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 I. 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

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 almost 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 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

<table>
<thead>
<tr>
<th>Case No.</th>
<th>Diagnosis Original (O)</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
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 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
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 observa-
tions (the trend) is relatively large but the variance
of the difference, as shown by the width of the histo-
gram, 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 collabo-
rating 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 histopatho-
ologists and found considerable and disturbing
disagreement in their diagnoses. This present inves-
tigation 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 patholo-
gists 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 corre-
lation 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 availa-
bility 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 ex-
perienced 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, Lan-
gy, 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 single labora-
tory 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

Committee for Histological Definitions (1962). Editorial. Acta cyt-
ol. (Philad.), 6, 235.
6, 168.
Govan, A. D. T., Haines, R. M., Langley, F. A., Taylor, C. W., and
Mather, K. (1943). Statistical Analysis in Biology, p. 55. Methuen,
London.
Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri
J. Cocker, H. Fox and F. A. Langley

doi: 10.1136/jcp.21.1.67

Updated information and services can be found at:
http://jcp.bmj.com/content/21/1/67

These include:

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/