# *PREDICTION OF VARIOUS GRADES OF CERVICAL PRENEOPLASIA AND NEOPLASIA ON PLASTIC EMBEDDED CYTOBRUSH SAMPLES*

## *DISCRIMINANT ANALYSIS WITH QUALITATIVE AND QUANTITATIVE PREDICTORS*

Jacqueline J. Meulman*
Pio Zeppa**
Mathilde E. Boon***
Wop J. Rietveld****

* Department of Data Theory, Faculty of Social Sciences, University of Leiden, Leiden, The Netherlands

** II Faculty of Medicine and Surgery, Naples, Italy

*** Leiden Cytology and Pathology Laboratory, Leiden, The Netherlands

**** Department of Physiology, University of Leiden, Leiden, The Netherlands

# *PREDICTION OF VARIOUS GRADES OF CERVICAL PRENEOPLASIA AND NEOPLASIA ON PLASTIC EMBEDDED CYTOBRUSH SAMPLES*

## *DISCRIMINANT ANALYSIS WITH QUALITATIVE AND QUANTITATIVE PREDICTORS*


### Summary

The purpose of this study was to investigate whether discrimination into 5 groups of various grades of cervical preneoplasi and neoplasia is possible using discriminant analysis models. Data were obtained for 242 cases diagnosed either as slight dysplasia (n=50), moderate dysplasia (n=50), severe dysplasia (n=50), carcinoma in situ (n=50) or invasive carcinoma (n=42), and consist of qualitative and quantitative features of cells derived from a repeated sample taken from the ectocervix as well as the endocervix using Cytobrushes. The samples were embedded in plastic, and thin sections were prepared resulting in a monolayer of cut nuclei. The percentages of expected correct prediction were obtained by using 10.000 double cross-validation samples; the mean percentage of correct prediction into 5 groups was 65% and into 2 groups (dysplasia versus carcinoma in situ and invasive carcinoma) 91%. The features do not classify the cases in the same way; the discriminant analyses suggest that the quantitative features play an important role in the discrimination of the dysplasias from the carcinoma cases, while the majority of the qualitative features are important in the discrimination within the three dysplasia groups.


**Key words**: Cytobrush, Cervical (Pre)Neoplasia, Canonical Discriminant Analysis, Cross-Validation, Discrimination, Prediction

### Introduction

The cytodiagnost can try to predict the histological diagnosis on the basis of the cells present, in the cervical smear; the prediction is made by weighting of subjective features of the cells. Although there is some correlation between the cytologic prediction and histological diagnosis, the results are far from satisfactory (Helmerhorst et al. 1987). In the Leiden

Cytology and Pathology Laboratory, percentages of correctly classified cases were attained as follows: for slight to moderate dysplasia (taken together into one histological group) 43%, for severe dysplasia 42%, for carcinoma in situ 61%, and for invasive carcinoma 65% (Boon and Bosch 1990).

Wheeler and co-workers used discriminant analysis models for predicting the histological diagnosis on basis of digitized nuclear images, analyzing specially prepared slides in which the cells were in a monolayer (Rosenthal and Suffin 1984; Wheeler et al. 1987). In a relatively small study (total number of cases $n = 27$), they achieved an overall correct prediction of 93% of the cases, when the cases were classified into three groups (moderate dysplasia, severe dysplasia/carcinoma in situ, and invasive carcinoma).

In 1985 a special sampling and preparation technique was introduced in the Leiden Cytology and Pathology Laboratory for women with an abnormal smear. In short, a repeated sample is taken from the ectocervix as well as the endocervix using Cytobrushes. The samples are suspended in a special fixative containing ethyl alcohol and polyethylene glycol. Thereafter the samples are embedded in plastic, and thin sections are prepared resulting in a 'monolayer' of cut nuclei. The advantage of this technique is that no cells are lost, all nuclei can be visualized, no nuclear overlapping occurs, there is minimal nuclear shrinkage, and optimal morphology is obtained (Boon et al. 1990). Subtle chromatin changes of early preneoplasia are visible in these thin slices of nuclei, and mitotic figures are easily discerned; also nucleoli are well visible. These plastic sections are therefore well-suited for analytic and qualitative studies of the morphologic spectrum of cervical (pre)neoplasia.

In the past years, over 200 cases of various grades of preneoplasia were collected, and diagnosed in a subsequently taken biopsy. The number of cases with invasive carcinoma, rather scarce in the Dutch population due to intense screening activities, could be enlarged to 42 by the cooperation of physicians in Indonesia where cervical carcinoma is very frequent. We have used these samples to investigate whether the error rate of prediction of the histological diagnosis can be decreased by using discriminant analyses of qualitative and

quantitative predictors on the cellular material obtained by the non-invasive Cytobrush method. It is known from the literature that it is very hard to distinguish among the 3 groups of dysplasia (Kraemer, 1990). Therefore, one of the main objectives of this study is to examine whether the data make such a fine discrimination possible. Several analyses were performed to explore the data with respect to different use of the predictor variables; they include an analysis with optimal transformations of the quantitative predictors, and the incorporation of a threshold variable that scores whether cases transcend pre-chosen thresholds for two of the quantitative predictors.

## Material and methods

*Patients*

From each patient, one ectocervical and one endocervical sample was obtained with a Cytobrush, bent at an angle of 90 degrees for the ectocervical sampling. The samples were suspended in Leiden fixative (Boon and Drijver 1986). The method to prepare the plastic sections thereof is described in detail in Boon et al. 1990. The histological diagnosis of each patient entered in this study was known from a subsequently taken biopsy. There are 50 cases with mild dysplasia (histological group 1), 50 cases with moderate dysplasia (histological group 2), 50 cases with severe dysplasia (histological group 3), 50 cases with carcinoma in situ (histological group 4), and 42 cases with invasive squamous cell carcinoma (histological group 5), so the total number of cases $n = 242$. The plastic sections were stained according to a modified Papanicolaou method (Boon et al. 1990).

*Establishing the qualitative and quantitative features*

Due to the "toothpick effect" of the Cytobrush used for the sampling (Boon et al. 1990), the abnormal cells were situated in epithelial fragments and easy to localize in the section. For each case, 7 qualitative features of the abnormal cells were determined, in one plastic section. The chromatin changes were evaluated using the models proposed by Patten (1983).

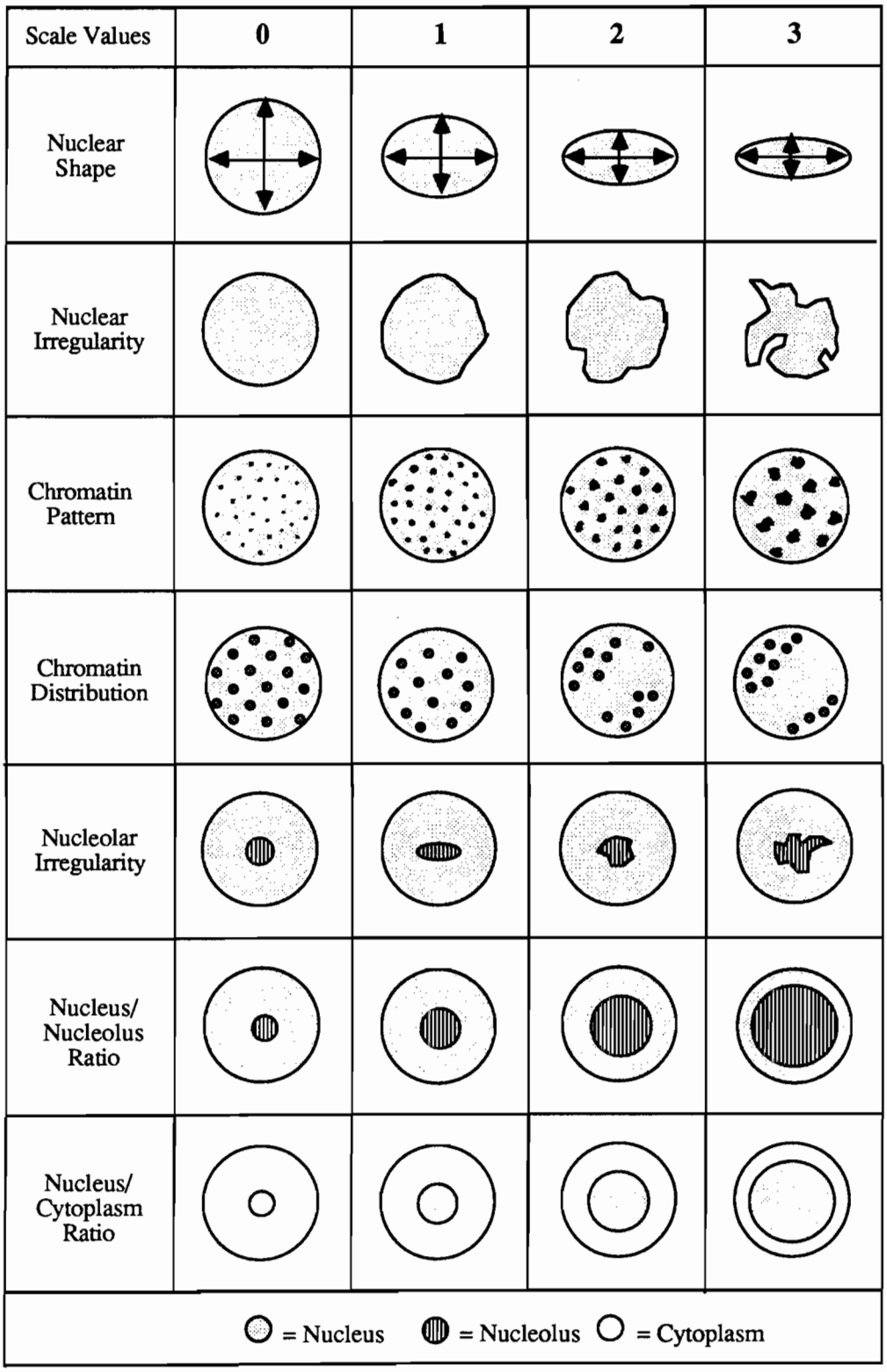


**Figure 1.** Chart for classifying the features

In Figure 1, the chart used for classifying the qualitative features is presented; a 4-point scale, ranging from 0-3, was used. After some experience the system proved to be highly reproducible. The qualitative variables are: *Nuclear Shape, Nuclear Irregularity, Nucleus/Cytoplasm Ratio, Nucleus/Nucleolus Ratio, Nucleolar Irregularity, Chromatin Distribution,* and *Chromatin Pattern.* In addition, 4 quantitative features of the sample were established; these are *Number of (abnormal) Cells per Fragment,* (the mean value was calculated), *Total Number of (abnormal) Cells, Number of Mitoses,* and *Number of Nucleoli* (mean value). Table 1 gives the mean values for the 5 different groups with respect to the qualitative variables, and in Table 2 they are given for the quantitative variables.

**Table 1.** Mean values of qualitative variables per group

| | Nuclear Shape | Nuclear Irregularity | Chromatin Pattern | Chromatin Distribution | Nucleolar Irregularity | Nucleus/ Nucleolus Ratio | Nucleus/ Cytoplasm Ratio |
|---|---|---|---|---|---|---|---|
| 1 | 1.58 | 1.40 | 1.28 | 1.42 | 0.94 | 0.84 | 1.34 |
| 2 | 1.92 | 1.66 | 1.44 | 1.78 | 1.28 | 0.98 | 1.96 |
| 3 | 2.10 | 1.92 | 1.84 | 2.20 | 1.30 | 1.12 | 2.22 |
| 4 | 2.12 | 1.96 | 2.34 | 2.56 | 1.60 | 1.30 | 2.54 |
| 5 | 2.24 | 2.14 | 2.07 | 2.55 | 2.24 | 2.07 | 2.05 |

**Table 2.** Mean values of quantitative variables per group

| | Number of Cells per Fragment | Total Number of Cells | Number of Mitoses | Number of Nucleoli |
|---|---|---|---|---|
| 1 | 21.39 | 62.60 | 0.02 | 1.12 |
| 2 | 32.85 | 121.08 | 0.14 | 1.37 |
| 3 | 32.77 | 111.44 | 0.26 | 1.24 |
| 4 | 106.41 | 481.88 | 4.48 | 1.53 |
| 5 | 125.45 | 485.81 | 6.10 | 1.77 |

It is important to notice that the group means do not always increase from group 1 to group 5. *Chromatin Pattern, Chromatin Distribution,* and *Nucleus/Cytoplasm Ratio* have the highest values for group 4, while for the latter mentioned variable even the mean of group 3, is larger than the mean of group 5. From the means of the quantitative variables, it can be seen that group 2 has a larger mean value than group 3 for the variables *Number of Cells per Fragment, Total Number of Cells,* and *Number of Nucleoli.* These deviations give an

indication that the structure of the data is not one-dimensional; when group means are not monotonic with increasing grades of (pre)neoplasia, prediction should be based on a criterion that goes beyond a single (weighted) sum of the variables.

*Methods*

The data have been analyzed by performing 4 different canonical discriminant analyses, and a Monte Carlo cross-validation study has been carried out pertaining to the results of the analyses. In a canonical discriminant analysis the data are regarded as consisting of two sets of variables: the set of the predictor variables, and the set of the group membership variables. The objective of the analysis is to predict the group membership variables from the predictor variables. The different discriminant analyses were performed to make two different types of comparisons. In the first place a discriminant analysis with all predictors treated as interval level variables has been compared with a so-called nonlinear analysis (Gifi 1990), where the categories of the qualitative variables were optimally scaled during the analysis. In the second place, a comparison has been made using different combinations of the predictors.

Graphically, a canonical discriminant analysis gives two low-dimensional spaces. The first space is the space of the predictor variables; the dimensions of this space are sometimes called canonical variables, and they are obtained by optimal linear combinations of the predictors. In the space of the predictor variables each individual case is represented by a point, and the coordinates of this point are given by the sums of the differentially weighted scores belonging to the case. The second space is the group space; when $K$ groups have to be discriminated, group $k$, where $k$=1,...$K$, defines a binary group membership variable that has the value 1 when a case belongs to group $k$, and has the value 0 otherwise. The dimensions of the group space are given by optimal linear combinations of the group membership variables, and each group is represented by a point in this space. The two spaces are perfectly related to each other: the point representing group $k$ in the group space is equal to the centroid (the center of gravity) of the points in the predictor space that belong

to group *k*. Because of this perfect relationship, the points in each of the two spaces may be represented in a single diagram. The size of the group space (the differential length of the axes) is such that optimal discrimination is obtained.

The 7 qualitative variables and the 4 quantitative variables together form the set of what will be called *source* predictors. The analyses will be discussed in the following order: (a) linear discriminant analysis with the 11 source predictors; (b) discriminant analysis with quantitative variables analyzed on interval level and optimal scaling of the qualitative variables; (c) analysis with source predictors, with addition of a binary *Threshold* variable, derived from *Number of Mitoses* and *Average Number of Cells per Fragment*; (d) analysis as in (c), but *Number of Mitoses* and *Average Number of cells per Fragment* omitted from the analysis; their influence is only channeled through the threshold variable. In the first part of the study, resubstitution has been used to compare the apparent error rates of prediction in the different analyses. Apart from the percentages of cases predicted correctly in the 5 groups, some other statistics have been computed. Since false negatives (classification in a group with a diagnosis less severe than the actual diagnosis) is considered a much more serious error than false positives (classification in a group with a diagnosis more severe than the actual diagnosis), the percentage correct including false positives was also considered.

Apart from the discrimination into 5 groups, we were also interested in the main distinction between the dysplasia groups on the one hand and the carcinoma groups (in situ and invasive) on the other hand. Therefore, a second order classification has been included in the analyses, which means that on the basis of the analysis with 5 groups one joint group point is computed for the dysplasia cases (the centroid of the 1-2-3 points), and one joint group point for the carcinoma cases (the centroid of the 4 and 5 points). The original points in the predictor space were used to re-assign the cases to one of the two new groups (1-2-3 or 4-5).

The resubstitution method might give too optimistic estimates of the expected actual error rate of prediction, because group membership is derived from the analysis in which the model parameters are optimally estimated. In the recent literature, the expected actual error

rate is often estimated by the 'leaving-one-out" method (for example, in Wheeler et al. 1990). Here an optimal linear combination of the predictors is determined for $n$-1 cases, and the assignment of the 'left-out-case' is on the basis of the weights obtained from the $n$-1 active cases. In the statistical literature, this procedure is known as the Jackknife (Miller 1974; Efron, 1979; 1982). Recent studies, however, indicate that the application of the Jackknife is this context does not necessarily give optimal results with respect to error rate estimation (Hirst, Ford, and Critchley 1990). Therefore, a Monte Carlo approach to cross-validation has been used to compare the analyses with respect to the error rate of prediction. Cross-validation on the basis of a randomly determined training sample (consisting of half of the data), determining the predictive power with respect to the other half (the cross-validation sample), might be criticized for its inefficient use of the available data (see Hand 1981). On the other hand, this procedure can be regarded as a very rigorous version of the Jackknife (leaving-half-out), and since the number of ways to split a large number of cases into two subgroups is very large, the cross-validation can be repeated for different split-half samples.

The present study combines the repeatedly performed rigorous Jackknife with double cross-validation (Tatsuoka 1976). In double cross-validation, the total available sample is split into two subsamples of equal size, and a canonical discriminant analysis is performed for each subsample. Then subsample 2 is used as the cross-validation sample for the prediction on the basis of the weights obtained for subsample 1 as training sample, and *vice versa*. 10.000 independent split-half samples were taken from the data, controlled so that each subsample contained 25, 25, 25, 25 and 21 cases respectively (half of the observations in the original groups). The discriminant weights were determined for both subsamples, and applied to the other half, and the number of primary and secondary correct predictions were computed, giving the number of correct classifications as well as the number correct including false positives.

## Results

*Analysis (a): the 7 qualitative and 4 quantitative predictors analyzed on interval level*

On the basis of the distribution of the canonical correlations, two optimal linear combinations of the source predictors and the group variables were considered. The two largest canonical correlations are .873 and .613. (the remaining two are .406 and .232). The apparent error rate of prediction was computed by using resubstitution in the following way: a case was re-assigned to a group by selecting the smallest distance from the case point to each of the *k* group points. The re-assignment results are given in the upper part of Table 3.

**Table 3.** Primary Prediction Based on Resubstitution (Apparent Error Rate)

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total |
| Analysis (a): 7 qualitative and 4 quantitative predictors (interval level) | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 37 | 7 | 0 | 0 | 50 |
| Histological | 3 | 0 | 13 | 31 | 5 | 1 | 50 |
| | 4 | 0 | 1 | 7 | 29 | 13 | 50 |
| | 5 | 0 | 1 | 0 | 7 | 34 | 42 |
| % Correct | | .82 | .74 | .62 | .58 | .81 | .71 |
| + false positives | | 1.00 | .90 | .74 | .84 | .81 | .86 |
| Analysis (b): 7 qualitative (ordinal level) and 4 quantitative predictors (interval level) | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 4 | 29 | 16 | 1 | 0 | 50 |
| Histological | 3 | 0 | 12 | 31 | 6 | 1 | 50 |
| | 4 | 0 | 1 | 10 | 30 | 9 | 50 |
| | 5 | 0 | 0 | 1 | 9 | 32 | 42 |
| % Correct | | .82 | .58 | .62 | .60 | .76 | .67 |
| + false positives | | 1.00 | .92 | .76 | .78 | .76 | .85 |
| Analysis (c): 7 qualitative, 4 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 31 | 10 | 2 | 1 | 50 |
| Histological | 3 | 0 | 14 | 30 | 5 | 1 | 50 |
| | 4 | 0 | 0 | 2 | 36 | 12 | 50 |
| | 5 | 0 | 1 | 1 | 6 | 34 | 42 |
| % Correct | | .82 | .62 | .60 | .72 | .81 | .71 |
| + false positives | | 1.00 | .88 | .72 | .96 | .81 | .88 |
| Analysis (d): 7 qualitative, 2 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 5 | 31 | 8 | 3 | 3 | 50 |
| Histological | 3 | 0 | 13 | 30 | 5 | 2 | 50 |
| | 4 | 0 | 0 | 2 | 38 | 10 | 50 |
| | 5 | 0 | 1 | 0 | 5 | 36 | 42 |
| % Correct | | .82 | .62 | .60 | .76 | .86 | .73 |
| + false positives | | 1.00 | .90 | .74 | .96 | .86 | .89 |

The error rate of prediction is largest for group 4 (only 58% is predicted correctly); the histological diagnoses of the cases in group 1 and 5 is predicted quite well: 82% and 81% respectively. The percentages correct including the false positives are also given in Table 3. This statistic shows that the result for group 4 improves considerably (84%), since most of the misclassified cases were assigned to group 5. On the basis of this finding it might be important to scrutinize the patients concerned again, since apparently they have features in common with the cases in group 5. The results for the secondary prediction into a joint dysplasia group and a joint carcinoma group is given in the upper part of Table 4. Although the analysis was not focused on the prediction into two groups, the percentages predicted correctly (87% and 91%) are quite satisfactory.

**Table 4.** Secondary Prediction Based on Resubstitution

| | | Predicted | | |
|---|---|---|---|---|
| | | 1-2-3 | 4-5 | Total |
| | | Analysis (a) | | |
| Histological | 1-2-3 | 131 | 19 | 150 |
| | 4-5 | 8 | 84 | 92 |
| % Correct | | .87 | .91 | .89 |
| | | Analysis (c) | | |
| Histological | 1-2-3 | 136 | 14 | 150 |
| | 4-5 | 3 | 89 | 92 |
| % Correct | | .91 | .97 | .93 |
| | | Analysis (d) | | |
| Histological | 1-2-3 | 133 | 17 | 150 |
| | 4-5 | 2 | 90 | 92 |
| % Correct | | .89 | .98 | .92 |

*Analysis (b): Optimal scaling of the 7 qualitative predictors with the 4 quantitative predictors analyzed on interval level*

Since 7 of the variables analyzed involve subjective ratings on a 4 point scale, the data have been re-analyzed in the following way. While fitting the parameters in the discriminant model, the categories of the 7 qualitative variables are subjected to an optimal scaling. Optimal scaling finds new scale values for the categories, and implies that the variables may

be nonlinearly transformed to maximize the sum of the canonical correlations; because the order of the categories is known, the scale values should maintain the original order. In the same analysis, the quantitative variables can be treated on an interval level.

Such a discriminant analysis can be performed by using the OVERALS program (Gifi 1990; implemented in SPSS 1990), which is a program for nonlinear generalized canonical correlation analysis. Canonical discriminant analysis is a special case of canonical correlation analysis: the canonical correlations (the correlations between the axes of the predictor space with the axes of the group space) are the same. To perform canonical discriminant analysis an ordinary canonical correlation analysis can be performed, provided that the axes of the group space are properly rescaled afterwards; the axes of the predictor space are also rescaled, so that the centroid principle is preserved, while the distinction between the group means is maximized.

The canonical correlations of this analysis are .884 and .700. Their sum (1.584) is larger than the sum from the previous analysis (1.486), but the improvement is not very substantial. Of utmost importance in this analysis is the transformation of the qualitative variables, obtained by the optimal scaling of the original scale values. These are shown in Figure 2: the original values are given at the horizontal axes, and the optimally scaled values on the vertical axes. An isolated dot at the coordinates (0, 0.0) indicates that the category 0 contains no observations for this variable.

The transformation of *Nuclear Shape, Chromatin Pattern, Chromatin Distribution, Nucleus/Nucleolus Ratio* and *Nucleus/Cytoplasm Ratio* distinguishes (0 and) 1 from 2 and 3. The categories 0 and 1 obtain scale values below the average (which is 0.0), and the categories 2 and 3 obtain scale values above the average. The transformation of *Nucleolar Irregularity* is different, since it contrasts the categories 0, 1 and 2 (below average) with category 3. The transformation of *Nuclear Irregularity* is dominated by the extreme scale value for category 1, which is caused by the fact that this category contains only one case. Due to this extreme value, the other scale values look quite similar; had they been plotted on

larger scale, however, the distinction between 1 versus 2 and 3 would be much more visible.

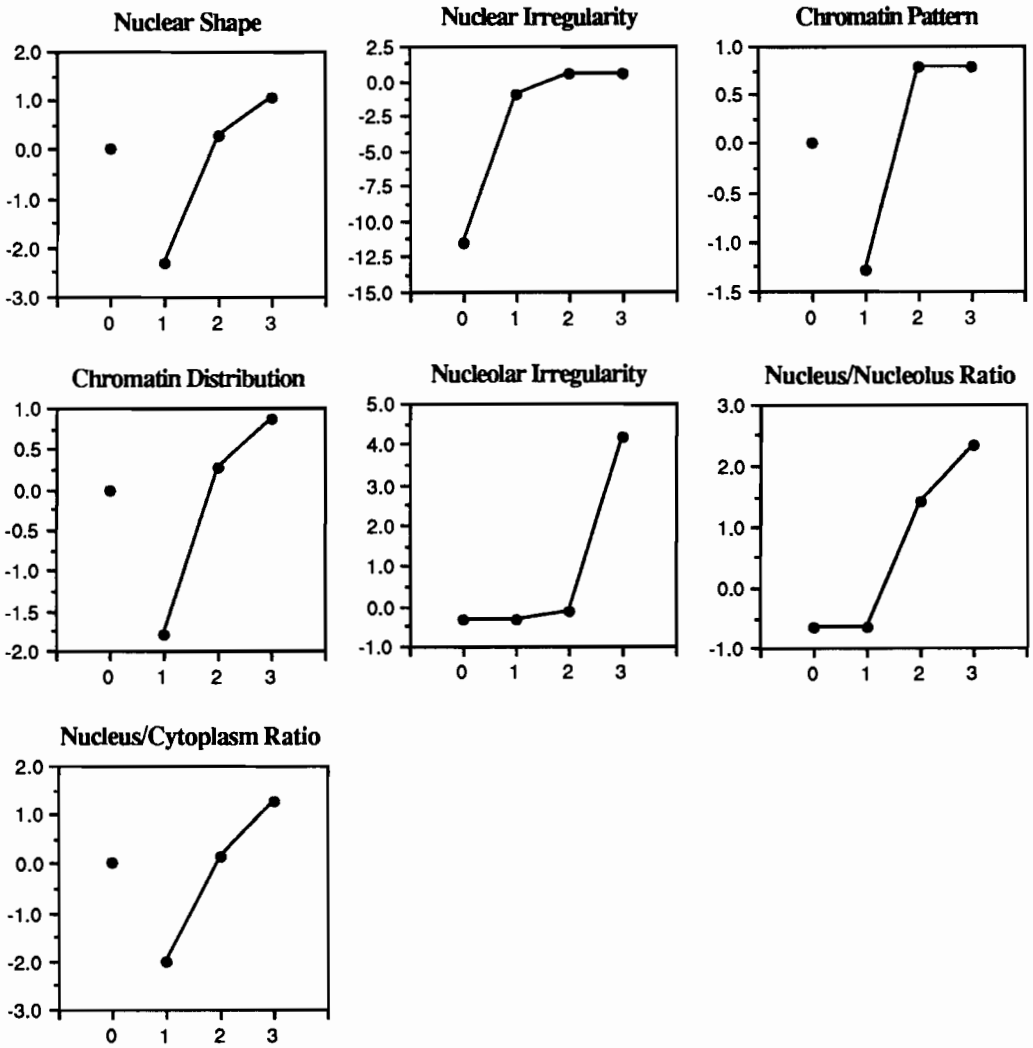


**Figure 2.** Monotonic transformations of the 7 qualitiative variables

It was remarked above that the optimal scaling of the qualitative variables shows a small improvement with respect to the sum of the first two canonical correlations. A similar improvement, however, is not obtained with respect to the percentages of cases that are predicted correctly, which is a different criterion; this is seen in the second part of Table 3.

The percentages of correct prediction for the groups 1 and 3 are identical to the results fo the first analysis, prediction for class 4 is only slightly better, and prediction for the groups 2 and 5 deteriorates. Therefore it was decided to continue the exploration of the data on the basis of the original source predictors.

*Analysis (c): 7 qualitative, 4 quantitative predictors and a threshold variable*

In the two previous analyses, the coordinates for a case in the predictor space are the sums of the (differentially weighted) predictors. In this sum a low score on one predictor variable may be compensated by a high score on another predictor. The present data contain information for which such a compensatory scheme might not be desirable; specifically, this concerns the variables *Number of Mitoses* and *Number of Cells per Fragment.* Therefore, in the third analysis the set of source predictors has been extended with a so-called *Threshold* variable, created in the following way. For *Number of Mitoses* and *Number of Cells per Fragment* a threshold was established (of 1 and 50.00 respectively). The threshold variable obtains the value 0 for those cases that score below the pre-chosen threshold for both mentioned variables, and has the value 1 for all other cases. The canonical correlations of the analysis are .900 and .616, so the sum is only a fraction larger than in the first analysis. The results for the primary prediction into 5 groups are given in the third part of Table 3.

Compared to the first analysis, the overall correct prediction is identical (71%), but there is a substantial increase in the correct prediction of group 4 (from 58% to 72%), and a decrease for group 2 (from 74% to 62%). When the false positives are included, the overall percentage is somewhat better (88%). Inspecting the secondary prediction into 2 groups (see Table 4), reveals an increase in correct prediction of both the dysplasia cases (from 87% to 91%) and the carcinoma cases (from 91% to 97%), so it is concluded that the *Threshold* variable contributes positively to the overall correct prediction (this conclusion will be scrutinized in the cross-validation study). Three cases that have the histological classification 4 or 5 were re-assigned according to the discriminant analysis into the joint dysplasia class

1-2-3. Interestingly enough, these three cases scored below both thresholds of *Number of Mitoses* and *Number of Cells per Fragment.*

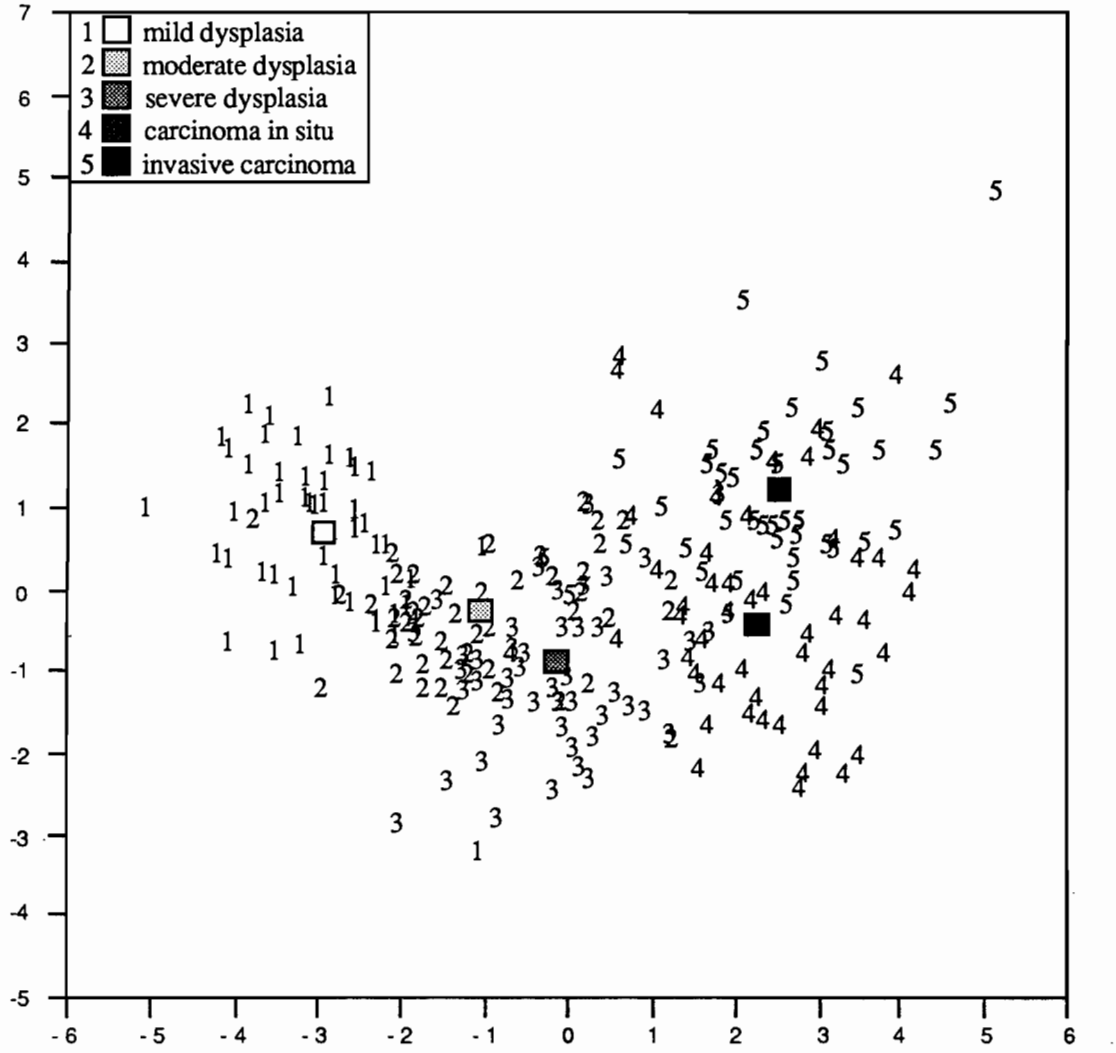


**Figure 3.** Two-dimensional space for cervical dysplasia and carcinoma cases: Individual points and group points

In Figure 3 the cases are displayed in the predictor space, labeled with their group number. Although the clouds of points for the different groups are clearly not perfectly separated, an explicit structure is established along the horizontal axis (the first dimension), going from the left hand side, with a majority of 1's, to the right hand side, with 4's and

5's, while the 2's and 3's are in the middle. The second dimension (the vertical axis) contributes to the discrimination, separating the 5's from the 4's. This structure is evident when the group points (the centroids of the individual cases) are inspected in Figure 3.

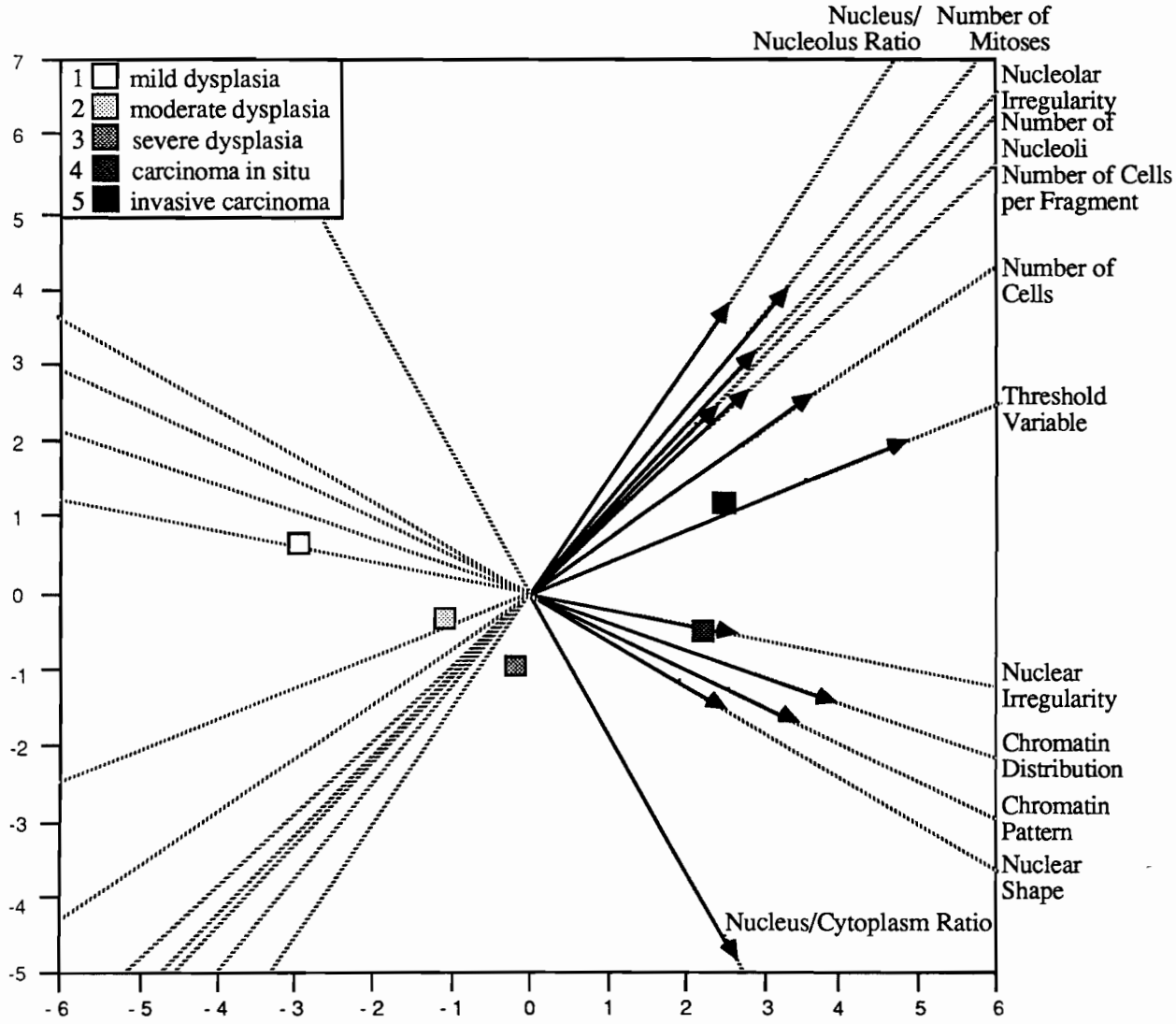


**Figure 4.** Two-dimensional solution for cervical dysplasia and carcinoma cases: Group points and predictor vectors

To investigate the structure of the set of predictors, they are depicted as vectors in the space in Figure 4, together with the group centroids (the Figures 3 and 4 are presented separately for the sake of clarity, but they show the very same space). The coordinates for

the vectors have been computed from 12 multiple regressions, using the two dimensional coordinates for the individual cases as independent variables and each predictor separately in the role of dependent variable. The orthogonal projection of the individual points onto the variable vectors gives an approximation of the data; the length of the vectors is proportional to their goodness-of-fit, which is the correlation of the approximation and the original variable. By using the centroids the relative position of the groups with respect to the predictor variables can be inspected.

The variables seem to consist of three subsets; the goodness-of-fit values are given in parentheses. Subset 1 contains *Nucleus/Nucleolus Ratio* (.73), *Number of Mitoses* (.83), *Nucleolar Irregularity* (.68), *Number of Nucleoli in 50 Cells* (.57), *Number of Cells per Fragment* (.63), *Total Number of Cells* (.72) and the *Threshold* variable (.84), ordering the groups from 5 (high values) to 1 (small values). This subset distinguishes predominantly between 4 and 5 on one side, and 1, 2 and 3 on the other. The second subset, consisting of *Nuclear Irregularity* (.40), *Chromatin Distribution* (.68), *Chromatin Pattern* (.61) and *Nuclear Shape* (.46), reverses the order of the groups 4 and 5, and gives a more clear distinction, compared to the first subset, between the groups 1, 2 and 3. The third subset, finally, consists of a single variable, *Nucleus/Cytoplasm Ratio* (.86), that clearly gives a completely different ordering, which is 4-3-5-2-1. This order is in perfect concordance with the group means of the original variable given in Table 1.

*Analysis (d): 7 qualitative, 2 quantitative predictors and the threshold variable*

The role of the threshold variable was put to a further test by omitting the original variables *Number of Mitoses* and *Number of Cells per Fragment* from the analysis, channeling their influence only through the binary *Threshold* variable. The canonical correlations are .898 and .570. Since the *Threshold* variable classifies the cases into two groups, it was expected that the present analysis could not improve upon the primary prediction into 5 groups, but would improve the secondary prediction into two groups; the results are given in lower parts of the Tables 3 and 4.

Surprisingly, compared to the previous analysis, the present one gives identical percentages correct for the groups 1 (82%), 2 (62%) and 3 (60%), but an improvement for group 4 (from 72% to 76%) and 5 (from 81% to 86%); the overall percentage improves from 71% to 73%. For the secondary prediction, however, the overall percentage decreases slightly from 93% to 92%, primarily due to more false positives.

*Monte Carlo Double Cross Validation*

As was explained in the introduction, double-cross validation can be repeated a large number of times; in the present study the data were split 10.000 times, so in total there are 20.000 predictions of the 5 groups for each of the analyses (a), (c), and (d). Each cross-validation sample consists of $n$ = 121 cases, and Figure 5 gives in a single diagram the distributions of the number of correct primary predictions out of these 121 cases; in Figure 6 the number of correct predictions includes the false positives. Both diagrams give a clear distinction between analysis (a), with only the 7 qualitative and 4 quantitative predictors, and the analyses (c) and (d) that include the threshold variable, where the latter two outrank the first mentioned. The cross-validation also confirms the slight superiority of the analysis (d), with the 2 source predictors omitted, over (c), with the 2 source predictors included. Figure 7 displays the distributions for the correct secondary prediction. With respect to analysis (a), results conform with the primary prediction results. Analysis (c), however, gives better results than (d). These observations are even more evident in the diagrams with the cumulative number of correct predictions in the Figure 8 (primary prediction into 5 classes) and 9 (secondary prediction into 2 classes). The modes of the distributions of the number of correct predictions are given in Table 5.

The accumulated re-assignment results are given in the Tables 6 and 7; the total number of observations in each cell was divided by 10.000, and rounded to the nearest integer. In this way the numbers can be easily compared to those in the Tables 3 and 4.

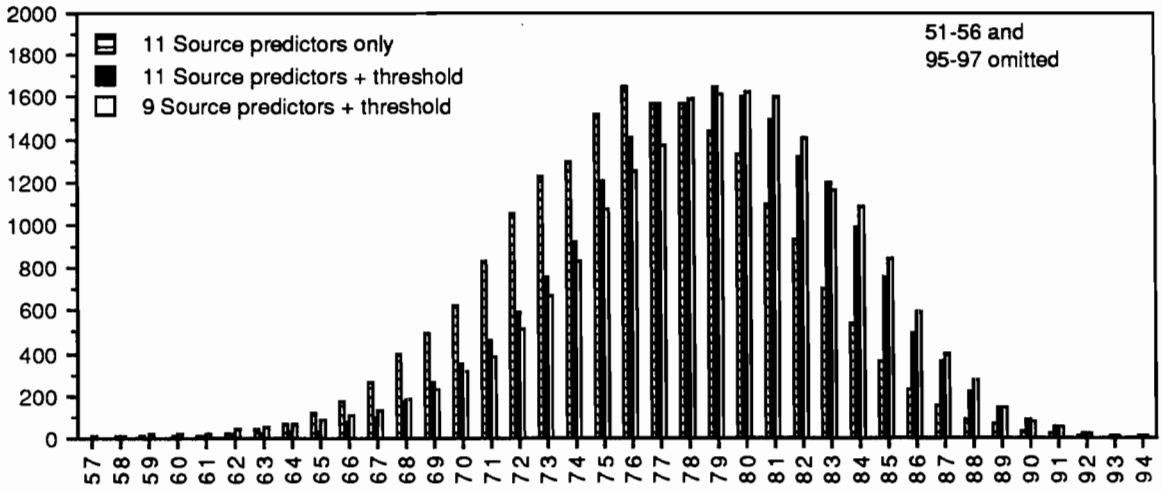


**Figure 5.** Distribution of number of cases predicted correctly (out of 121) in 20.000 cross-validation samples: primary prediction into 5 groups

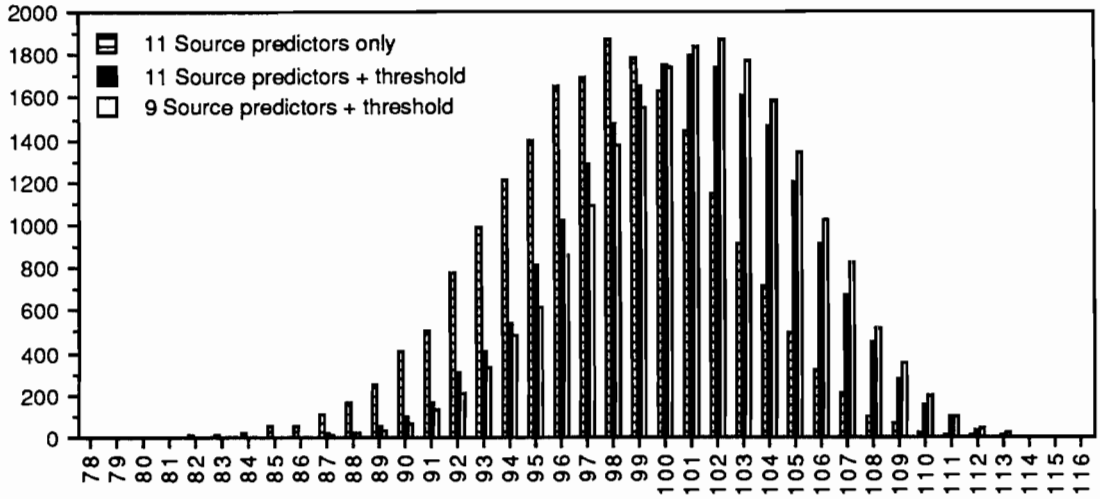


**Figure 6.** Distribution of number of cases predicted correctly plus false negatives

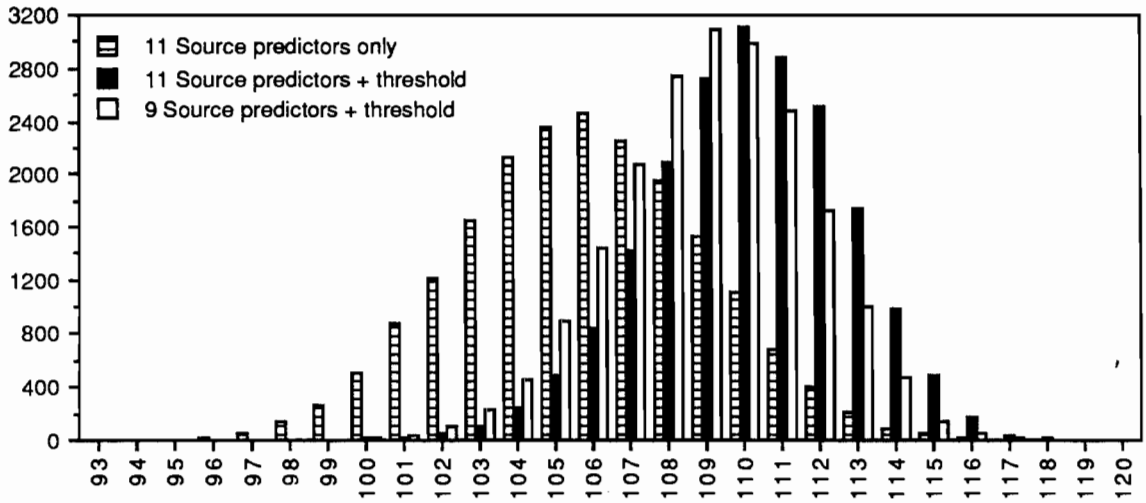


**Figure 7.** Distribution of number of cases predicted correctly: classification into 2 groups

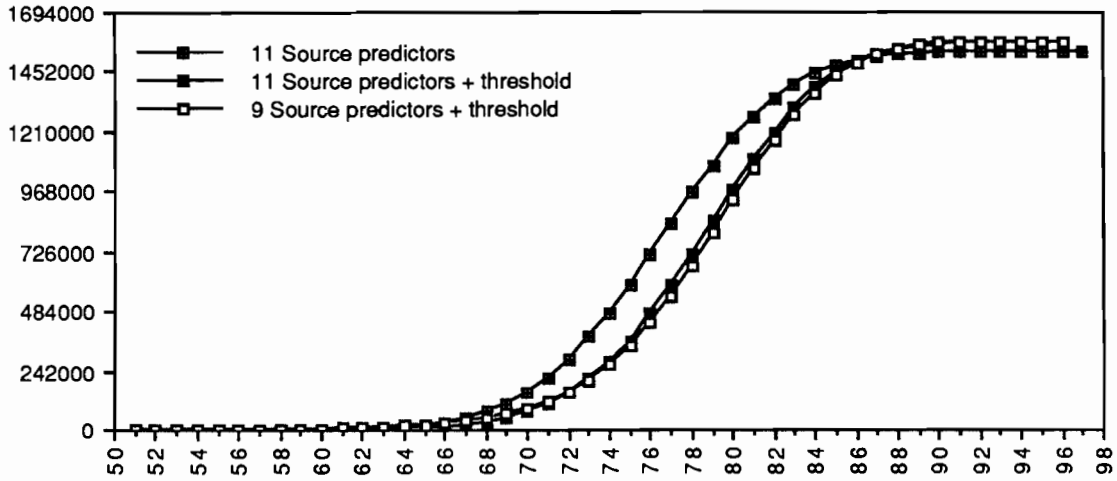


**Figure 8.** Cumulative number of cases predicted correctly (maximum is 121x20.000) in 20.000 cross-validation samples: primary prediction into 5 groups

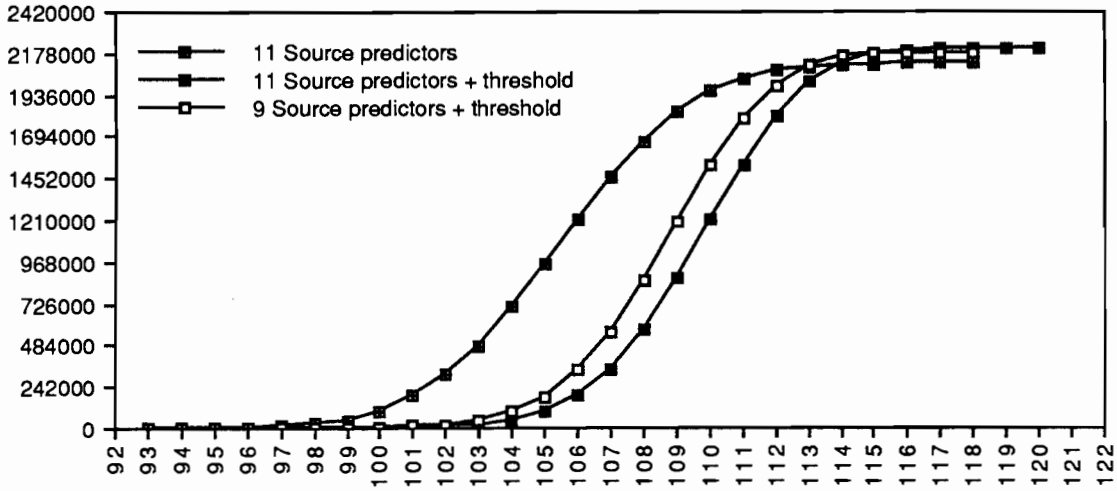


**Figure 9.** Cumulative number of cases predicted correctly (maximum is 121x20.000) in 20.000 cross-validation samples: secondary prediction into 2 groups

**Table 5.** Modes of the Distributions of Percentage Correct in Cross Validation

| Analysis | (a) | (c) | (d) |
|---|---|---|---|
| % Correct (classification into 5 groups) | .63 | .65 | .66 |
| % Correct including false positives | .81 | .83 | .85 |
| % Correct (classification into 2 groups) | .88 | .91 | .90 |
| % Correct including false positives | .96 | .98 | .98 |

**Table 6.** Primary Prediction Based on Monte Carlo Double Cross Validation; Average over 10.000 Samples Rounded to Nearest Integer (Expected Actual Error Rate)

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total |
| Analysis (a): 7 qualitative and 4 quantitative predictors | | | | | | | |
| | 1 | 39 | 10 | 1 | 0 | 0 | 50 |
| | 2 | 5 | 33 | 10 | 1 | 0 | 50 |
| Histological | 3 | 0 | 15 | 27 | 7 | 1 | 50 |
| | 4 | 0 | 2 | 10 | 26 | 12 | 50 |
| | 5 | 0 | 2 | 2 | 10 | 28 | 42 |
| % Correct | | .78 | .66 | .54 | .52 | .67 | .63 |
| + false positives | | 1.00 | .88 | .70 | .76 | .67 | .81 |
| Analysis (c): 7 qualitative, 4 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 29 | 10 | 3 | 1 | 50 |
| Histological | 3 | 0 | 16 | 27 | 6 | 2 | 50 |
| | 4 | 0 | 1 | 4 | 32 | 13 | 50 |
| | 5 | 0 | 1 | 1 | 11 | 28 | 42 |
| % Correct | | .82 | .58 | .54 | .64 | .67 | .65 |
| + false positives | | 1.00 | .86 | .70 | .90 | .67 | .83 |
| Analysis (d): 7 qualitative, 2 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 41 | 8 | 1 | 0 | 0 | 50 |
| | 2 | 7 | 28 | 10 | 2 | 3 | 50 |
| Histological | 3 | 0 | 16 | 26 | 6 | 3 | 50 |
| | 4 | 0 | 1 | 4 | 33 | 12 | 50 |
| | 5 | 0 | 1 | 1 | 10 | 29 | 42 |
| % Correct | | .82 | .56 | .52 | .66 | .69 | .65 |
| + false positives | | 1.00 | .86 | .70 | .90 | .69 | .84 |

**Table 7.** Secondary Prediction Based on Double Cross Validation

| | | Predicted | | |
|---|---|---|---|---|
| | | 1-2-3 | 4-5 | Total |
| Analysis (a) | | | | |
| Histological | 1-2-3 | 130 | 20 | 150 |
| | 4-5 | 11 | 81 | 92 |
| % Correct | | .87 | .88 | .87 |
| Analysis (c) | | | | |
| Histological | 1-2-3 | 133 | 17 | 150 |
| | 4-5 | 5 | 87 | 92 |
| % Correct | | .89 | .95 | .91 |
| Analysis (d) | | | | |
| Histological | 1-2-3 | 130 | 20 | 150 |
| | 4-5 | 4 | 88 | 92 |
| % Correct | | .87 | .96 | .90 |

To show that the Monte Carlo double cross validation gives a somewhat pessimistic impression, the error rates of the use of the different subsets of predictors was also estimated by the leaving-one-out method, the results of which are given in Table 8. It is clear that the latter method gives a slightly more optimistic overall result for each of the three possible analyses.

**Table 8.** Primary Prediction Based on Leaving-One-Out Method (Expected Actual Error Rate)

| | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | Total |
| Analysis (a): 7 qualitative and 4 quantitative predictors | | | | | | | |
| | 1 | 40 | 9 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 34 | 9 | 1 | 0 | 50 |
| Histological | 3 | 0 | 15 | 29 | 5 | 1 | 50 |
| | 4 | 0 | 1 | 9 | 27 | 13 | 50 |
| | 5 | 0 | 3 | 2 | 8 | 29 | 42 |
| % Correct | | .80 | .68 | .58 | .54 | .69 | .66 |
| + false positives | | 1.00 | .88 | .70 | .80 | .69 | .82 |
| Analysis (c): 7 qualitative, 4 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 40 | 9 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 30 | 9 | 3 | 2 | 50 |
| Histological | 3 | 0 | 16 | 27 | 6 | 1 | 50 |
| | 4 | 0 | 0 | 5 | 32 | 13 | 50 |
| | 5 | 0 | 1 | 1 | 9 | 31 | 42 |
| % Correct | | .80 | .60 | .54 | .64 | .74 | .66 |
| + false positives | | 1.00 | .88 | .68 | .90 | .74 | .84 |
| Analysis (d): 7 qualitative, 2 quantitative predictors + threshold variable | | | | | | | |
| | 1 | 40 | 9 | 1 | 0 | 0 | 50 |
| | 2 | 6 | 29 | 8 | 2 | 5 | 50 |
| Histological | 3 | 0 | 16 | 27 | 5 | 2 | 50 |
| | 4 | 0 | 0 | 2 | 37 | 11 | 50 |
| | 5 | 0 | 1 | 1 | 7 | 33 | 42 |
| % Correct | | .80 | .58 | .54 | .74 | .79 | .69 |
| + false positives | | 1.00 | .88 | .68 | .96 | .79 | .86 |

## Discussion

First, it was tested whether a linear discriminant analysis results in a reasonably good prediction into the 5 histological groups on the basis of 4 quantitative variables and 7 qualitative variables, which were ratings on a 4 point scale. The results demonstrate that the prediction is quite satisfactory, particularly for the dysplasias (see Table 3); in total, out of 242 cases, the histological diagnosis of 71% of the cases is predicted correctly. When afterwards only a distinction is made between groups of dysplasias on the one hand and carcinoma in situ and invasive carcinoma on the other hand, the percentage of correct prediction is 87% and 91% respectively (Table 4). In a second analysis, the subjective, qualitative criteria were optimally scaled; on the basis of the established error rates (Table 3), it was decided to continue the exploration of the data with all variables treated on an interval level.

The number of mitotic figures gives an indication of the mitotic activity. In the dysplasias they are rather rare, but in carcinoma in situ and invasive carcinoma they can be found in the histological section throughout the epithelium (Burghardt, 1970). In cytologic smears the mitotic figures are difficult to see in the thick epithelial fragments in which they are mostly situated (Boon and Tabbers-Boumeester 1980), but in the plastic section they are easy to detect. The number of cells per epithelial fragment gives an indication of the cohesiveness of the abnormal cells. We experienced that in smears made from Cytobrushes these fragments are large in cases of carcinoma in situ and invasive carcinoma, and small in cases of dysplasias. Thus for both variables, there seems to be a division line between the group of dysplasias on the one hand and the carcinoma group on the other hand (see also Table 2). Therefore we tested whether threshold values for these two features could be used. It was decided to transform the original variables *Number of Mitoses* and *Number of Cells per Fragment,* into a two-category *threshold* variable, cases scoring 0 (when below both thresholds) or 1 (above one of the two or both). The percentage of cases predicted correctly in the carcinoma in situ group increased (from 58% to 72%), but for the moderate

dysplasia group there were found more false positives (Table 3). These findings might indicate that some dysplasias have many mitotic figures, and/or are very cohesive. When the original variables *Number of Mitoses* and *Number of Cells per Fragment* are excluded from the analysis, channeling their influence through the threshold variable only, the total percentage predicted correctly is 73% for the 5 groups, and 89% for contrasting the dysplasia groups with the carcinoma groups.

To compare these results with the routinely made predictions on the basis of the cervical smear, they are paralleled in Table 9 (note that the results are expressed here in percentages only, in contrast to the other Tables, because the number of cases in the routine prediction is much larger than used for the current study). First the predictions on basis of the cervical smear are compared to a secondary prediction from analysis (d) that includes the threshold variable and the 7 qualitative and 2 quantitative predictors, calculating a joint group point for the slight and moderate dysplasia cases, and re-assigning all cases. They are also compared to an analysis with the same variables, but now with optimal discrimination into 4 groups (primary prediction). In the histological group 1-2 correct prediction increased from 43% to 80%; in group 3 the increase is from 42 to 78%, in group 4 from 61% to 78%, and in group 5 from 65% to 86%.

**Table 9.** Comparison of Cervical Smear Results and Discriminant Analyses

| | | Predicted | | | |
|---|---|---|---|---|---|
| | | 1-2 | 3 | 4 | 5 |
| Cervical Smear | | | | | |
| Histological | 1-2 | .43 | .24 | .16 | .01 |
| | 3 | .23 | .42 | .25 | .03 |
| | 4 | .08 | .18 | .61 | .08 |
| | 5 | .00 | .04 | .32 | .65 |
| Analysis (d): Secondary prediction into 4 groups | | | | | |
| Histological | 1-2 | .75 | .18 | .03 | .04 |
| | 3 | .04 | .82 | .10 | .04 |
| | 4 | .00 | .06 | .76 | .20 |
| | 5 | .00 | .02 | .12 | .86 |
| Primary prediction into 4 groups | | | | | |
| Histological | 1-2 | .80 | .13 | .02 | .05 |
| | 3 | .08 | .78 | .10 | .04 |
| | 4 | .00 | .06 | .78 | .16 |
| | 5 | .02 | .00 | .12 | .86 |

Because the resubstitution method might give a too optimistic estimate of percentage correctly predicted, a Monte Carlo approach to double cross-validation has been included to compare the analyses with respect to the expected actual error rate of prediction. This was done both for using the original 11 source predictors as well as using the threshold variable (both including and omitting the original 2 source predictors). The main results were confirmed in the cross validation (Table 5). In the latter the correct prediction rate is, as can be expected, slightly less but still far better than on basis of the evaluation of the cervical smears. The superiority of the analyses using the threshold variable is confirmed, not only for prediction into two groups (dysplasias versus carcinomas), but also into 5 groups.

Two points clearly emerge from this study. Firstly, if the visualization of cell features in epithelial fragment sampled by the Cytobrush is optimized (by using thin slicing of plastic blocks) a canonical discriminant analysis enables us to distinguish not only between the dysplasia and the carcinoma cases, but also between the three grades of dysplasia. The percentage of correct prediction outnumbers that obtained in the routine smear technique. Secondly, both qualitative and quantitative variables contribute to this prediction; the qualitative variables *Nuclear Shape, Chromatin Pattern, Chromatin Distribution , Nuclear Irregularity, and Nucleus/Cytoplasm Ratio* contribute especially to the discrimination between the 3 dysplasia groups. The variables *Number of Mitoses* and *Number of Cells per Fragment* are very powerful when transformed into a threshold variable; the experience presented in this paper can be used in diagnostic practice.

It seems clinically not a disadvantage that a classification results in a few cases of dysplasia placed in the carcinoma group: we even can argue that this subgroup of dysplasia should be treated as carcinomas (in situ) because they have important features characteristic for carcinoma.

**Acknowledgement**

The authors would like to thank Dr. Inne Susanti (Bali, Indonesia) and Dr. J.H. Lunardhi (Surabaya, Indonesia) for their cooperation in collecting the data.

## References

Boon ME, Bosch MMC (1990) Het einde van de Pap-classificatie in zicht ? [The end of Pap-classification in sight ?] *Nederlands Tijdschr Geneesk* 134: 1016-1017 (in Dutch)

Boon ME, Drijver JS (1986) *Routine cytological staining techniques: Theoretical background and practice*. Mac-Millan Education, London

Boon ME, Tabbers-Boumeester ML (1980) *Gynaecological Cythology; A textbook and atlas*. Mac-Millan Press Ltd, London

Boon ME, Zeppa P, Ouwerkerk-Noordam E, Kok LP (1990) Exploiting the toothpick effect of the cytobrush by plastic embedding of cervical samples. *Acta Cytol* (in press)

Burghardt E (1970) Latest aspects of precancerous lesions in squamous and columnar epithelium of the cervix. *Int J Gynaecoll Obstet* 8:573-580

Efron B (1979) Bootstrap methods: another look at the Jackknife. *Annals of Statistics* 7:1-26

Efron B (1982) *The Jackknife, the Bootstrap, and other resampling plans*. CMBS-NSF regional conference series in applied mathematics, monograph 38. SIAM, Philadelphia

Gifi A (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester-New York

Glenthoj A, Bostofte E, Rank F (1986) Brush cytology from the uterine cervix. *Acta Obstet Gynaecol Scand* 65:689-691

Helmerhorst TJM, Dijkhuizen GH, Calame JJ, Kwikkel HJ, Stolk JG (1987) The accuracy of colposcopically directed biopsy in diagnosis of CIN. *Eur J Obstet Reprod Boil* 24:221-229

Hand D (1981) *Discrimination and Classification*. Wiley, Chichester-New York

Hirst D, Ford I, Critchley F (1990) An alternative approach to the estimation of error rates in discriminant analysis. In Sidák S, Eben K (eds) *Proceedings of DIANA III*. Czechoslovak Academy of Sciences, Prague (in press)

Kraemer BB (1990) Does the Bethesda System promote or endanger the quality of cervical cytology? *Acta Cytol* 34:458-459

Miller RG (1974) The Jackknife: a review. *Biometrika* 61: 1-15

Patten SF (1983) Morphologic subclassification of pre-invasive cervical neoplasia. In: Wied GL et al. (eds) *Compendium on Diagnostic Cytology,* Fifth ed. Tutorials of Cytology, Chicago, pp 108-117

Rosenthal DL, Suffin SC (1984) Predictive value of digitized cell images for the prognosis of cervical neoplasia. In Greenberg SD (ed) Computer-Assisted Image Analysis Cytology. In Wied GL (ed) *Monographs in Clinical Cytology,* Ninth volume. S Karger, Basel, pp 163-180

SPSS (1990) *CATEGORIES.* SPSS Inc, Chicago

Tatsuoka (1976) *Validation Studies.* Selected Topics in Advanced Statistics, Institute of Personality and Ability Testing, Champaign IL

Wheeler N, Suffin SC, Hall TL, Rosenthal DL (1987) Prediction of cervical neoplasia diagnosis groups. Discriminant analysis on digitized cell images. *Analyt Quant Cytol Histol* 9:169-181