Observer variability in endometrial cytology using $\kappa$ statistics

F Tezuka, T Namiki, H Higashiwai

Abstract

Aims: To analyse the current state of endometrial cytology practice by determining the interobserver variability in grading the atypicality of endometrial cells using $\kappa$ statistics.

Methods: A series of 70 clusters of benign, hyperplastic, and malignant endometrial cells on cytology specimens were examined by 19 experienced Japanese cytopathologists. They assigned each cluster to one of three diagnostic categories according to increasing cellular atypicality: negative for, suspicious of, and positive for malignancy. Each observer used their subjective judgement in grading without receiving any discriminatory criteria.

Results: Most classified the series of 70 clusters of endometrial cells into 33 negative, 20 suspicious, and 17 positive grades. However, observer variation was considerable, $\kappa$ statistics showing unfavourable overall agreement ($\kappa = 0.36$). Although there was good agreement on negative and positive ($\kappa = 0.46$ and 0.47, respectively), poor reproducibility occurred in grading suspicious ($\kappa = 0.15$).

Conclusions: Currently, Japanese cytopathologists are not reliable in grading the atypicality of endometrial cells. This could be attributed mainly to the situation in endometrial cytology practice in which there are not generally accepted criteria for defining cells in which malignancy is suspected.

Introduction

Endometrial carcinoma has a poor prognosis if malignant cells have been disseminated widely before treatment, but the prognosis is relatively favourable if the cells are confined exclusively within the endometrial mucosa. There has been widespread interest in developing a screening programme to diagnose the disease cytologically at an early stage when it should be most amenable to treatment.

Early detection of carcinoma depends on the validity of the diagnostic cytology and in the diagnostic practice of endometrial cytology there seems to be considerable interobserver variability. It is important for cytopathological diagnosticians to check the reliability of their reporting. This study therefore examined the level of agreement among Japanese cytopathologists in grading the cellular atypicality of endometrial cells.

Methods

Nineteen cytopathologists, considered to be representative of cytologic practice in northern Japan, participated. They were from the cytology divisions of Tohoku University Hospital (3), National Sendai Hospital (2), Miyagi Cancer Society (12) and Nippon Telegraph and Telephone Cooperative Tohoku Hospital (2), and represented a range of consultant experience (five to 20 years).

The study comprised 29 patients in which both cytological and histological diagnoses on endometrial specimens were almost simultaneously made from 1985 through 1990 at National Sendai Hospital. Histopathological examination classified these patients as follows: 13 with well differentiated adenocarcinoma; three with adenomatous and atypical hyperplasia; and 13 controls without any neoplastic changes. Cytological samples were obtained by the endocyte technique, smeared immediately, fixed in 95% ethanol and stained by the Papanicolaou procedure. From these samples, a series of 70 clusters of epithelial cells were selected microscopically at random for analysis of interobserver variability. True diagnosis was previously unknown for each of these clusters, because histopathological diagnosis could not be applied to individual clusters as the “gold” standard. Clusters selected from samples obtained from patients with histopathological diagnosis of adenocarcinoma might include all grades of atypicality from normal to malignancy.

The series of 70 clusters were circulated in batches to 19 observers, who independently made assessments “blind”. They assigned each cluster to one of the following three diagnostic categories according to increasing cellular atypicality: (1) negative for malignancy; (2) suspicious of malignancy; and (3) positive for malignancy. Each observer used their subjective judgement in grading without receiving any discriminatory criteria.

The method for the analysis of interobserver variability using $\kappa$ statistics has been described in detail. $\kappa$ statistics are a measure...
of agreement which requires no assumption concerning the "true" diagnosis or the "gold" standard and incorporates a correction for the amount of agreement to be expected by chance. It is provided by:

\[ \kappa = \frac{(\overline{P} - \overline{P_e})}{(1 - \overline{P})} \]

where \( \overline{P} \) and \( \overline{P_e} \) are a measure and the expected extent of agreement, respectively. Given that a series of \( N \) clusters is reviewed by \( n \) observers and each of them assigns each cluster to one of three categories of classification, then \( \overline{P} \) is defined as:

\[ \overline{P} = \frac{1}{N} \sum_{i=1}^{N} n_{ij} (n_{ij} - 1) \]

where \( n_{ij} \) is the number of observers who assigned the i-th cluster to the j-th category.

If it were assumed that the observers made their assignment at random, \( \overline{P_e} \) would be:

\[ \overline{P_e} = \sum_{j=1}^{k} p_j^2 \]

where \( p_j \) is the proportion of the total number of assignments which were made to the j-th category, given by:

\[ p_j = \frac{1}{N} \sum_{i=1}^{n} n_{ij} \]

The overall value of \( \kappa \) for more than two categories is the weighted average of the values for individual categories. The value of \( \kappa \) can range from -1.0 to +1.0—a value of 0 indicates chance agreement alone, while a value of +1.0 indicates perfect agreement. A negative value would indicate systematic disagreement among observers. It is generally accepted that a value of 0.75 or above reflects excellent agreement, 0.75-0.4 fair to good agreement, and values of less than 0.4 poor agreement. Formulae for the standard error of \( \kappa \) were derived by Fleiss et al.

**Results**

The grades of cellular atypicality allocated by 19 observers are presented in table 1. The diagnosis which occurred most often on the 19 report forms classified a series of 70 clusters of endometrial cells into 33 negative (47.1%), 20 suspicious (28.6%), and 17 positive grades (24.3%). Considerable variability was easily recognised from a comparison between allocated proportions of grades. One observer (K) allocated more positive grades (71.4%) and less negative (4.3%) than most, while another observer (1) allocated more negative (72.9%) and less positive (11.8%) grades. Grading of cellular atypicality agreed completely on only five out of 70 clusters. Figs 1 and 2 show the cytological features of clusters which all observers graded as negative and positive, respectively, and figs 3 and 4 show those features on which the poorest agreement occurred. Figure 3 was assigned as positive by eight observers, as suspicious by eight, and as negative by three. Most of the observers recognised an irregular overlapping of nuclei on this figure, but assessed its grade very differently. Figure 4 was assigned as positive by one observer, as suspicious by nine, and as negative by nine. On this figure, all observers agreed there was a pronounced increase in nuclear density, but differed on grading its atypicality.

The results of the statistical analysis are presented in table 2. The overall proportion of agreement was 58.5% with a \( \kappa \) value of 0.36, corresponding to a poor level of agreement. The proportions of agreement for individual categories were 69.2% for the negative grade, 61.1% for positive grading and 40.3% for the suspicious grade, \( \kappa \) values being 0.46, 0.47 and 0.15, respectively. There was relatively good agreement on negative and positive grades, but not on suspicious of malignancy grade.

**Discussion**

Analysis using \( \kappa \) statistics showed considerable
Figure 3  Cluster of endometrial cells from a 69 year old woman with adenocarcinoma. The poorest agreement occurred on this figure which was graded positive by eight observers, suspicious by eight, and negative by three (Papanicolaou stain).

Figure 4  Cluster of endometrial cells from a 52 year old woman with atypical hyperplasia. This figure was graded positive by one observer, suspicious by nine, and negative by none (Papanicolaou stain).

<table>
<thead>
<tr>
<th>Cytological grading</th>
<th>Proportionate distribution of assignments</th>
<th>Proportion of agreement</th>
<th>$k$ statistic</th>
<th>Standard error of $k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.430</td>
<td>0.462</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Suspicious</td>
<td>0.302</td>
<td>0.403</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>Positive</td>
<td>0.268</td>
<td>0.611</td>
<td>0.47</td>
<td>0.03</td>
</tr>
<tr>
<td>Overall</td>
<td>1.000</td>
<td>0.585</td>
<td>0.36</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 2  Proportions of agreement and kappa statistics of 19 observers in grading the cellular atypiaclity of 70 clusters of endometrial cells

Interobserver variability indicating that Japanese cytopathologists are not currently reliable in grading the atypicality of endometrial cells. There was relatively good agreement on positive and negative grades, but considerable inconsistency concerning the suspicious grade. This unfavourable result can be partly attributed to the current situation in endometrial cytology in which there are no generally accepted criteria for classifying endometrial cells into benign, hyperplastic, and malignant conditions. Misinterpretations may often occur, especially when attempting to discriminate well differentiated adenocarcinoma from adenomatous and atypical hyperplasia. These two lesions seem to form a continuous spectrum and morphological differences in the cellular atypiaclity are so subtle that even experienced cytopathologists can make a mistake. Some authors have claimed that nuclear size can be the most important indicator of malignancy, because the mean and deviation of nuclear sizes are significantly larger in cancer cells. However, classification of endometrial cells into normal, hyperplastic, and malignant conditions based on the mean values of nuclear area and related parameters has serious limitations. In a study by Skaaland a cutoff of mean nuclear area of 45 $\mu m^2$, which provided the best discrimination for normal and malignant endometrial cells, was associated with 17% false positive and 20% false negative classifications. Fu et al also claimed that a mean nuclear area of 50 $\mu m^2$ gave the highest accuracy with 14% of normal samples, 43% of hyperplasias, and 67% of carcinomas exceeding that value. The scatter of values in different conditions may overlap to such a degree as to render nuclear size unreliable as a diagnostic criterion when used alone. This suggests that correct diagnosis of malignancy should not be based on single parameters, but on comprehensive recognition of multiple variables related to cellular and structural atypia. To improve interobserver reproducibility, criteria should be established for classifying endometrial cells using multivariate analysis.

The result of this study also indicates a need for education among Japanese cytopathologists to improve their ability to classify endometrial cells. Continuing reviews of cells of different grades of atypia to identify and eliminate extraneous sources of inconsistency will ensure that the eventual level of agreement is greater than that reported here.

This study is funded by the Kurokawa-Toshio Grant for Cancer Research 1991.

6 Landsis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
Observer variability in endometrial cytology using kappa statistics.

F Tezuka, T Namiki and H Higashiiwai

doi: 10.1136/jcp.45.4.292

Updated information and services can be found at:
http://jcp.bmj.com/content/45/4/292

Email alerting service

These include:
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/