Multivariate discriminant analysis of normal, intraepithelial neoplasia and human papillomavirus infection of the uterine cervix samples

E. Artacho-Pérula1, R. Roldán-Villalobos1, J. Salas-Molina2 and R. Vaamonde-Lemos1

1Department of Morphological Sciences and 2Reina Sofia Hospital, University of Córdoba, Córdoba, Spain

Summary. The present investigation studies the role of multivariate statistical methods on quantitative histopathological features of cells in uterine cervix epithelium to discriminate between normal and abnormal uterine cervix samples. 143 histological specimens were included in the study involving normal cervix, cervical intraepithelial neoplasia (CIN) lesions and cervical human papillomavirus (HPV) infection with and without CIN (condyloma-CIN and condyloma-NCIN groups, respectively). Deep, middle and superficial regions of the cervical squamous epithelium were morphometrically analyzed. Identification of normal cervix from pathological cases was highly achieved with a specificity of 100%. The application of discriminant statistical method within pathological specimens showed an acceptable percentage of cases correctly classified; thus, an efficiency of 83.0% and 74.6% was obtained in order to discriminate within CIN and condyloma-CIN grades respectively. These percentages increased when differentiation between each grade of CIN versus condyloma-CIN were considered, using only 1-3 morphometrical parameters. Our findings indicate that the combination of nuclear and cytoplasmic size parameters, specially size parameters, permit a high correct percentage classification of cervix samples. The discrimination process was better when few diagnostic categories were included; however, 100% specificity for normal samples was always reached.

Key words: Uterine cervix, Human papillomavirus, Cervical intraepithelial neoplasia, Discriminant analysis

Introduction

There is a considerable amount of information available on cervical human papillomavirus (HPV) infections (Syrjänen, 1983; Gissmann et al., 1984; Kadish et al., 1986; Syrjänen et al., 1987) and cervical intraepithelial neoplasia (CIN) (Fu et al., 1989; Montag et al., 1989). Recently, morphometry, a technique used extensively in biological sciences, has quantitatively described the nuclear and cytoplasmic features of the intraepithelial cells of the uterine cervix for both control and pathological biopsies (Twiggs et al., 1981; Hartman et al., 1986; Hall et al., 1988; Bibbo et al., 1989; Mariuzzi et al., 1989).

Many authors have reported the quantitative detection of uterine cervix disorders (Abdul-Karim et al., 1982; Boon and Kok, 1985; Hall et al., 1988; Montag et al., 1989). Thus, measurements of nuclei and cytoplasm in intraepithelial cells have shown that the use of nuclear area as a diagnostic criterion to differentiate between normal and abnormal cervix is adequate (Tosi et al., 1988). However, isolated quantification of the uterine cervix cells are of little value for discrimination within pathological processes.

In the present study, the diagnostic value of nuclear and cytoplasmic size and shape in the detection of HPV infections and CIN lesions is assessed together with multivariate discriminant analysis to obtain a better differentiation of these processes. Thus, morphometric quantification of the deep, middle and superficial layers of stratified uterine cervix epithelium cells is performed and their features investigated in order to provide a better diagnostic criterion.

Materials and methods

The study was performed on paraffin-embedded cervical tissue specimens from archival cases from the Reina Sofia Hospital (Córdoba, Spain). Sections from paraffin blocks were cut at 6 μm and stained with haematoxylin and eosin. All cases were reviewed and the diagnoses were confirmed independently by two pathologists. Several weeks later, the material was reassessed by the same pathologists. Samples were excluded from the study when they showed autolitic and
deteriorated images, or when there was disagreement between the pathologists.

A total of 143 samples were used for further study, comprising the following groups: 1) Normal: 16 samples; 2) Cervical intraepithelial neoplasia (CIN) lesions: 55 samples (including 7, 11 and 37 samples for CIN I, II and III, respectively); 3) Cervical human papillomavirus (HPV) infection without CIN (condyloma-NCIN): 10 samples; and 4) HPV infection with CIN (condyloma-CIN): 62 samples (including 23, 29 and 10 samples for condyloma-CIN I, II and III, respectively). The classification of the CIN I, II and III was based on histopathological criteria: CIN 1, mild dysplasia; CIN II, moderate dysplasia; and CIN III, severe dysplasia and carcinoma in situ (Abdul-Karim et al., 1982).

The quantitative analysis was performed by one observer experienced in the use of the graphic tablet (E.A.-P). The Leitz A.S.M. semi-automatic image analyser equipped with a cursor was used for all measurements. The cursor is fitted with a light-emitting diode, which can be observed through a Leitz Dialux 20 microscope with a drawing tube for measurements of microscopic images on the graphic tablet. Calibration of the magnification of this microscope/drawing system was performed by using an object microscope graticule of (10 μm) lines. The contours of cytoplasm and nucleus were traced and the data was calculated by microcomputer and recorded in a personal computer.

For each diagnosis and sample, an average of six microscopic fields (range 3-6) were chosen at random within the epithelium where the whole thickness of the epithelium was seen. Within each field, stratified squamous epithelium cells were examined at a final magnification of x 2,200 at the three levels: deep, middle and superficial layers. Thirty cells per layer (range 25-40) were measured at random. This number of cells was selected according to studies by Tosi et al. (1986, 1988).

The morphometric parameters selected for analysis were the area, the perimeter, the maximal, minimal and equivalent circle diameters, nuclear/cytoplasmic ratio (N/C ratio) and several shape factors (Meijer et al., 1980; Marchevsky et al., 1987; Tosi et al., 1988) which included the following: a) Form factor PE (4πarea/perimeter²), its value is 1.00 for a circle and <1.00 for ellipse and irregular structures; b) Form factor ELL (minimal diameter/maximal diameter), its value is 1.00 for a circle and <1.00 for elliptical structures; c) Form factor AR (area/(π/4)maximal diameter/minimal diameter), its value is 1.00 for a circle and an ellipse and <1.00 for irregular structures; d) Form factor CO (perimeter/area)¹/², this parameter takes the minimum value of 3.54 for a circular contour; and e) Form factor CI (perimeter/πequivalent circle diameter), its value is 1.00 for a perfect circle and >1.00 for irregular structures. The quantitative information was obtained for both nuclear and cytoplasmic profiles of the cells.

The mean, standard deviation and coefficient of variation of each feature were computed. Normal distribution was tested with the Kolmogorov-Smirnov test; an alpha level of <0.05 was accepted for significance.

The ability of morphometry to distinguish between different processes was investigated by multivariate discriminant analysis. A total of 63 features were included in each discriminant analysis (11 size parameters and 10 shape parameters in each cervix epithelium region). The 12 discriminant analyses performed are shown in Figure 1. Analyses 1-6 include normal group for discrimination with pathological processes. Analyses 7-12 allow discrimination within pathological processes. The efficiency, sensitivity and specificity was also calculated according to Collan (1989).

All statistical analyses were performed on a Mitac MPC 2000SL personal computer using the SPSS/PC Statistical Software Package procedure DSC. The DSC performs linear discriminant analysis for two or more groups. The aim of discriminant analysis is to classify cases into one of several mutually exclusive groups, based on their values for a set of predictor variables. The classification rule is developed on cases for which group membership is known. The rule can then be used to classify cases for which group membership is not known. A stepwise analysis was carried out where Wilks’ method was used for entering variables into the analysis phase; at each step, the variable which minimizes the overall Wilks’ lambda is entered. Thus, classification function coefficients (Fisher’s linear discriminant functions), discriminant scores and classification information of the cases, and classification results table are displayed.

Results

Figure 1 summarizes the number of variables chosen by multivariate discriminant analyses and the accuracy of the classification for each preselected group. It is evident that the normal group could indeed be discriminated from pathological cases. The first to sixth discriminant analyses showed a 100% recognition rate of normal group samples. The features chosen were mainly size parameters.

The worst overall correct recognition rate appeared when eight cytopathological diagnoses were included (discriminant analysis number one -DA1-). In spite of the high number of variables selected, there was only efficiency discrimination in 48.55% of the cases. If the samples were divided into four groups (normal, CIN, condyloma-NCIN and condyloma-CIN) the total percent of diagnostic concordance was 61.76% (DA2). When the discrimination was performed between three groups (normal, CIN and HPV infection with/without CIN) the overall accuracy rate increased to 68.38% through six variables (five corresponding to size parameters and one to shape factor parameter) (DA3).

The power of discrimination between groups increased when the statistical analyses were performed
Discriminant analysis of uterine cervix
to determine which features separated normal group from CIN group (DA4), from condyloma-NCIN group (DA5) and from condyloma-CIN group (DA6). The overall accuracy rates were 89.86%, 87.50% and 85.54% respectively, while the specificity was always 100%.

These results were obtained with descriptors concerning nuclear size and nuclear/cytoplasmic ratios. The discrimination functions and classification results are shown in Table 1.

The 7th to 12th discriminant analyses were performed

Table 1. Classification matrix for DA4, DA5 and DA6, and coefficients of the discriminant functions.

**Discriminant Functions**

<table>
<thead>
<tr>
<th>CYTOPATHOLOGIST'S GROUPS</th>
<th>PREDICTED GROUP</th>
<th>PERCENT OF CASES CORRECTLY CLASSIFIED</th>
<th>NORMAL</th>
<th>CIN</th>
<th>EQUIVALENT WILK'S LAMBDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>14</td>
<td>14</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>CIN</td>
<td>55</td>
<td>7</td>
<td>48</td>
<td>87.3%</td>
<td></td>
</tr>
<tr>
<td>Total % concordance: 89.86%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>14</td>
<td>14</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Condyloma-NCIN</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>Total % concordance: 87.50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>14</td>
<td>14</td>
<td>0</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Condyloma-CIN</td>
<td>69</td>
<td>12</td>
<td>57</td>
<td>82.6%</td>
<td></td>
</tr>
<tr>
<td>Total % concordance: 85.54%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NMaD-M: nuclear maximal diameter for intermediate cells; N/C ratio-S: nuclear/cytoplasmic ratio for superficial cells; N/C ratio-B: nuclear/cytoplasmic ratio for basal cell; NMiD-S: nuclear minimal diameter for superficial cells; NEdC-S: nuclear equivalent circle diameter for superficial cells.

Fig. 1. Diagnostic groups and discriminant analyses (DA) included in our study. Each box represents a preselected group. For instance, DA1 performs a discriminant analysis between eight groups whereas DA2 analyzes discrimination between four groups (CIN grades and condyloma-CIN grades joined). The number of variables for each DA is shown. The percentage of cases correctly classified is printed in order for each preselected group; thus, DA2 classify correctly 60.4%, 100.0%, 60.0% and 64.2% of CIN, normal, condyloma-NCIN and condyloma-CIN cases, respectively.
to differentiate within pathological groups. The DA7 (considering six sets of patients -three CIN and three condyloma-CIN-) did not clarify these entities sufficiently, since the percentage of diagnostic concordance was only 50%.

In practice, it is interesting to discriminate within CIN group and within condyloma-CIN group. Eight and seven variables were selected respectively (DA8 and DA9). The discrimination was better when considering CIN stages (83.02% of the total cases) than within condyloma-CIN lesions (74.58%). In both analyses, the cases correctly classified decreased with the increase of tumoral dedifferentiation (Fig. 1).

Another decision was to make a correct classification between CIN and condyloma-CIN within the same stage of dedifferentiation. Thus, DA10 to DA12 were carried out (Fig. 1). The variables selected corresponded to both cell size and shape characteristics of superficial and middle layers of the cervix epithelium. The percentage of diagnostic concordance in relation to cytopathological classification was greater when the CIN I versus condyloma-CIN I were considered (89.66% of the total cases); this percentage decreased to 80.85% when CIN III versus condyloma-CIN III were computed. The discrimination functions and classification results are shown in Table 2.

### Discussion

The present study has shown that multivariate statistical techniques are useful for the routine diagnosis and/or classification of normal and pathological samples. The nuclear and cytoplasmic features, both size and shape, are the most common parameters for application in Pathology (Boon and Kok, 1985; Baak et al., 1988; Fu et al., 1988; Norris et al., 1989). After a selection of the problem of interest, the application of an appropriate general procedure and measurements of the morphological characteristics of the samples (Artacho-Pérula et al., 1993), the researchers performed several multivariate analyses. The morphometrical data were analyzed to provide a powerful discriminating tool to determine which parameters were the most suitable to classify the individual cases into specific groups. Furthermore, discriminant analyses provided a probability statement for the morphometric classification of the cases.

Classification rules are seldomly 100% valid. This percentage is obtained with extreme difficulty (Baak et al., 1982). Whereas changes in tissues moving away from normality are more easily detected, the correct classification within pathological entities is lower. This fact was reported by Selkäinaho and Collan (1984) who claim that a reliable distinction between various grades

### Table 2. Classification matrix for DA10, DA11 and DA12, and coefficients of the discriminant functions.

<table>
<thead>
<tr>
<th>CYTOPATHOLOGISTS GROUPS</th>
<th>PREDICTED GROUP</th>
<th>PERCENT OF CASES CORRECTLY CLASSIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CIN I</td>
<td>CONDYLoma-CIN I</td>
</tr>
<tr>
<td>CIN I</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Condyloma-CIN I</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Total % concordance: 89.66%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CYTOPATHOLOGISTS GROUPS</th>
<th>PREDICTED GROUP</th>
<th>PERCENT OF CASES CORRECTLY CLASSIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CIN II</td>
<td>CONDYLoma-CIN II</td>
</tr>
<tr>
<td>CIN II</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Condyloma-CIN II</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>Total % concordance: 84.21%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CYTOPATHOLOGISTS GROUPS</th>
<th>PREDICTED GROUP</th>
<th>PERCENT OF CASES CORRECTLY CLASSIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CIN III</td>
<td>CONDYLoma-CIN III</td>
</tr>
<tr>
<td>CIN III</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>Condyloma-CIN III</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Total % concordance: 80.85%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NFFELL-S: nuclear form factor ELL for superficial cells; CFFAR-S: cytoplasmic form factor AR for superficial cells; NMaD-M: nuclear maximal diameter for intermediate cell; CFFCI-M: cytoplasmic form factor CI for intermediate cells; CArea-M: cytoplasmic area for intermediate cells; NArea-S: nuclear area for superficial cells; CMD-M: cytoplasmic minimal diameter for intermediate cells.
of pathological changes with a single morphometric measurement may be difficult due to overlap between grades. When several parameters are applied at the same time, the possibility of obtaining a clear-cut distinction between grades is more possible.

The results of the discriminant analyses in the present investigation lead to several conclusions. In general, the features chosen to better discriminate between groups were mainly cytoplasmic and, specially, nuclear size. When the discrimination process was performed on a high number of groups, the percentage of correctly classified cases decreased. The decrease in the number of diagnostic categories allows better discrimination. The quantitative microscopic classification rules increased to 100% of specificity for normal samples while the efficiency was 85-90% in correctly classifying pathological diagnosis.

The same happened when condyloma-CIN grades were included in our study. In both cases, with the increase of tumoral dedifferentiation the sensitivity to correctly classify cases diminished. Most of the quantitative misclassified cases, were included in phases which were close together. For instance, from five misclassified cases of condyloma-CIN I lesions (22.7% of the total condyloma-CIN I cases), 4 were included in condyloma-CIN II group and only one in condyloma-CIN III group.

An objective of our study was to evaluate the changes in uterine cervix epithelium when it is infected by papillomavirus. The valuation of quantitative data between normal versus condyloma-NCIN groups, CIN I versus condyloma-CIN I groups, CIN II versus condyloma-CIN II groups and CIN III versus condyloma-CIN III groups makes it clear that there is an increase in both nuclear and cytoplasmic size when human papillomavirus infection occurs (Artacho-Pérola et al., 1993). Tosi et al. (1988) reported a progressive increase of nuclear area in the cells of deep, intermediate and superficial layers parallel to that of tumoral dedifferentiation; the association with human papillomavirus includes a higher rate or progression. Our findings indicate that the combination of features can be reliable in correctly classifying these entities, considering that 80-90% of the cases were identified whereas Tosi et al. (1988) could not differentiate CIN III from condyloma-CIN III using bivariate analysis (Student’s t-test).

References
Discriminant analysis of uterine cervix


Accepted September 17, 1993