Quality assessment in cervical cytology: a pilot study

G D H THOMAS,* C HEAD, J THOROGOOD

From the *Department of Pathology, Royal Halifax Infirmary, Halifax and the Data Processing Department, Yorkshire Regional Cancer Organisation, Cookridge Hospital, Leeds

SUMMARY  A pilot study of an external quality assessment scheme was run between February 1985 and September 1986 to (i) assess the feasibility of running one from a district general hospital; (ii) to estimate the time required to organise and run it with a computer; (iii) to provide sound statistical results with which future schemes could be compared. Seven laboratories participated, and the 20 smears selected from each laboratory were circulated in three rounds in batches of seven, seven, and six according to a prearranged order. Results analysed using the $\kappa$ statistic showed moderate levels of interlaboratory agreement, with complete agreement emerging only on a small proportion of cases.

Quality assessment schemes are widespread in most branches of pathology but are uncommon in the more subjective areas of histopathology and cytology. A need for a quality assessment scheme in cervical cytopathology has been identified and the format for such a scheme suggested. We carried out a quality assessment scheme in cervical cytology to assess its feasibility in a district general hospital, to estimate the time required to organise and run it using a computer, and to provide sound statistical results which could form a basis for comparing future schemes.

Material and methods

Six district general hospital laboratories and one teaching hospital laboratory with gynaecological cytology workloads between 15 000 and 35 000 annual cases participated in the pilot study. The seven laboratories were each asked to select 20 smears to reflect the full range of their cytological reporting, including normal smears, but with about two thirds of the smears showing some degree or type of neoplasia. Participants were discouraged from selecting problematical smears but the final choice of material was left up to the individual laboratories. It was hoped that by choosing smears in this way a full spectrum of roughly equally distributed diagnostic categories would be seen by all participants, thus permitting sound statistical analysis.

The smears, with relevant clinical details, were circulated in three rounds in batches of seven, seven, and six according to a prearranged order. Laboratories were given one week to report on a set of smears and send the results to JT and the Yorkshire Regional Cancer Organisation before sending them to the next laboratory.

A results proforma was designed with diagnostic categories similar to those on the HMR101 and a comments section to be used as required. Once all the reports for a round had been received laboratories were instructed to recirculate their smears accompanied by relevant histological material. It was hoped this would act as a confirmatory and educational exercise. Tables depicting results of a completed round were distributed to all the participating laboratories. GDHT was responsible for contacting laboratories about inadequate or unclear information.

To measure the amount of interlaboratory agreement about degree or type of neoplasia we wanted to compare the reports of an individual laboratory with those of all the other participating laboratories and calculate some appropriate measure of agreement. The measure also needed to take into account the possibility of agreement due to chance. Such a measure is the $\kappa$ statistic introduced by Cohen. A discussion of its application to histological diagnosis is given by Silcocks. Kappa can be calculated by:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o =$ the observed proportion of agreement and $P_e =$ the expected proportion of agreement. Table 1 shows an example calculation. Kappa takes the value +1·0 for perfect agreement and zero for chance agreement. The maximum value of $\kappa (\kappa_m)$ is reduced...
216  
Thomas, Head, Thorogood  

Table 1  Example of κ calculation from one laboratory  

<table>
<thead>
<tr>
<th>Laboratory diagnosis</th>
<th>Other laboratories' diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benign</td>
</tr>
<tr>
<td>Benign</td>
<td>160</td>
</tr>
<tr>
<td>CIN I/II/III</td>
<td>46</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>208</td>
</tr>
</tbody>
</table>

\[ P_o = \frac{160 + 397 + 71}{786} = 0.80 \]

\[ P_e = \left[ \frac{(208 \times 192) + (464 \times 480) + (114 \times 114)}{786} \right] / 786 = 0.45 \]

\[ \kappa = \frac{0.80 - 0.45}{1 - 0.45} = 0.64 \]

when the contingency tables used to calculate it have unequal margins. To calculate \( \kappa \) we used the formula:

\[ \kappa_m = \frac{P_{om} - P_e^3}{1 - P_e} \]

where \( P_{om} \) is found by taking the smaller of the row or column totals for each category, summing these values, and dividing by the number of items categorised. The proportion of disagreements explained by systematic bias (\( P_s \)) is given by \( P_s = 1 - P_{om} \).

The disagreement that cannot be explained by bias is the difference between \( P_o \) and \( P_{om} \), and this is referred to as haphazard disagreement (\( P_h \)). Therefore \( P_o + P_{om} + P_h = 1 \). An approximation to the standard error of \( \kappa \) is given by:

\[ \sigma_\kappa = \sqrt{\frac{P_e(1 - P_e)}{N(1 - P_e)^2}} \]

where \( N \) is the total number of observations. With \( N > 100 \) the distribution of \( \kappa \) will approach normality allowing confidence intervals to be defined in the usual way—that is, 95% confidence = \( \kappa \pm 1.96 \).

Results

At the end of the study 140 cervical smears had been circulated. This should have produced a total of 980 reports, but two slides were broken part way through the scheme and it was therefore impossible to obtain a complete set of reports for these smears. For this reason we decided to remove these smears from the analysis altogether. This left us with a total of 966 reports.

In the interim, results tables were compiled at the end of each round and reports of the individual smears were totalled and combined with a condensed set of comments to produce a table for each of the participating laboratories. Each laboratory received a complete set of tables which had their own reports highlighted. Final analysis was performed on a combined data set containing the reports from all three rounds. To enable us to calculate \( \kappa \), contingency tables were

Table 2  \( \kappa \) statistics using seven categories

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Observed ( \kappa ) value (95% confidence limits)</th>
<th>( \kappa_m )</th>
<th>( P_s )</th>
<th>( P_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.45 (0.41-0.49)</td>
<td>0.84</td>
<td>0.12</td>
<td>0.31</td>
</tr>
<tr>
<td>B</td>
<td>0.45 (0.41-0.49)</td>
<td>0.84</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>C</td>
<td>0.49 (0.45-0.53)</td>
<td>0.86</td>
<td>0.11</td>
<td>0.29</td>
</tr>
<tr>
<td>D</td>
<td>0.47 (0.43-0.51)</td>
<td>0.92</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>E</td>
<td>0.35 (0.30-0.39)</td>
<td>0.89</td>
<td>0.08</td>
<td>0.42</td>
</tr>
<tr>
<td>F</td>
<td>0.46 (0.42-0.51)</td>
<td>0.87</td>
<td>0.10</td>
<td>0.32</td>
</tr>
<tr>
<td>G</td>
<td>0.40 (0.36-0.45)</td>
<td>0.80</td>
<td>0.16</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 3  \( \kappa \) statistics using three categories

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Observed ( \kappa ) value (95% confidence limits)</th>
<th>( \kappa_m )</th>
<th>( P_s )</th>
<th>( P_h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.59 (0.54-0.64)</td>
<td>0.95</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>B</td>
<td>0.64 (0.59-0.69)</td>
<td>0.81</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.69 (0.65-0.74)</td>
<td>0.92</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>D</td>
<td>0.60 (0.54-0.65)</td>
<td>0.94</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>E</td>
<td>0.50 (0.45-0.56)</td>
<td>0.94</td>
<td>0.02</td>
<td>0.26</td>
</tr>
<tr>
<td>F</td>
<td>0.68 (0.63-0.73)</td>
<td>0.92</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td>G</td>
<td>0.64 (0.59-0.69)</td>
<td>0.96</td>
<td>0.02</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Quality assessment in cervical cytology: a pilot study

Table 4  No of laboratories in agreement

<table>
<thead>
<tr>
<th>No of laboratories in agreement</th>
<th>Diagnosis</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Benign</td>
<td>CIN I/II</td>
<td>CIN III</td>
<td>Invasive squamous</td>
<td>Adenocarcinoma</td>
</tr>
<tr>
<td>All 7</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6 or more</td>
<td>25</td>
<td>14</td>
<td>10</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5 or more</td>
<td>31</td>
<td>28</td>
<td>17</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4 or more</td>
<td>36</td>
<td>44</td>
<td>27</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

produced in which the 138 reports of an individual laboratory were compared with the 828 reports from the others. We compiled these tables in two ways. Firstly, we used seven categories:

1 Inadequate
2 Benign
3 CIN I/II (mild/moderate dysplasia)
4 CIN III (severe dysplasia/carcinoma in situ)
5 Invasive squamous carcinoma
6 Adenocarcinoma
7 Not otherwise specified or other

Secondly, some categories were combined to give the following:

1 Benign
2 CIN I, II, or III
3 Invasive squamous or adenocarcinoma

These produced two \( \kappa \) statistics, their associated confidence intervals, and values for \( p_a \), \( p_o \), and \( \kappa_m \) for each laboratory (tables 2 and 3).

The number of laboratories making the same assessment of a smear for each diagnostic category was calculated (table 4). In only 22 cases were all seven laboratories in complete agreement, almost half of these smears being classified as benign. Five, six, or seven laboratories agreed on the diagnosis in 85 smears; 31 of these were assessed as benign, 28 as CIN I/II, 17 as CIN III, and so on.

Discussion

We have shown that the cluster design of a quality assessment scheme works in a district general hospital, but we found that the expected completion time of nine months was actually two years. This was partly due to the great increase in cervical cytology workload experienced during 1985–6 by all laboratories, and partly to circulating too many smears in each batch. The main reason for trying to carry out this scheme quickly was to try and ensure that participating laboratories did not change their diagnostic criteria as a result of receiving interim results while the scheme was in operation. Bird et al showed no change in the concordance rates of expert panel members among themselves and submitting pathologists in the diagnosis of lymphomas over a period of five years. From this it would seem that we could expect little change in reporting habits of participating laboratories over the two years it took to complete this scheme and that the results could be taken as a baseline against which to assess changes in the participating laboratories' diagnostic criteria in future schemes.

The value of the \( \kappa \) statistic has been discussed by Silcocks, and he suggests that it is a statistical tool applicable to laboratory quality control which can provide a uniform criteria of repeatability.

In the study by Hicklin et al use is made of the Pearson correlation coefficient. This is not the ideal statistical method for analysing agreements of opinion because it assumes that the report of one observer influences another. From their study, calculation of the \( \kappa \) statistic for the two participating laboratories can be made (value 0·73 if five categories similar to ours are used or 0·43 if their full 10 are used). The correlation coefficient was also used by Ooms et al in their study of bladder tumour grading. They also illustrated their results pictorially, which did not show how pathologists differed among themselves as regards a consensus view. Quoting percentage agreements fails to take account of the degree of agreement which could arise by chance.

Husain et al admitted that their numbers were small, and in some instances too small to calculate the \( \kappa \) statistic. Because of the large numbers in our study, the statistical results can be justified, and this is reflected in the smaller 95% confidence limits for the value of \( \kappa \). Our statistical method measures each laboratory's opinion against all the other laboratories' diagnoses in turn and thus does not assume that the opinion of any one laboratory is a better reference point than another, as has been the case in previous studies.

Quality assessment should give a reflection of pathologists' and technicians' performance over the full range of a laboratory's practice; in an observer variation study, at the outset all observers agree that the material is adequate for the diagnosis under study. In both sets of results our \( \kappa \) values were lower than those of Husain et al. Those authors admit that the laboratories taking part in their study were self selected, highly motivated and were regarded as "elite" (Husain OAN, personal communication). Although none of the laboratories participating in our
study would regard itself as expert, it is both interesting and encouraging to note that our values for $\kappa$ were similar to those found in studies of observer variation.\textsuperscript{11,12}

Tables 2 and 3 show that haphazard disagreement ($p_3$) forms the larger proportion of disagreement. Haphazard disagreement arises because observers do not use the same criteria for categorisation. In theory this element of disagreement should be reduced by training, and the role of quality assessment schemes in this respect has yet to be proved. Systematic disagreement ($p_2$) which forms a smaller proportion of the disagreement arises from observers' expectation as to the numbers of smears in each diagnostic category. In routine cervical cytology screening this element of variation is unlikely to be reduced because screeners do not know in advance what sort of distribution of diagnostic categories to expect in any one day.

It has been suggested that a scoring system should be established in histopathology quality assessment schemes.\textsuperscript{13} To score a diagnosis as "right," "close," or "wrong," there has to be a way of defining the "right" diagnosis. Table 4 shows that if a "right" diagnosis is taken as a majority diagnosis (four or more laboratories in agreement) then a "right" diagnosis emerges in 123 of 138 possible cases (89%). If five or more laboratories in agreement were to be taken as indicating the "right" diagnosis such a diagnosis emerges in 85 cases (62%). This would mean that in either 15 (11%) or 53 (38%) cases, depending on how a "right" diagnosis is defined, such a scoring system could not be applied. These percentages are very similar to those that can be derived from table 1 of the paper by Holmquist et al.\textsuperscript{14}

By using the $\kappa$ statistic with the derived values of haphazard and systematic disagreement,\textsuperscript{2} problems of scoring do not arise and individual laboratories can see how, in general terms, their diagnostic criteria differ from those of their colleagues. This, taken with the feedback of interim results, gives a clear picture of how each laboratory reports in relation to the other laboratories.

We found that circulating histological material proved impractical. The intention was to provide some sort of confirmatory or educational aspect to the scheme, but the extra material circulating became too bulky and was too time consuming to select and examine. Histological diagnosis of subsequent biopsy material cannot be taken as supplying a "correct" cytological diagnosis, and the histological diagnosis and grading of cervical neoplasia is also subject to observer variation.\textsuperscript{16} The problems this may cause are mentioned by Whitehead and Woodford.\textsuperscript{15}

Use of a computer in quality assessment is rare. We found the advantages are that the large amount of data produced by this quality assessment scheme were easily stored and reproduced, and the time to input results into the computer was no greater than that required for a manual method. In this study the REPORT facility of the statistical package for social sciences (SPSS-X) was used to generate the interim results. The contingency tables required were also produced by SPSS-X but the calculations of the $\kappa$ statistics were performed manually from these tables as there was not suitable software readily available.

Sherwood and Hunt assessed pathologist, secretarial and technical time in the coordinating laboratory at 22 hours per round in a histopathology quality assessment scheme involving six laboratories, each submitting two cases per round.\textsuperscript{16} This could be extrapolated to 66 hours for three rounds. Their figures do not mention development time. We also found that input by a pathologist was advantageous; about one telephone call a week encouraged circulation of smears and gave a chance to air and sorting out of problems. Our estimate of time involved includes the initial development period, day to day running, and time required to produce results. Averaged over the two years the scheme took to run, our workload worked out at around 40 minutes for each of the three authors a week. In future schemes this development time should be reduced.

Most of our values for $\kappa$ lie in the range 0.41 to 0.60, which Landis and Koch\textsuperscript{17} defined as reflecting moderate agreement. Brown and Brown suggested that all that is required from a cytological report is a comment on the presence of intraepithelial neoplasia or invasion.\textsuperscript{18} In the recommendations of the British Society for Clinical Cytology Working Party follow up of some grades of abnormality can be done cytologically.\textsuperscript{19} To avoid completely overloading a colposcopy service some reliance needs to be placed on the grading of cytological abnormalities. We were unable to show complete agreement at the benign/evidence of neoplasia (no matter what grade) or at the evidence of intraepithelial neoplasia/invasion levels, so perhaps it is these areas in cervical cytology quality assessment schemes on which attention should be focused in future.

We thank the YRCo and the histopathology group for their support for this scheme and gratefully acknowledge a grant from the DHSS. Gratitude is also expressed to all the cytoscreeners, medical laboratory scientific officers, pathologists, and clerical staff who worked in the scheme, and to Miss D M N Hartley for typing the manuscript.

References

2 Husain OAN, Butler EB, Woodford FP. Combined external
Quality assessment in cervical cytology: a pilot study


Requests for reprints to: Dr G D H Thomas, Department of Pathology, Royal Halifax Infirmary, Halifax HX1 2YP, England.
Quality assessment in cervical cytology: a pilot study.

G D Thomas, C Head and J Thorogood

doi: 10.1136/jcp.41.2.215

Updated information and services can be found at:
http://jcp.bmj.com/content/41/2/215

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/