	+1
7	3	4	−1
8	1	1	0
9	6	6	0
10	3	4	−1
11	1	1	0
12	3	5	−2
13	3	3	0
14	5	5	0
15	6	6	0
16	1	1	0
17	3	3	0
18	3	3	0
19	6	6	0
20	6	6	0
21	5	5	0
22	6	6	0
23	1	2	−1
24	5	5	0
25	3	3	0
26	3	4	−1
27	6	6	0
28	3	4	−1
29	3	4	−1
30	6	6	0

Mean of differences = −0·267
Variance = 0·340
Standard deviation = 0·1063
t = 0·267 ÷ 0·1063 = 2·51
Therefore P lies between 0·02 and 0·01.

than in series 1, but still not statistically significant. The most interesting observation is the comparison between A on two occasions separated by a year and a half (Table IV). Of the 30 cases, the same diagnosis was made on both occasions 21 times. On the second examination the diagnosis was one category less severe in seven cases and two categories less severe in one case; once only was it more severe. These

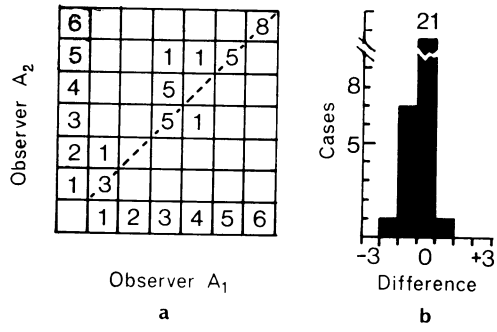

FIG. 4*a*. *Correlation between the diagnoses of observer A on two occasions. b. Histogram making similar comparison.*

differences are illustrated in the histogram (Fig. 4). Because the changes are mostly in one direction the difference in the means of the two series of observations (the trend) is relatively large but the variance of the difference, as shown by the width of the histogram, is small. Hence it is large and thus the difference between the two series of observations is significant. This suggests that A's diagnostic criteria had shifted in the year and a half. This might be accounted for by the fact that A had been collaborating with another group of pathologists during this period in defining more precisely the diagnostic criteria of epithelial abnormalities of the cervix.

DISCUSSION

Both Siegler (1956) and Kirkland (1963) sent sections of cervical lesions to competent histopathologists and found considerable and disturbing disagreement in their diagnoses. This present investigation shows that even in one laboratory, in which it is supposed that the same diagnostic criteria apply, there can be significant differences in the application of these criteria not only by the different pathologists but by one pathologist at different times.

The two series of sections examined were not entirely comparable since in the first series there were more cases of dysplasia and carcinoma *in situ* than in the second series. It might be thought that it is more difficult to distinguish these lesions than either the more bland lesions or invasive carcinoma and that they would therefore give rise to more diagnostic discrepancies. Examination of the correlation diagrams shows that this is what happened, for there is greater scatter in categories 3 and 4 than in other categories. Nevertheless, in the more difficult first series A and C were in closer agreement than in the less difficult second series. The availability of standard photographs of cervical lesions in the second test of series 2 did not make agreement closer between A and C. Fletcher and Oldham (1951) used standard radiographs for comparison in their studies of pneumoconiosis and found that although they helped the less experienced person to be consistent they did not help the more experienced observer. This is probably because the more experienced worker has a more clearly and firmly defined mental picture of the condition than the less experienced and therefore he does not rely on the standard photographs. Firmly defined as this mental picture is, it is not immutable. This is illustrated by the shift in A's diagnoses in the course of 18 months when he was working with other pathologists on the problem of histological criteria.

Kirkland (1963) states that: 'In recent years it has been suggested that anything from 4% to 65% of these atypical changes (in the cervix) precede or progress to carcinoma *in situ*.' This discrepancy is most probably caused by variation in diagnostic criteria. The recent papers of Govan, Haines, Langley, Taylor, and Woodcock (1966) and Grubb and Janota (1967) illustrate this. For example, Grubb and Janota include under the term 'intraepithelial' carcinoma lesions which Govan *et al.* term 'severe dysplasia'. Until such discrepancies are resolved, either by common agreement or as a result of further knowledge, Ashley's (1966) contention that any competent pathologist can diagnose carcinoma *in situ* requires qualification. It would seem unwise to combine, as he does, results from different centres for the purpose of epidemiological survey. The type of analysis used in this paper can be employed to show the diagnostic discrepancies that exist between different laboratories so that in any large-scale survey diagnostic criteria can be standardized. This type of analysis can also be used in a single laboratory to test the consistency and agreement obtaining between the different pathologists and especially of pathologists in training, and, as A's experience shows, in testing and revealing any slight change in the criteria of an individual pathologist with the passage of time and increasing experience.

We wish to thank Dr. A. M. Adelstein for statistical advice and Dr. R. Ollerenshaw for preparing the histograms.

REFERENCES

Ashley, D. J. B. (1966). *J. Obstet. Gynaec. Brit. Cwlth*, **73**, 372.
Committee for Histological Definitions (1962). Editorial. *Acta cytol. (Philad.)*, **6**, 235.
Fletcher, C. M., and Oldham, P. D. (1949). *Brit. J. industr. Med.*, **6**, 168.
——, —— (1951). *Ibid.*, **8**, 138.
Govan, A. D. T., Haines, R. M., Langley, F. A., Taylor, C. W., and Woodcock, A. S. (1966). *J. Obstet. Gynaec. Brit. Cwlth*, **73**, 883.
Grubb, C., and Janota, I. (1967). *J. clin Path.*, **20**, 7.
Kirkland, J. A. (1963). *Ibid.*, **16**, 150.
Mather, K. (1943). *Statistical Analysis in Biology*, p. 55. Methuen, London.
Siegler, E. E. (1956). *Cancer (Philad.)*, **9**, 463.