Journal of Cellular Biochemistry, Supplement 23:33-42 (1995)

Cervical Cell Recognition and Morphometric Grading by Image Analysis

James W. Bacus, PhD

Bacus Research Laboratories, Inc., Elmhurst, IL 60126

Abstract Cervical cell recognition by morphometric image analysis was compared to human visual cell recognition on the same 6,375 cells from 40 dysplastic, CIS, invasive, and 10 normal pap smears. The experimental approach defined receiver operating characteristic (ROC) curves for morphometric image analysis which could be rigorously compared to previously established human visual cell recognition ROCs on the same cells. Overall performance was measured by A_z, the area under the ROC curves in the two instances. For morphometric image analysis cell recognition, $A_z = 0.91$, and for human visual cell recognition, $A_z = 0.87$. These results clearly demonstrated that morphometric image analysis is equivalent to experienced human observers in ability to recognize isolated cells from cervical smears.

An approach was also developed to link the ROC analytic methods of this study to a cytopathological or histopathological grading system, or "scale", that could be expressed in terms of normal deviate units of morphometric descriptors. This approach has the advantage of describing the grading scale in terms of its ROC characteristics; in essence, it describes performance for that grading scale at any decision point along the scale, if used for two-category classification. Additionally, this concept provides for a uniform final scale, regardless of which cells or tissues are graded. Also, this type of grading scale would automatically adjust itself for measurement variance for different types of cells or tissue, by reference to normal cells or tissues, so that a standard reference could be maintained.

Key words: Cervical cancer, morphometry, Pap smears, receiver operator characteristic (ROC) curves

One of the issues related to computerized morphometric image analysis of cells and tissues is determining performance criteria. This is important whether one is using this technology to classify cells or to grade cells along a scale of maturation or atypia. Since morphometric image analysis often replaces a human visual assessment, there is often a requirement to characterize the "performance" of the visual assessment, *e.g.*, the ability of the human observer to recognize malignant and normal cells or tissue, in a manner consistent with morphometric image analysis, and then to use the "human performance" as a benchmark standard for comparison purposes. In this regard, a previously reported study determined cervical cell discrimination ability by the human observer using conventional psychophysics and signal detection theory methodology, with a Gaussian signal detection model [1]. This study rigorously defined human observer receiver operating characteristic curves (ROCs), both for isolated individual malignant cell detection and for the visual assessment of the entire slide, *i.e.*, slide screening. This paper reports results of morphometric image analysis and cell classification, performed on the same cells used in the human recognition studies. Also, a method of morphometric grading, conceptually linked to multicategory cell classification and two category ROC analysis, is proposed. These results were obtained using a Gaussian classification model,

Address correspondence to James W. Bacus, PhD, Bacus Research Laboratories, Inc., 910 Riverside Drive, Unit 8A, Elmhurst, IL 60126.

which corresponded to the analysis assumptions used in the human recognition studies. Thus, morphometric image analysis ROC characteristics were compared in a directly analogous manner to previously obtained human performance ROCs to document the capability of computerized morphometric image analysis.

MATERIALS AND METHODS

Data Acquisition

Case material collection and cell acquisition for photomicroscopy has been described previously [1]. Briefly, 50 consultant-reviewed cases of normal, dysplastic, carcinoma *in situ* and invasive cervical cases were studied. To obtain the microscope slide screening ROC performance, these cases were randomly mixed with 1,147 routine cytology cases; all of the microscope slides were screened in the conventional manner by 10 experienced cytotechnologists. The ROCs were determined for this process by using these results and other literature studies relating to slide screening, but with varying false positive and false negative decision criteria. Secondly, 6,375 individual cell photomicrographs were acquired in a stratified random sampling study design from these same specimens. These were used to obtain human cell recognition ROCs. The photomicrographs of each cell were obtained with and without background and associated surrounding cells in order to compare the effects of surrounding background on cell recognition. Thus the recognition studies involved the independent "blinded" use of both types of photomicrographs. One such pair of photomicrographs is shown in Figure 1.

A primary aim of these studies was to obtain descriptions of cell recognition independent of the arbitrary selection of decision thresholds. Detection capabilities were thus determined using ROC methodology from signal detection theory [2,3]. The area under the ROC curve, A_z, was chosen as the standard of comparison between individuals and between slide screening and cell recognition. This allowed a comparison of entire ROC curves instead of individual points. However, the point at which an individual could operate, *i.e.*, which performance specification in terms of the probability of false positives or false negatives is adhered to, could still be examined later to understand the relationships between individual cell recognition and slide screening.

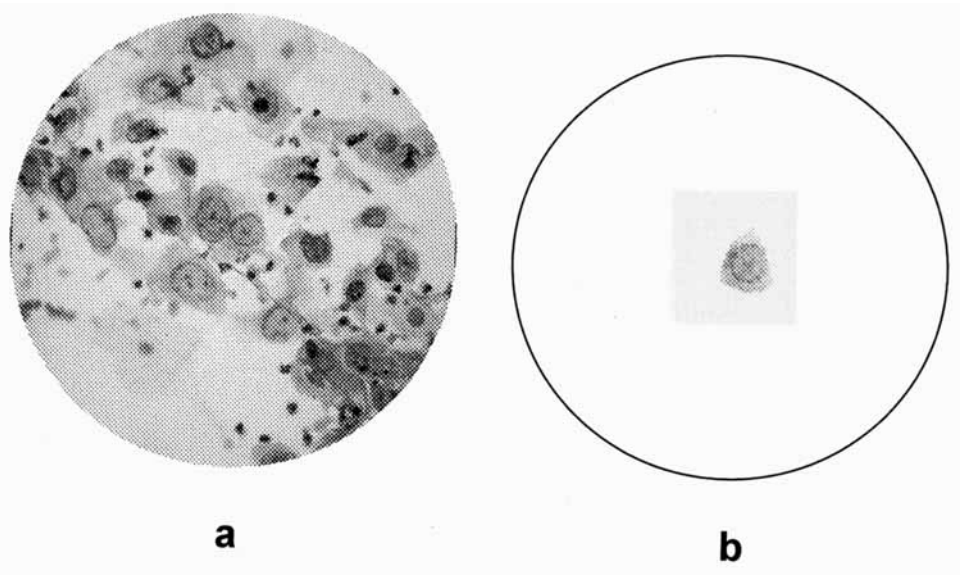


Fig. 1. Examples of photomicrographs used in the human observer visual cell recognition studies. Phomicrograph (a) included the entire microscope field of view, while photomicrograph (b) was electro-optically altered during the data acquisition process to blank-out the surrounding background, leaving only the cell of interest for the observer to classify.

The photomicrographed cells were also digitized at several different spectral wavelengths and resolutions using a cell acquisition system centered around a Leitz Orthoplan microscope. The optical path included a 100 x Plan Apo objective and a 1.25 x nosepiece. A standard 100 watt tungsten halogen lamp was used as a light source. The microscope was equipped with an electronic scanning stage capable of two dimensional translation over an area of 2 cm by 2 cm in 10 μm steps. Modifications to the microscope included a special stage clamp to enable most of the 24 mm by 50 mm specimen slides to be accessed by the scanning stage, a 6-position computer-controlled filter wheel mounted below the condenser, and replacement of the standard observation tube with a specialized optical system necessary to acquire the data. The filter wheel was configured with six filters to aquire digital images at several narrow bandpass wavelengths, to aquire photomicrographs, and for observation by cytotechnologists during data acquisition. A separate optical path from the objective to a solid state camera was used to obtain the digitized cellular images. A swing lens was used to obtain digital images at two resolutions, a low resolution of 80 μm by 80 μm mapped onto a 128 x 128 pixel digital image (*i.e.*, 0.625 μm between pixels), and a higher resolution of 16 μm x 16 μm (resulting in 0.125 μm between pixels). Four digitized images were obtained from each cell. Three images, 528 nm, 576 nm and 632 nm, each at a bandpass of 9 nm, were at low resolution, including cell and background, with different color filters chosen to match the Papanicolaou stain [4]; one image of the nucleus only was at high resolution with no color filter.

Morphometric Image Analysis

As indicated above, digital images were acquired at two resolutions and at different spectral wavelengths. A number of features were analyzed in an interactive fashion to determine a reasonable set. Table I indicates the set of features finally used. These features represented standard, state-of-the-art parameters for cell description, and there was ample precedence in the literature for their use [5,6]. They included standard measurements of area, density, color, shape and texture. As indicated above, we were particularly interested in employing a multivariate Gaussian classifier in the morphometric studies because of potential comparison to the Gaussian signal detection model and performance results of the psychophysical studies. One of the problems with this approach was that the features chosen, *i.e.*, those listed in Table I, were not all distributed in a Gaussian manner. In order to correct this, a number of normalizing scale transformations were tried [7]; the resulting distributions were tested for normality by the chi square goodness of fit test. The transformation with the lowest chi square value was selected for each feature. Table I also lists the normalizing transformations that were finally used.

For the multivariate Gaussian classification experiment, a normal versus abnormal classifier

TABLE I. Summary of Individual Cell Measurement

Cell Measurements	Resolution	Transformation
Nuclear area	low	$y = \ln(x)$
Nuclear/cytoplasm ratio	low	$y = [(x/(1-x)]$
Cytoplasm shape (perimenter^2/area)	low	$y = [(x/(1-x)]$
Average nuclear density at 528 nm	low	$y = [(x/(1-x)]$
Average nuclear density at 632 nm	low	none
Markovian angular second moment at 0.125 μm	high	$y = \arcsin(x)^{1/2}$
Markovian sum entropy at 0.375 μm	high	$y = \ln(x)$
Markovian maximum correlation coefficient at 1.375 μm	high	$y = [(x/(1-x)]$

was constructed for the eight dimensional transformed feature space. The mean vectors and covariance matrices of each group were obtained from training on one half of the data (the training set), chosen randomly. Several classification runs were then performed on the other half of the data (the testing set), each time varying the *a priori* probabilities in the multivariate Gaussian model. This is exactly analogous to a human observer changing decision criteria, or changing detection thresholds and rereading the entire data set in a psychophysical experiment. The computer could of course perform this simulation very rapidly and precisely to obtain multiple points defining its ROC curve, whereas it is often impractical to do this with human observers.

ROC Analytic Methods

As indicated above, one of the aims of these studies was to obtain measures of cell recognition free of judgemental bias, *i.e.*, to obtain measurements of recognition ability independent of the selection of decision criteria. In this type of cell recognition, the selection of decision criteria is related to judgments regarding the progression of normal maturation sequences or judgments related to the progression of atypia to carcinoma. One way of handling problems of decision criteria and judgemental bias in data analysis is the ROC analytic method [2]. Figure 2 illustrates some of the concepts of the ROC method. In the classical sense, the method is applied when considering two overlapping distributions such as shown in Figure 2a. The distribution on the left represents the results of measuring some characteristic of normal individuals, and the distribution on the right represents the results of measuring the same characteristic on abnormal individuals. The intent is to develop a decision rule based upon the measurements to characterize or distinguish normal from abnormal. In psychophysical signal detection theory, where the ROC method was first employed, the distribution on the left is considered noise and the one on the right is the distribution of possible signals. The abscissa is shown in standardized units scaled to zero mean and a standard deviation of one for the normal category. Typically the abnormal category has a higher mean value and a higher standard deviation than the "noise" or normal distribution. The decision rule is simply a criterion value on the abscissa, such that individuals with values above the criteria are abnormal and below it are normal. Clearly, overlapping distributions can cause false positive (FP) or false negative (FN) errors, and different decision rules will produce different FP and FN error rates. The correct responses in each case are called true positive (TP) and true negative (TN). As an example, the ROC curve shown in Figure 2b graphically depicts the complete set of FP and FN responses for the distributions shown in Figure 2a. The shape of the curve is governed by the distance between the respective means and the standard deviations of the signal distribution compared to the noise distribution. An increased separation of the means produces increased curvature towards lower FP and FN values; increased differences in the standard deviations increases the asymmetry of the ROC curve. In psychophysical studies, ROC curves are usually depicted as in Figure 2c and Figure 2d. Figure 2c is analogous to the FP versus FN curve of Figure 2b, where (1-FN), *i.e.*, TP, is plotted instead. This curve is often transformed to the binormal coordinate space for analysis purposes, shown in Figure 2d. The advantages of transforming to binormal coordinates is that the ROC is linear for normally distributed data. The straight line ROC has the parameters Δm for the x intercept and s for the slope. These parameters can be estimated for experimental data by fitting a straight line to plotted points. Also, the slope of the straight line ROC is the reciprocal of the standard deviation of the abnormal distribution, expressed in normal deviate units of the normal distribution. The intercept of the straight line ROC on the binormal abscissa is equal to the difference between the means of the distributions.

When comparing different ROC curves, it is convenient to use an index of performance, a single number to characterize how well the measurement variable can separate normal from abnormal. The index of performance commonly used in psychophysics experiments is A_Z, the area beneath the original TP, versus FP ROC curve before conversion to binormal coordinates. Figure 2c illustrates the meaning of the detection index A_z, its relation the original false positive (FP) and true positive (TP) coordinate space, and the transformed binormal coordinate space. Clearly, A_z ranges from 0.0 to 1.0, and an A_z of 0.5, *i.e.*, with the ROC curve lying on the major

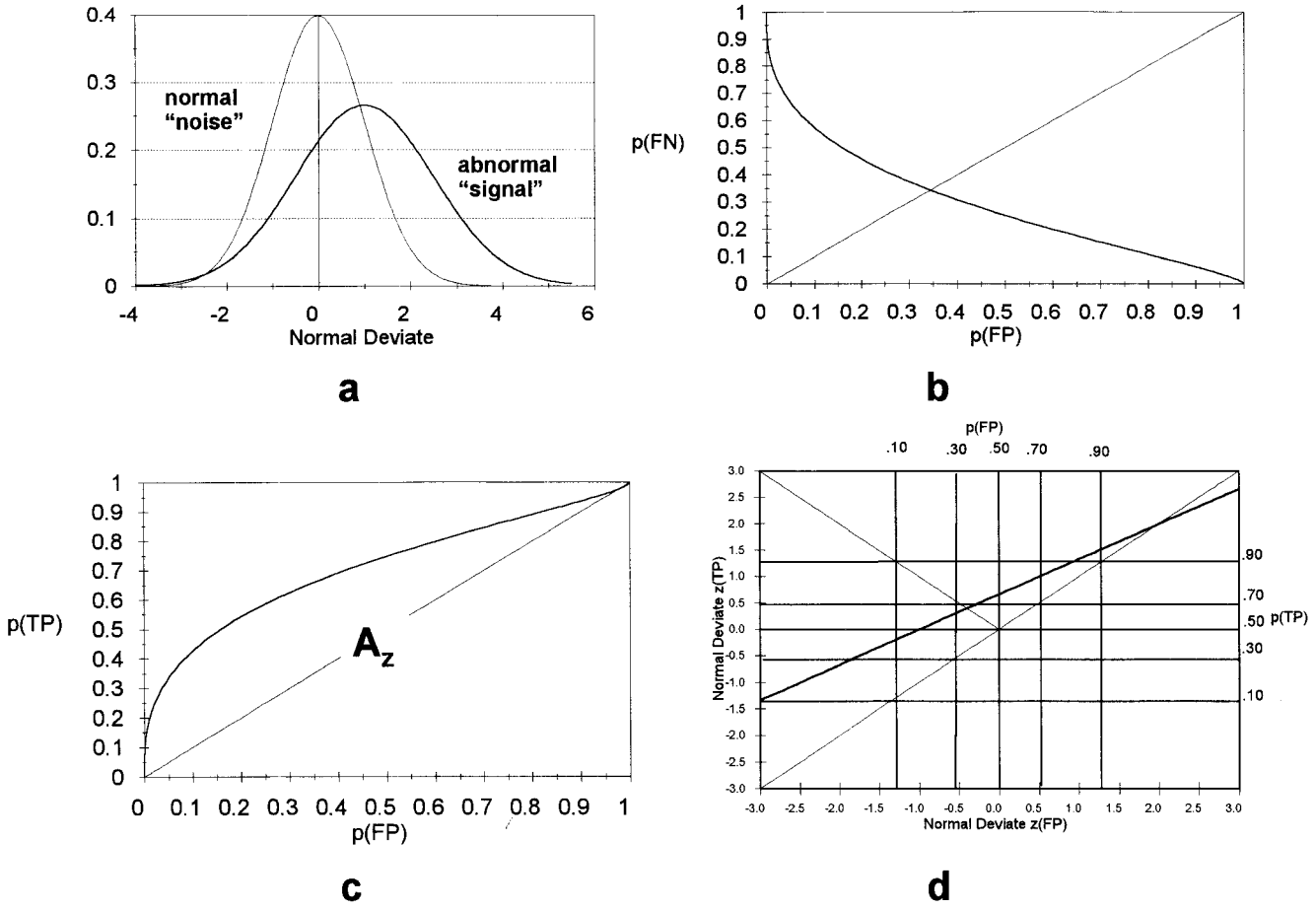


Fig. 2. (a) Measurement distributions, where the abscissa or normal deviate axis is normalized to zero mean and unit standard deviation using estimates derived from the normal or "noise" distribution, allowing the abnormal "signal" distribution to shift and scale by the mean and standard deviation estimates derived from the noise distribution. These distributions define the standard false negative (FN) and false positive (FP) ROC curve shown in (b), which is often used in medical studies. A single point on the ROC curve is generated from the distributions by selecting a "criterion", or cutoff point, along the normal deviate axis and plotting the area above that point from the noise distribution (the FP error), and below that on the signal distribution (the FN error), as a single point on the curve. The entire curve is generated by choosing different criterion points and plotting the resulting pairs of FN and FP errors. Similarly, the standard false negative (FN) and true positive (TP) ROC curve often used in psychophysical studies is shown in (c). The ROC curve from (c) replotted in binormal coordinates is shown in (d). The ROC curve is linear in this plot, and the standard deviation of the signal distribution compared to the unit standard deviation of the noise distribution is reflected in the slope of the linear ROC. A slope of less than 1.0 indicates a greater, equal to 1.0 the same, and greater than 1.0 a smaller standard deviation, respectively.

diagonal, represents chance behavior. An A_z of 1.0, *i.e.*, with the curve showing p(TP) = 1.0 and p(FP) = 0, represents perfect cell recognition. The value A_z is obtained by computing $z(A) = s(\Delta m/1 + s^2)$ and referring to the standard normal distribution. A_z is the area under the normal distribution up to the value of z(A). The experimental data points are fitted in the binormal coordinate space because of the linearity in that space; however, the computed performance index actually relates back to the original p(FP), p(TP) space.

In a strict sense, conventional ROC analysis is limited to two stimuli (or classification) alternatives, *e.g.*, normal versus abnormal for a cell type. Thus, one of the experimental problems in this study was to allow the observers to inspect and describe the visual material in the traditional classification form and then, through analysis procedures, transform their responses into a two-response form. This was accomplished by condensing the multicategory confusion matrix of the initial recorded classifications to obtain different "groupings" of cell categories [1]. Thus, multiple points along the ROC curve could be obtained from the same recorded data. This is analogous to using "rating method" analysis procedures in psychophysics experiments [2]. An alternative method of achieving multiple points would have been to resubmit the cell photomicrographs to multiple observations after instructing the observers to change their decision criteria relating to definitions of cell type progression indicated above. In plotting ROC data and subsequently computing performance indices, a normal probability distribution was assumed. This assumption of normality, and also the cell ranking actually used with the rating method, was generally validated in this case since when the ROC data was displayed on binormal coordinates, the experimental data points evidenced a close fit to a straight line [1].

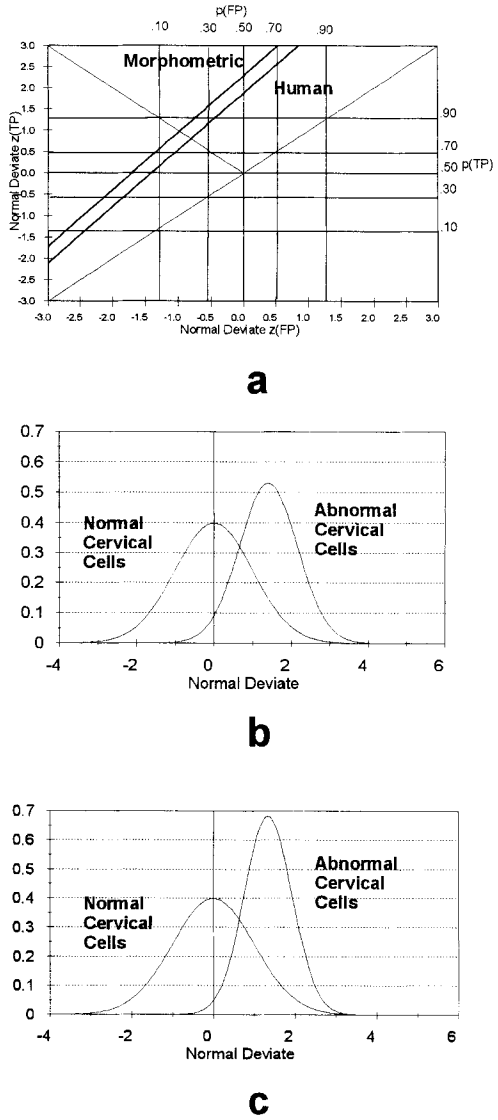
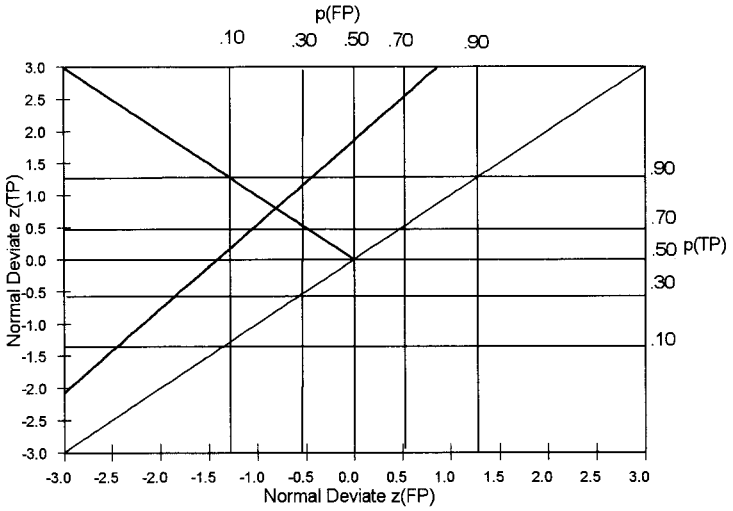


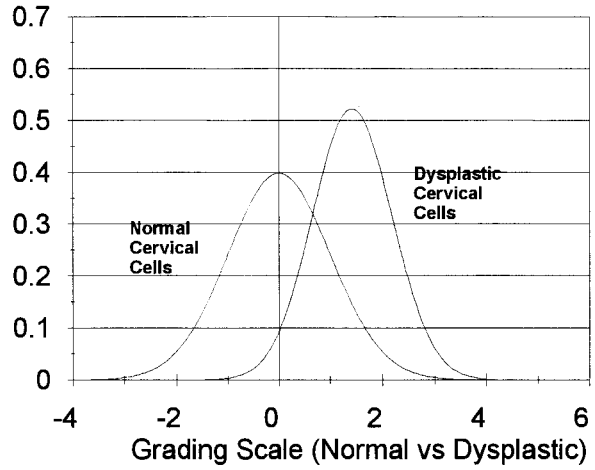
Fig. 3. (a) The composite binormal ROC for human visual cell recognition (lower line, A_z = 0.87) compared to that obtained on the same cells by morphometric image analysis (upper line, A_z = 0.91). The overlapping Gaussian distributions implied by the two ROC curves are shown in (b) and (c) for human visual cell recognition and morphometric image analysis respectively.

RESULTS

The ROC curve results, comparing the morphometric cell classifier to the human observer cell classifications on the same cells from the psychophysical experiment, are shown in Figure 3a. The lower curve is for human visual cell recognition, which is the composite ROC from 10 observers on 50 cases, Δm = 1.4, s = 1.33, and A_z = 0.87 [1]. The upper curve is from results of the computerized morphometric image analysis using the same features listed in Table I, on exactly the same cells. In this case, Δm = 1.71, s = 1.34 and A_z = 0.91. Figures 3b and 3c show the corresponding overlapping Gaussian distributions implied by the ROC curves. The study design enabled comparing many different



a



b

Fig. 4. (a) The composite (10 observers) binormal cell recognition ROC for normal versus dysplastic cervical cells (Δm = 1.31, s = 1.42, A_z = 0.86) [1]. (b) The implied overlapping distributions of normal versus dysplastic cervical cells, where the normal deviate is also considered as a morphometric grading scale.

classes individually with each other. Some of the results were compiled previously [1]. As an example, Figure 4 indicates the composite (for 10 observers) pairwise comparison of normal versus dysplastic cervical cells in a format similar to Figure 3. The linear ROC is shown in Figure 4a; Δm = 1.31, s = 1.42 and A_z = 0.86. The implied distributions are shown in Figure 4b. In this instance, the normal deviate scale is labeled as a "grading scale" for reasons discussed below.

DISCUSSION

As indicated in Figure 3a, these results clearly demonstrate that morphometric image analysis is equivalent to experienced human observers with regard to the ability to recognize isolated cells from cervical smears. This level of ROC performance is similar to that of other difficult medical visual inspection tasks [2]. Methods to link this level of ROC performance to the overall higher level of pap smear slide screening performance were addressed previously [1], and it was shown that this individual cell recognition ability is sufficient to achieve slide screening with the number of cells available for inspection. It was also shown that immediately surrounding background cells did not increase the recognition capability, *i.e.*, cells with and without background had similar ROCs. It is important that individual cell recognition capability by morphometric image analysis, and equivalence to experienced human observer capability, is documented in a definitive way for at least one organ system, since morphometric image analysis is under serious consideration as a generalized surrogate endpoint biomarker (SEB) in many organ systems [8]. These specific ROC curves provide a documented standard of potentially achievable performance for other algorithm development with regard to pap smear screening systems using image morphometry.

Classification data in cell or tissue image analysis by image morphometry is typically tabulated in a "confusion matrix" table. The rating method, which is commonly used in psychophysical studies, is a unique way of analyzing confusion matrix data, where categories in the matrix are organized to relate to maturity (or atypia) transitions from one type of cell to another. This is often the case for cellular or tissue material, since the cells, or lesions, develop from one category to the next along a continuum. A "trained" observer should have the basic discriminatory ability to recognize subtle cues in mapping the transition from one stage to another, but may place the cells or image structures into arbitrary categories. Thus, the rating method can condense the multicategory confusion matrix into multiple p(FP) and p(FN) points, defining the analytic goals (the ROC curve) for the task. This method of analysis provides a better description of discriminatory ability than others commonly used, such as a single set of p(FP) and p(FN) points, or other indices of performance such as the *kappa* statistic. Other indices have been reviewed and compared to each other by Swets [3].

These results also suggest an approach to a cytopathological or histopathological grading system, or "scale", that could be expressed in terms of normal deviate units of morphometric descriptors. It is common practice in multivariate statistical analysis to transform individual measurements to normal deviate units by subtracting the mean and dividing by the standard deviation for each measurement axis. This transformation simply shifts each measurement scale to zero mean and unit standard deviation using all of the data, and still preserves individual group or category differences. This transforms the data into "standard deviation units" and is sometimes referred to as the z variate, or z score. It is particularly helpful if the different measurement scales occur in units with widely differing absolute values because it tends to put the different scales of the multivariate space in the same numerical range, but preserves group differences. Also, statistical techniques such as linear discriminate analysis "project" multivariate measurements onto a one dimensional decision axis to make decisions that occur with a two category ROC analysis, such as normal versus abnormal. If the above-described transformation to normal deviate units is applied using the mean and standard deviation of the normal category only, and applied to all measurements prior to projection onto a one-dimensional axis, the result after projection would be a uniform morphometric scale expressed in normal deviate units of the "noise" or normal category, similar to that shown in Figure 2a. This suggestion is more strictly correct with uncorrelated variables. However, in a more detailed implementation, the principle compo-

nents of the distributions would be accounted for, in order to account for possible correlation among the descriptive measurements. Sometimes this is automatically accomplished with the method of projection on a discriminant axis. Gaussian transformations, if necessary, *e.g.*, as used in this study, should be performed before this "grading transformation."

A limitation of this proposed grading scale is an implied assumption of similar correlation between normal and abnormal categories which might not hold. However, a very positive advantage is that of relating ROC curve methodology to "grading scale" methodology; quite often there is reasonably similar correlation between the same measurements. This would allow further characterization of scales for different cells or tissue types by A_z and Δm, thus precisely describing the grading scale in terms of its ROC characteristics, and in essence describing performance for that grading scale at any decision point along the scale, if it were used for two category classification. This concept also provides for a uniform final scale, regardless of which cells or tissues are graded. Many grading schemes are reported in the literature, partly because of increased capabilities in computerized morphometric image analysis; there is currently no correspondence between grading for different cell or tissue types. For example, no common scales or units of measurement link cervical cell nuclear grading with breast cancer nuclear grading, and there is no effective way to compare them to each other. This method would help with this through standardization of scales and the link to ROC descriptions of performance as described above. Finally, and very importantly, this type of grading scale would also incorporate the concept of measurement variance by reference to normal cells or tissues into the concept of a grading scale. The scale would automatically adjust for different cell types, for example cervical versus breast, which might have different variances for individual normal categories under presumably different measurements employed. This would be an aid to interpretation since grading numbers, *e.g.*, + 1.3, *etc.*, would already be adjusted for biological variation of the normal reference cell population.

The overlap between normal and abnormal shown in Figure 3c for the morphometric image analysis is similar to the overlap in Figure 3b for the human observers on the same cells, indicating that this may be close to the best performance that can be obtained for morphometric cell classification. As an example of the above discussion, both in terms of overall cell recognition ROC performance and grading, if grading and not classification performance is of interest, Figure 4b might define a grading scale for transitions from normal to dysplastic for cervical cells. The high degree of overlap between distinct categories, and the implied high FP and FN error rates, is only important if the cell recognition task is artificially constrained to a classification task. If the cell recognition task is redefined as a morphometric grading task, where the aim is to place individual cells on a continuum, then a high degree of categorical overlap is not a problem. In this case, an individual cell could have a grading score of + 1.5 on this scale and be completely defined by its 8 morphometric measurements, placing it reliably in the progression from normal to atypia rather than in the overlap region of high FP and FN errors. Thus, fixed categorization cell recognition tasks by the older visual descriptive techniques of microscopy, which evidence low performance ROC characteristics, and where there is a biological rationale for subtle progression between the previously defined visual categories, may actually be considered "ideal" for quantitative morphometric grading, since the redefined scale can actually define a maturation or atypia sequence quantified in terms of the morphometric measurements. In summary, these visual inspection tasks could be considered as graded maturation situations instead of multicategory classification situations, where by definition the cell is at the stage indicated by the scale. This would eliminate the implication of high error rates.

The methods of study design and analysis shown here may provide a useful paradigm for future SEB studies. For example, if a morphometric image analysis is developed for nuclear grade in breast cancer as a SEB, the discriminatory ability of the pathologist in determining normal, atypia and premalignant nuclear grades would be important. Likewise, in evaluating premalignant lesions in esophageal cancer, the ability of the trained observer to detect and grade the continuum of developing patterns would be a reference point for morphometric image analysis of the histological tissue structure.

REFERENCES

1. Bacus JW, Wiley EL, Galbraith W, Marshall PN, Wilbanks GD, Weinstein RS: Malignant cell detection and cervical cancer screening. Anal Quant Cytol Histol 6:121–130, 1984.
2. Swets JA: Measuring the accuracy of diagnostic systems. Science 240:1285–1293, 1988.
3. Swets JA: Form of empirical ROCs in discrimination and diagnostic tasks: Implication for theory and measurement of performance. Psychol Bulletin 99:181–198, 1986.
4. Galbraith W, Marshall PN, Lee ES, Bacus JW: Studies on Papanicolaou staining I. Visible-light spectra of stained cervical cells. Anal Quant Cytol Histol 1:160–168, 1979.
5. Pressman NJ: Markovian analysis of cervical cell images. J Histochem Cytochem 24:138–144, 1976.
6. Bacus JW: Automation of morphology in hematology. In Schmidt RM (ed): "CRC Handbook Series in Clinical Laboratory Science, Section I: Hematology." Vol. II. New York: CRC Press Inc., 1980, pp 79–89.
7. Natrella MG: Experimental Statistics, National Bureau of Standards Handbook 91, Chapter 20-I Normality and Normalizing Transformations, 1966, pp 20-1 to 20-3.
8. Boone, CW, Kelloff GJ, Freedman LS: Intraepithelial and postinvasive neoplasia as a stochastic continuum of clonal evolution, and its relationship to mechanisms of chemopreventive drug action. J Cell Biochem 17G (Suppl):14–25, 1993.