J. clin. Path., 1974, 27, 945-950

# Observer variation and quality control of cytodiagnosis

D. M. D. EVANS, GLENYS SHELLEY, B. CLEARY, AND YVONNE BALDWIN

*From the University Hospital of Wales Department of Cytology, St David's Hospital, Cardiff*

SYNOPSIS The aim of quality control of a laboratory investigation is to ensure that similar results are obtained on the same material at different centres. To investigate its practicability in cytodiagnosis, the same cytological material was examined independently at six centres. Each centre supplied material from 20 cases, providing a total of 120 cases, ie, 100 cases excluding the donor centre's own material. The degree of agreement between the centres was studied using (a) the standard National Health Service cytology report terminology, (b) the centre's own terminology, and (c) the recommended recall time. The results revealed close agreement between five out of six centres in the reports obtained in relation to dysplasia and malignancy, namely, less than 3% false negative results and not more than 1·7% false positive results. The recommended recall time provided a similar order of agreement after discrepancies due to the management of inflammatory conditions had been eliminated. There was marked disagreement in the diagnosis of both presence and type of infection. The results indicate that improvement in the quality of cytological material would increase the consistency of cytodiagnosis. Cytodiagnosis itself, being an expression of opinion, does not appear to be an appropriate field for quality control.

The aim of quality control of a laboratory investigation is to ensure that similar results are obtained on the same material at different centres. To investigate its practicability in cytodiagnosis the same cytological material was examined independently at six centres. The present investigation differs from a previous study in this field (Evans and Sanerkin, 1970) in that the material studied was provided in equal proportion by each of the centres concerned instead of being supplied by only one centre. This was in order to diminish bias arising out of the nature of the material, eg, due to staining characteristics which might be more acceptable to one centre than to others.

## Materials and Methods

The six centres participating in the study were The Postgraduate Medical School, Hammersmith, London, The Royal Free Hospital, London, The Radcliffe Infirmary, Oxford, The Queen Elizabeth Hospital, Birmingham, The Southmead Hospital, Bristol, and The University Hospital of Wales, Cardiff.

Received for publication 14 August 1974.

Each centre provided 20 slides and was requested to select the material from the routine intake to the cytology department, removing from this a sufficient number of normal slides to provide a relatively high proportion of abnormal slides. The material supplied was from cervical scrapes with the exception of that from centre D where approximately half the slides were from vaginal aspirates. Every centre was asked to examine the slides in essentially the same way that they would adopt when examining the material from their own clinics. Accompanying each slide was the usual clinical information provided on the National Health Service (NHS) cytology form and the report was made using the NHS coding. In addition each centre gave a written report using its own terminology and also gave a recommended recall time and a report on the presence and type of inflammation.

The centre providing the cytological material was deemed the reference centre for that case and had access to the clinical follow up together with corresponding histological material where this was available. The reports of the reference centres were placed in a sealed envelope until after the relabelled material had been screened by all centres. The sealed envelopes were then opened and the reference reports used to evaluate the results.

## Results

The results were assessed using the following criteria: (1) NHS cytology form coding; (2) recommended recall time; (3) inflammation.

### 1 NHS CYTOLOGY FORM CODING

The analysis of the reports at centre A compared with those of the reference centres is shown in table I.

Where agreement is complete the results fall between the two diagonal lines. The immediately adjoining squares indicate reasonably close agreement, ie, not more than one grade difference. If a smear reported by the reference centre as 2 (normal) was considered by the reporting centre to be 4 (severe dysplasia/carcinoma *in situ*), 5 (carcinoma *in situ* or more advanced lesion), or 6 (glandular neoplasia) this was called a false positive, the converse being a false negative. The false positives were estimated as a percentage of the number of the reference centres' negative reports (coding 2), excluding those which were considered unsatisfactory (coding 1) by the screening centre. The false negatives were estimated as a percentage of the reference centres' reports in coding 3 (mild dysplasia), 4, 5 and 6 after excluding those considered unsatisfactory by the screening centre.

| | | Centre A | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | | |
| R | | | | | | | | | |
| e | 1 | 1 | 1 | 1 | | | | 3 | |
| f | 2 | 4 | 60 | 1 | | | | 65 | (61) |
| e | 3 | 1 | 6 | 4 | 4 | | | 15 | (14) |
| r | 4 | | 2 | 1 | 8 | 7 | 1 | 19 | (19) |
| e | 5 | 1 | 1 | 2 | 4 | 3 | | 11 | (10) |
| n | 6 | 1 | | | | 2 | 4 | 7 | (6) |
| c | | | | | | | | | |
| e | | | | | | | | | |

Table 1 *Analysis of NHS coding results from centre A compared with those of the reference centres*[1]

False positive 0/61 (0%)
False negative 3/49 (6·1%)
[1]The vertical columns of figures represent the results from centre A and the horizontal lines of figures represent the reference centres' reports on the same slides. The figures on the right of the large square are the totals for each of the reference centres' categories. The figures in parentheses are the ones used in the study after exclusion of slides deemed unsatisfactory (coding 1) by the screening centre.

A similar assessment was also undertaken for each of the other five centres, the results being summarized in table II.

It can be seen that, with the exception of centre D, there is a false positive rate of 0-1·6% and a false negative rate of 6-11%. The results from centre D have a noticeably different pattern, with a higher false positive rate (17%) and a lower false negative rate (2%).

| *Screening Centre* | *False Positive (%)* | *False Negative (%)* |
|---|---|---|
| A | 0 | 6·1 |
| B | 0 | 8·5 |
| C | 0 | 11·1 |
| D | 17·1 | 2·1 |
| E | 0 | 5·9 |
| F | 1·6 | 8·0 |

Table II *Summary of results on 120 slides examined independently by the six centres*

If each centre's own slides were excluded and it was assessed on only the 100 slides contributed by the other five centres, the false positive and false negative rates were marginally increased (by about 1%) for each centre, the overall pattern remaining remarkably similar.

If, however, centre D's slides and results were excluded there was a considerable alteration in the results obtained (table III).

| *Screening Centre* | *False Positive (%)* | *False Negative (%)* |
|---|---|---|
| A | 0 | 0 |
| B | 0 | 2·9 |
| C | 0 | 2·9 |
| E | 0 | 0 |
| F | 1·7 | 2·8 |

Table III *Summary of results on 100 slides after exclusion of slides and results from centre D*

It can be seen that the false negative rate for each centre is now reduced to a level which corresponds to a disagreement on not more than one slide. The false positive rate is also very low. These results indicate a close measure of agreement between five out of six centres.

When only these five centres were considered, the false positives and false negatives involved only four slides. The various reports on these four slides were as follows (table IV).

| *Reference Centre* | *Slide Identification Numbers* | | | |
|---|---|---|---|---|
| | *36* | *97* | *103* | *91* |
| | 2 | 6 | 4 | 5 |
| A | 3 | 6 | 4 | 3 |
| B | 2 | 2 | 4 | 1 |
| C | 2 | 6 | 1 | 2 |
| D | 2 | 6 | 5 | 1 |
| E | 3 | 6 | 5 | 3 |
| F | 6 | 6 | 2 | 3 |

Table IV *Analysis of significant differences on individual slides using NHS codine*

It should be mentioned that the comments relating to the first two slides showed greater similarity than the equivalent NHS coding categories. The great disparity of results on the fourth slide is probably a reflection of the poor quality of the material.

The quality of the smears was in fact often criticized (62 smears out of 120) but centres often did not agree as to which smears were unsatisfactory (coding 1). Out of the total of 120 smears, 46 were considered unsatisfactory by at least one centre, 16 by two centres, seven by three centres, four by four centres, and only one smear was considered unsatisfactory by all six centres. The main reasons given for a smear being unsatisfactory were scantiness, bloodiness, poor fixation, and poor staining. Other reasons were air drying, absence of endocervical cells, excessive thickness, and the smear being a vaginal aspirate. The quality of the smears was also found to deteriorate as the study proceeded, sometimes fading to an unacceptable degree.

## 2 RECOMMENDED RECALL TIME

Five categories of recall time were defined, as follows: R1, normal recall, R2, one year or less than normal, R3, six months, R4, one to three months, R5, immediate-abnormal NHS coding. Immediate recall was further subdivided into R5a inadequate specimen, R5b, repeat after treatment, R5c, miscellaneous reasons, eg, antepartum bleeding, abnormally high cell maturation.

The analysis of the recall times recommended by centre A compared with those recommended by the reference centres is shown in table V.

As previously, the figures between the two diagonal lines represent complete agreement and those in the immediately adjoining squares represent reasonably close agreement. The recommendation of a normal recall (R1) instead of one to three months (R4) or immediate (R5) was termed 'delayed' and the converse 'premature'.

The premature recalls were estimated as a percentage of the number of the reference centre's normal recalls (R1) after excluding those recorded R5a, 5b, or 5c by the screening centre. The delayed recalls were estimated as a percentage of the number of the reference centres' recalls in categories 2, 3, 4, and 5 after excluding those recorded R5a, 5b, or 5c by the screening centre.

A similar assessment was also undertaken for each of the other five centres, the results being summarized in table VI. The results reveal a substantial disagreement between the centres on recommended recall times, ranging from 45% premature recalls at centre D to 20% delayed recalls at centre B. As previously, it was found that the errors were marginally increased if each centre's own slides were excluded. Only a partial reduction in the disparity was achieved when the slides and results from centre D were excluded (table VII).

| | | *Centre A* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *5a* | *5b* | *5c* | | |
| R | | | | | | | | | | | |
| e | 1 | 36 | 4 | | | | | | | 40 | (40) |
| f | 2 | 2 | 11 | | | | | | | 13 | (13) |
| e | 3 | | 1 | 1 | 2 | 1 | 1 | | | 6 | (5) |
| r | 4 | 2 | 5 | 3 | 4 | 4 | 2 | | | 20 | (18) |
| e | 5 | | 3 | 2 | 1 | 26 | 1 | | | 33 | (32) |
| n | 5a | 1 | | | | | 1 | | | 2 | |
| c | 5b | 1 | 2 | | 1 | 1 | | | | 5 | |
| e | 5c | | | | 1 | | | | | 1 | |

Table V *Analysis of recall times recommended by centre A compared with those of the reference centres*[1]

Premature 0/40 (0%)
Delayed 2/68 (2·9%)
[1]The vertical columns represent the recall times recommended by centre A and the horizontal lines of figures represent the recall times recommended by the reference centre on the same slides. The figures on the right of the large square are the totals for each of the reference centres' categories. The figures in parentheses are the ones used in the study, after exclusion of 5a (unsatisfactory), 5b (inflammation), and 5c (miscellaneous immediate recall).

| *Screening Centre* | *Premature (%)* | *Delayed (%)* |
|---|---|---|
| A | 0 | 2·9 |
| B | 10·3 | 20·6 |
| C | 16·7 | 9·8 |
| D | 45·8 | 0 |
| E | 20·5 | 15·9 |
| F | 5·3 | 3·4 |

Table VI *Summary of results on 120 slides based on recall times recommended by the six centres*

| *Screening Centre* | *Premature (%)* | *Delayed (%)* |
|---|---|---|
| A | 0 | 1·8 |
| B | 11·4 | 12·0 |
| C | 11·1 | 11·1 |
| E | 17·1 | 10·9 |
| F | 5·9 | 2·1 |

Table VII *Summary of results on 100 slides based on recommended recall times after exclusion of slides and results from centre D*

Since the disparity between the centres was greater than would have been anticipated from the NHS coding figures we decided to investigate the relationship between NHS coding and recall time for each centre. We found that there was a good correlation between the two for coding 4 (severe dyskaryosis/carcinoma *in situ*) as shown in figure 1.

A similar degree of correlation was demonstrated for codings 5 (carcinoma *in situ* or more severe lesion) and 6 (glandular neoplasia) but there was a

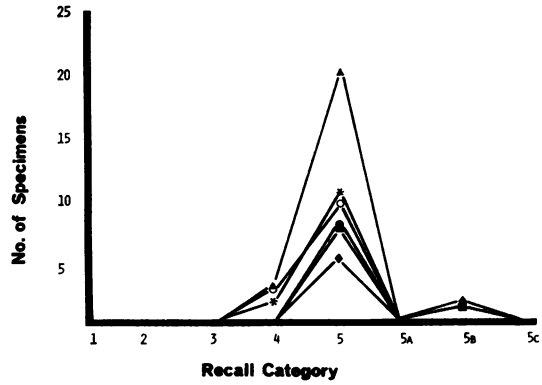


Fig 1 *Correlation between coding 4 (severe dyskaryosis/ carcinoma* in situ) *and recommended recall time.*

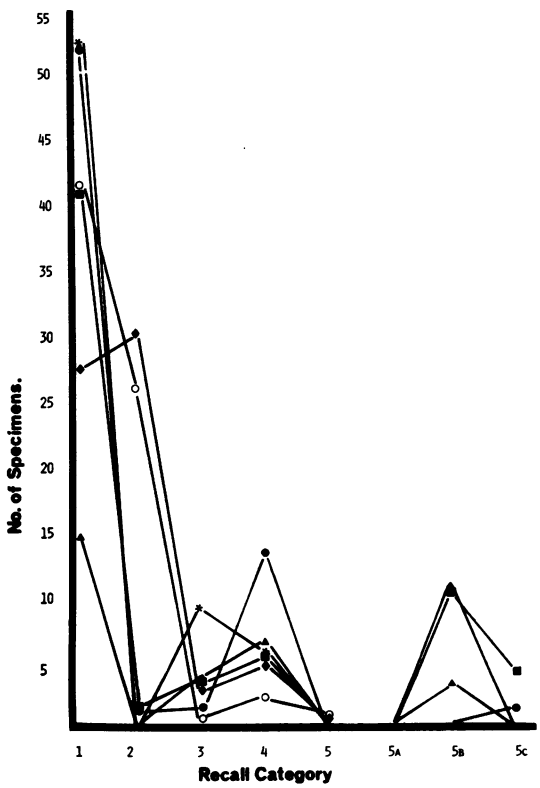


Fig 2 *Correlation between coding 2 (normal smear) and recommended recall time.*

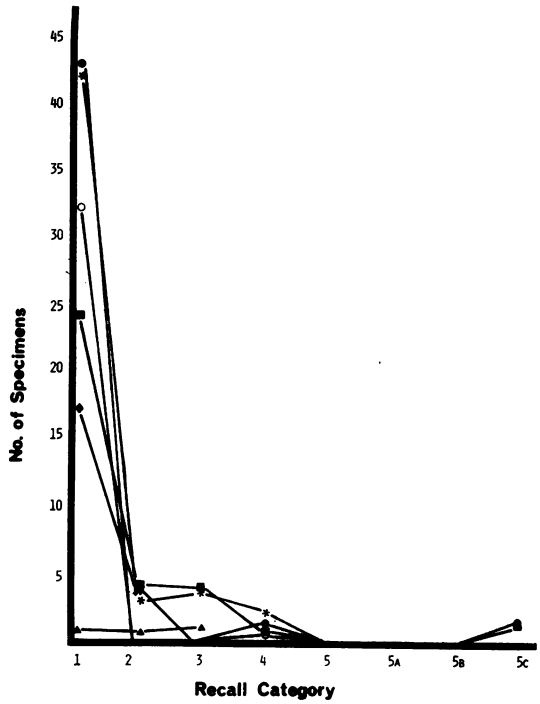


Fig 3 *Correlation between coding 2 (normal smear) and recommended recall time after excluding slides in which inflammation was reported.*

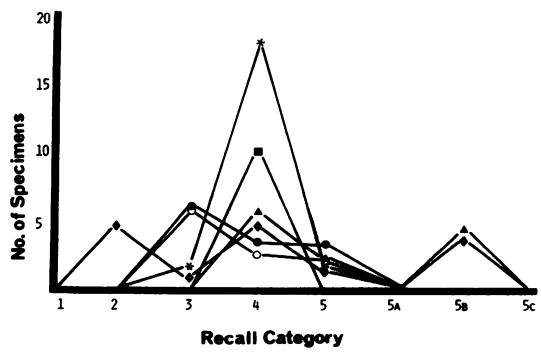


Fig 4 *Correlation between mild dyskaryosis (coding 3) and recommended recall time.*

surprising disagreement over recall time for coding 2 (normal smear) as shown in figure 2.

It was found that the disparity was greatly reduced if all the slides in which inflammation was reported were excluded (fig 3).

Slides reported as mild dyskaryosis (coding 3) also gave rise to a wide range of recommended recall times (fig 4).

Exclusion of the slides in which inflammation was reported did not produce a significant alteration in the pattern apart from eliminating R5b (immediate recall for inflammation).

3 INFLAMMATION

In table VIII the detection of all types of inflammation by each screening centre is compared with that reported by the reference centre. From table VIII it is clear that there was marked disagreement on the existence of inflammation. The extent of agreement between centres on specific types of inflammation was also investigated (table IX). It can be seen that there are again unacceptably large differences between the centres.

| *Inflammation Reported by* | *Screening Centre* | | | | | |
|---|---|---|---|---|---|---|
| | *A* | *B* | *C* | *D* | *E* | *F* |
| a Reference centres only | 41·0 | 48·7 | 30·8 | 10·3 | 46·1 | 15·4 |
| b Reference and screening centres | 59·0 | 51·3 | 69·2 | 89·7 | 53·9 | 84·6 |
| c Screening centre only | 35·9 | 25·6 | 46·1 | 74·3 | 17·9 | 115·4 |

Table VIII *Analysis of reports on presence of inflammation, expressed as a percentage of slides in which inflammation was reported by the reference centres*

| *Tr vaginalis or Monilia Reported by:* | *Screening Centre* | | | | | |
|---|---|---|---|---|---|---|
| | *A* | *B* | *C* | *D* | *E* | *F* |
| a Reference centres only | 25 | 20 | 20 | 5 | 30 | 15 |
| b both reference and screening centres with: | | | | | | |
| (i) exact agreement | 55 | 65 | 60 | 90 | 50 | 75 |
| (ii) partial agreement[1] | 20 | 15 | 20 | 5 | 20 | 10 |
| c Screening centre only | 10 | 10 | 30 | 80 | 45 | 25 |

Table IX *Analysis of reports on type of inflammation, expressed as a percentage of the slides in which Trichomonas vaginalis or Monilia albicans was reported by the reference centres*

[1]The partial agreements are those where the screening centre agreed with the reference centres on the presence of inflammation but disagreed on the cause.

## Discussion

The results were both encouraging and disquieting. It was encouraging to discover that five out of six centres were in very close agreement over the detection and grading of dysplasia and neoplasia. It was disquieting to find that inclusion of material from the sixth centre (centre D) led to a significant increase in the number of false negative results from the other five centres. The reports from the sixth centre appeared to be appropriately graded in relation to its own cytological and histological material but included a significant number of false positive results on the material supplied from the other five centres. This centre clearly had an enhanced sensitivity for less obvious cytological abnormalities.

An important factor in this apparent anomaly could well be the nature of the routine cytological material commonly submitted to the screening centres, vaginal to centre D and cervical to the other five centres. If such material characteristically contains relatively few abnormal cells and the changes though significant, are of relatively slight degree, as often with vaginal aspirates, then a heightened sensitivity is essential if screening is to be effective. When a screener who is used to reporting on vaginal aspirates examines cervical scrape smears, which characteristically show more frequent and severe changes when abnormal, there should be a conscious lowering of sensitivity to compensate for this difference in presentation. If this conscious lowering of sensitivity by the screeners of centre D were not as great as that required, one would expect this centre to overestimate cytological abnormalities compared with the centres dealing mainly with cervical material. This could explain the differences between centre D and the other centres.

Concerning the consistency of detection and grading of dysplasia and neoplasia by the other five centres, there was significant disagreement between them about only four slides, the greatest disagreement being about a single slide whose quality was criticized by nearly every centre.

Maintenance of the quality of cytological material takes place in two stages: first, there is a standardization of the techniques for the collection of material and for the preparation, fixation, and staining of the smear; secondly, there is rejection of unsatisfactory specimens to ensure that those selected for cytodiagnosis are adequately fixed and stained, have plenty of cells and are free from excess blood, etc. An analysis of the results shows that the standards chosen by the different centres for this second stage vary enormously.

Recommended recall time was originally chosen as a means of comparison between centres because it appeared free of ambiguity, unlike commonly used terms such as 'atypical' or 'suspicious' which were found to have different meanings at different centres. The results, however, revealed an unexpected disparity, ranging from 45% premature to 20% delayed recalls. This disparity was caused in the main by two factors, inflammation and mild dysplasia.

When inflammation was diagnosed, an earlier recall time was often recommended independently of the NHS coding for dysplasia and neoplasia. Frequently centres did not agree on the presence of inflammation and often their opinions differed over the type of infection. To a lesser extent the recall for

cases with the same type of inflammation varied between centres. The results indicate a need for improved accuracy in the cytological detection of infection.

The diagnosis of mild dysplasia also gave rise to a wide range of recommended recall times. This may partly have reflected differences in the interpretation of the significance of this finding. Differences in pressure on the available cytological resources were also relevant. Where the resources were sufficient to enable women with normal smears to be recalled annually there was a higher proportion of early recalls for various reasons, including mild dysplasia. Where the resources were hard pressed and normal recalls could only be undertaken every five years, the number of early recalls tended to be fewer.

The close agreement between the centres in the management of severe dysplasia and neoplasia was reassuring. The apparently wide variation in the management of mild dysplasia is seen on closer examination to resolve itself into a difference of months between the recommended recall times. In the great majority of cases this should not be of serious significance, provided that the lesion is in fact mild dysplasia.

The quality of the cytological material was clearly an important factor in cases where there was significant disagreement. It is therefore desirable that there should be agreement between the various centres throughout the country on minimum standards for the content, fixation, and staining of smears. It is probable that its enforcement would involve the rejection of a significant proportion of the type of material at present being used for cytodiagnosis. But this seems unavoidable in view of the significant proportion of false negative results arising from present techniques of collection and preparation (Graham and Meigs, 1949; Cuyler, Kaufmann, Carter, Ross, Thomas, and Palumbo, 1951; Friedell, Hertig, and Younge, 1960; Richart, 1964; Fidler, Boyes, and Worth, 1968; Evans and Sanerkin, 1970; Yule, 1973).

The near agreement achieved at five out of the six centres might suggest that quality control of cytodiagnosis was a practical possibility. However a cytological diagnosis is an expression of opinion and, while it is of considerable value to investigate the consistency of opinions offered by different observers, the subjective nature of opinion makes it an inappropriate field for the application of quality control. The latter should be applied only to tests which are objective and numerical as exemplified by Autonalyzer results. Also, in practice, the circulation of 'standard' smears would appear to be precluded by the inevitable deterioration in smear quality which this study has shown to result from repeated examination under the microscope.

We wish to acknowledge the essential part played by Professor Erica Wachtel, Dr Chandra Grubb, Dr A. Spriggs, Mr M. M. Boddington, Miss Betty Attwood, and Dr D. H. Johnson, Dr Elizabeth McKenzie, and the cytology staff of the six centres participating in this study. We would also like to express our appreciation to Miss D. Symons for typing the manuscripts and to the department of medical illustration for the graphs.

**References**

Cuyler, W. K., Kaufmann, L. A., Carter, B., Ross, R. A., Thomas, W. L., and Palumbo, L. (1951). Genital cytology in obstetric and gynecologic patients: a four-year study. *Amer. J. Obstet. Gynec.*, **62**, 262-278.

Evans, D. M. D., and Sanerkin, N. G. (1970). Cytology screening error rate. In *Cytology Automation: Proceedings of the 2nd Tenovus Symposium*, edited by D. M. D. Evans, pp. 5-13. Livingstone, Edinburgh.

Fidler, H. K., Boyes, D. A., and Worth, A. J. (1968). Cervical cancer detection in British Columbia. *J. Obstet. Gynaec. Brit. Cwlth.* **75**, 392-404.

Friedell, G. H., Hertig, A. T., and Younge, P. A. (1960). *Carcinoma in situ of the Uterine Cervix*, p. 102. Thomas, Springfield, Illinois.

Graham, R. M., and Meigs, J. V. (1949). The value of the vaginal smear. *Amer. J. Obstet. Gynec.*, **58**, 843-850.

Richart, R. M. (1964). Evaluation of the true false negative rate in cytology. *Amer. J. Obstet. Gynec.*, **89**, 723-726.

Yule, R. (1973). The prevention of cancer of the cervix by cytological screening of population. In *Cancer of the Uterine Cervix*, edited by E. C. Easson, pp. 11-25. Saunders, London.