Respiratory cytodiagnosis: study in observer variation and its relation to quality of material

DMD EVANS, GLENYS SHELLEY

From the Area Cytology Laboratory, St David’s Hospital, Cardiff

ABSTRACT Sensitivity and accuracy of cytodiagnosis were assessed in a multicentre study. Six centres each provided sputum cytological material from 20 cases. Each centre screened and reported on the 100 slides provided by the other five centres. The reports were assessed against consensus reference diagnoses, reached by discussion with transparencies, histological sections, and closed-circuit television. False positive rates of 0-4% (average 1·3%) and false negative rates of 0-12% (average 5·0%) of slides examined were recorded. The order of agreement on the three common cell types was adenocarcinoma 75% (50-91%), squamous cell carcinoma 80% (59-94%) and small carcinoma 95% (71-100%). The effect of quality of material on cytological opinion was assessed by comparing disagreement rates on each of the different sets of 20 slides. Disagreement varied from 1% to 23% depending on which set of material was examined.

Although the prognosis for lung cancer is generally poor, men with resectable squamous cell carcinoma have a 45-60% three-year survival after resection.1-3 Cancer cell detection and cell type recognition are critical factors in determining resectability. The aim of the present study was to assess observer variation at six different centres in the interpretation of sputum cytology preparations. Similar studies have been undertaken for cervical cytology4,5 but to our knowledge, this is the first in the field of respiratory cytodiagnosis. It also appears to be the first to undertake a quantitative assessment of the effect of quality of material on observer variation.

Methods

The material examined consisted of 120 sputum cytology slides, stained by the Papanicolaou technique, which were provided by the centres involved. The six centres participating in the study were St Mary’s Hospital, Manchester; St Stephen’s Hospital and St Mary’s Hospital, London; Brighton General Hospital; Southmead Hospital, Bristol; and the Area Cytology Laboratory, Cardiff. Each centre provided 20 sputum cytology slides, of which the numbers of slides containing neoplastic cells were 11, 12, 15, 9, 11, and 11 (average 11) in the sets from centres A to F respectively. It was requested that slides with neoplastic cells (grades 4 and 5) should have histological or clinical confirmation or both. They were selected from the routine diagnostic material without an undue proportion of rare or difficult cases. At five centres the routine material was prepared by direct sputum smear and at the sixth by a sputum concentration technique involving methyl cysteine hydrochloride.6 The centre providing each set of 20 slides was deemed the reference centre for that set. The reports of the reference centres were placed in sealed envelopes until after the study material had been screened.

Each centre was asked to examine the material by the same procedure used for routine cytodiagnosis. Experience with a similarly conducted study on cervical cytology6 had shown that it was advisable to exclude each centre’s own set of slides from its test material: some centres recognised their own slides and some did not, thus introducing a variable factor which could cause disparity in results. So each centre screened the 100 slides provided by the five other centres. This was considered the best way to provide comparable study material for each centre.

After the material had been screened, a meeting was held, attended by at least one representative from each centre, to review all the relevant material and information on cases in which there was difference of opinion, using colour transparencies and the test slides with related histological sections projected on closed-circuit television. No alterations of the reporting centres’ original diagnoses were
permitted. In more than 90% of the cases, the original diagnoses of the reference centres were confirmed. In a few instances where the cell type of the original histological diagnosis was open to question or where subsequent relevant information had been obtained, there was an agreed revision of the original reference centre’s diagnosis, the following results being based on the consensus diagnoses. It was evident from this review that most of the difficulties were encountered with poor preparations and not with rare or unusual cases.

Results

Degree of Abnormality
The grading of the severity of abnormality was based on each centre’s usual reporting nomenclature. Six clearly defined categories were identified: 0 unsatisfactory specimen; 1 normal; 2 hyperplasia, squamous metaplasia, or changes of the order seen in inflammation; 3 dyskaryosis or atypical metaplasia; 4 suspicious of malignancy; 5 diagnostic of malignancy. The results of comparing each centre’s reports with the reference diagnoses are summarised in Table 1. A similar table was prepared for each of the other five centres.

<table>
<thead>
<tr>
<th>Screening centre</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The vertical columns represent the reports of the screening centre and the horizontal lines represent the reference diagnoses on the same slides. The false positive and false negative reports are represented by the figures enclosed in circles. Overestimates and underestimates include the latter together with those enclosed in squares.

<table>
<thead>
<tr>
<th>Reference diagnosis</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>13</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Analysis of reports on grades of malignancy by one screening centre

<table>
<thead>
<tr>
<th>Centre</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>Overestimates</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1.3</td>
</tr>
<tr>
<td>False positives</td>
<td>12</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>5</td>
<td>5.0</td>
</tr>
<tr>
<td>False negatives</td>
<td>15</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>1</td>
<td>5</td>
<td>7.3</td>
</tr>
</tbody>
</table>

Table 2 Summary of erroneous reports

The erroneous reports as defined above are summarised in Table 2.

From this it is clear that in three of the six centres, false positive results and overestimates were significantly fewer and in two centres, marginally fewer than false negative results and underestimates. This pattern was reversed by centre E which had the lowest false negative and underestimate rates but also had the highest false positive and overestimate rates.

Analysis of False Positive Reports

The relationship of false positive reports to each of the different sets of material is represented by Table 3.

The vertical columns represent the false positive errors made by each screening centre. Three of the six centres gave no false positive reports. The average false positive rate for the total number of slides screened was 1.3% (range 0-4%).

The horizontal lines indicate the number of false positive errors made on the set of material provided.

Table 3 Analysis of false positive reports

The vertical columns relate false positive errors to their relevant screening centres; the horizontal lines relate to the source of the material examined.
by each reference centre. No false positive results were made on the material from centres B and E. The material from centre E was made by direct smear. That from centre B was prepared by a concentration technique using methyl cysteine hydrochloride, producing cell morphological characteristics which were unfamiliar to the other five centres. The absence of false positive reports on this material is therefore of particular note. The false positive reports were all made on standard direct smears.

**ANALYSIS OF FALSE NEGATIVE REPORTS**

A similar analysis of false negative reports is represented by table 4.

From an examination of the vertical columns, it can be seen that two of the six screening centres gave no false negative reports. The average false negative rate for the total number of slides screened was 5-0% (range 0-12%).

From the horizontal lines it can be seen that at least one false negative error was made on the material from each centre. Both the lowest (1%) and the highest (12%) false negative error rates were made on standard direct smears with an intermediate (7%) rate on smears prepared by the concentration method.

**ANALYSIS OF DISAGREEMENTS**

On the basis that overestimates (including false positives) and underestimates (including false negatives) are "disagreements", these have similarly been analysed. The results are represented by table 5.

From the vertical columns, it can be seen that the disagreements scored by the screening centres ranged from 3% to 17% with an average of 9-3%. From the horizontal lines it can be seen that comparison of the results on the different sets of material revealed an even wider range of disagreement rate, from 1% to 23% with an intermediate rate of 8% on the material from centre B, prepared by the concentration method.

**CELL TYPE RECOGNITION**

An examination of the reports revealed that the following designations were required in order to analyse the cancer cell types reported: squamous cell carcinoma; small cell carcinoma (including oat cell); adenocarcinoma; other types of designated tumour (for example, melanoma, lymphoma); untyped carcinoma; negative; unsatisfactory smear.

A comparison of one centre's reports with the reference diagnoses is summarised in table 6. A similar table was prepared for each of the other five centres.

The figures between the diagonal tramlines indicate the number of smears of each cancer cell type where the reports from the screening centre agreed with...
those of the reference centres. All the figures outside the tramlines indicate disagreement. Unlike the tables analysing the degree of abnormality, there is no acceptably close agreement on cell type. Discrepancies in cell identification varied with tumour cell type. Agreements on the three most common tumour cell types are tabulated (table 7).

**Analysis of Tumour Cell Type Recognition**

The relationship of cell type disagreements to each of the different sets of material is represented by table 8.

The number of disagreements on cell type recorded by each screening centre range from 3% to 17% with an average of 10.2%. The number of disagreements made on the material from each reference centre range from 2-22%. The lowest disagreement rates of 2%, 3% and 6% were on direct smears with which the screening centres were familiar and the highest disagreement rate of 22% was also on direct smears. The 15% cell type disagreement rate on concentrated material is significantly greater than the 8% rate on degree of abnormality, indicating that with this unfamiliar material, screening centres identified malignant features more readily than cell types.

**Discussion**

The average false positive rate was approximately 2.5% of the negative smears. This would be too high for a population screening procedure and might be considered unsatisfactory for a clinical diagnostic test with its possible implications of unnecessary bronchoscopy, lung biopsy, and subsequent uncertainty. The error rate would presumably have been less if each centre had been reporting on material prepared in its own department. It is, however, of note that where the material was prepared by the concentration technique using methyl cysteine hydrochloride, the false positive rate for all five screening centres was nil.

The average false negative rate was approximately 10% of the positive smears. This implies that under the conditions of the study, 10% of tumours with detectable cells in the sputum were missed. This may be considered too high for efficient cancer detection. One method of increasing the detection rate is by an increased sensitivity as to the significance of minor abnormalities in the cells. Such an increased sensitivity may be the reason why centre E had a false negative rate of 0% and an underestimate rate of about only 2% of positive smears but why they also had the highest false positive rate of about 8% of negative smears.

A similar pattern can be seen in a comparable study on cervical cytology, where centre D had the lowest false negative rate of 2% compared with the average of 7% but also had the highest false positive rate of 17% compared with the average of 3%. This process of increased sensitivity to improve cancer detection rate at the expense of an increased false positive error rate has been described by Langley as moving the "decision line" to the left.

An alternative method of increasing sensitivity, which on the evidence of the present study does not increase the number of false positive results, is to increase the amount of diagnostic material on the slide by a concentration technique. This can be achieved by using a mucusolytic agent such as methyl cysteine hydrochloride, in order to obtain a centrifuge deposit. A comparison of the results using both direct preparation and concentrated material from the same sputum specimens, examined independently by the same screening staff revealed that 30% more cases were detected by the concentration method. In view of the increased sensitivity, all sputum cytology at centre B has since been prepared by this technique, including that provided for the present study. The screeners at this centre actually preferred it to the direct smears because of its greater uniformity, with less need to adjust the focus.

Several screening centres in the study expressed difficulty in examining this unfamiliar material (and also in examining direct smears from at least one other centre) but their results on the concentrated material were slightly better than average, particularly in regard to the absence of false positive results. Cell type identification might be expected to improve on increased familiarity with the slightly different cell morphology produced by this technique. Familiarity with the type of material screened must be considered an important factor in screening accuracy and it is possible that the error rates recorded by the various centres would be significantly lower on their own material than in this study.
Respiratory cytodiagnosis: study in observer variation and its relation to quality of material

These figures underline the need for improvement, both in the preparation of material and in pattern recognition. The material should be prepared in a way which maximises the chance of encountering an identifiable abnormal cell and minimises false positive diagnoses. To improve pattern recognition, there is at present no substitute for the routine checking of cell appearance against the histological section but specimen preparation is at least as important as pattern recognition for accurate diagnosis.

This study differed from our multicentre studies on cervical cytology in having a final assessment meeting attended by representatives of all the participating centres. It was found to be of great value to be able to scrutinise material on which there had been differences of opinion. In a few instances, the final cell type diagnosis was revised by general consensus. No alterations of the screening centres’ original diagnoses were however permitted. At this meeting it was evident that many of the errors and disagreements occurred primarily because of the inadequacy of the specimen rather than the inadequacy of the screener.

We would like to thank Dr EB Butler, Dr DV Coleman, Dr OAN Husain, Dr E McKenzie, Dr DH Melcher, and the Cytology Staff of the six centres, for their participation and help in this study. We would also like to thank Mrs J Jellings for typing the paper. The cost of the final assessment meeting was funded by TENOVUS.

References