Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gynecologic Oncology 99 (2005) S24 – S31

Gynecologic Oncology

www.elsevier.com/locate/ygyno

ELSEVIER

Original Research Report

Classification using the cumulative log-odds in the quantitative pathologic diagnosis of adenocarcinoma of the cervix

Richard J. Swartz[a], Loyd A. West[b,1], Iouri Boiko[c], Anais Malpica[c], Martial Guillaud[d], Calum MacAulay[d], Michele Follen[e,f,g,*], E. Neely Atkinson[h], Dennis D. Cox[i]

[a]*Department of Behavioral Science, The University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd - Unit 243, Houston, TX 77098, USA*
[b]*Department of Obstetrics and Gynecology, Naval Medical Center Portsmouth, 620 John Paul Jones Circle, Portsmouth, VA 23708, USA*
[c]*Department of Pathology, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. Box 85, Houston, TX 77030, USA*
[d]*Department of Cancer Imaging, British Columbia Cancer Research Center, 601 West 10th Avenue, Vancouver, British Columbia, Canada V5Z 1L3*
[e]*Biomedical Engineering Center, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. Box 193, Houston, TX 77030, USA*
[f]*Department of Gynecologic Oncology, Center for Biomedical Engineering, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA*
[g]*Department of Obstetrics, Gynecology and Reproductive Sciences, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[h]*Department of Biostatistics and Applied Math, University of Texas M. D. Anderson Cancer Center, 1515 Holcombe Blvd. Box 193, Houston, TX 77030, USA*
[i]*Department of Statistics, Rice University, P.O. Box 1892, Houston, TX 77005, USA*

Available online 26 September 2005

Abstract

Introduction. This study develops a method that discriminates between normal and cancerous tissue sections (i.e., populations of cells) using a statistical model applied to high-dimensional quantitative measurements made on a sample of cells.

Materials and methods. We use a cumulative log-odds model to create a score for a tissue section using the information from the cells within that tissue section. Then, a threshold is determined using receiver operating characteristic (ROC) curve analysis. The method was tested using data from cervical adenocarcinomas, adenocarcinoma in situ, and normal columnar tissue.

Results. Using 120 potential features, we analyzed the data for staining-independent features. Twenty-two features were statistically significant. We then calculated the log-odds and created a score, followed by ROC curve analysis. The operating point which maximizes the sum of the specificity and sensitivity achieved a sensitivity of 100% with a specificity of 85%.

Conclusion. The cumulative log-odds performs well in classifying tissue sections using high-dimensional data measured at the cellular level, like that of quantitative pathology. This methodology potentially has applications in pathology, radiology, and optical technologies.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Logistic regression; Population classification; Receiver operating characteristic (ROC) curve analysis; Adenocarcinoma; Cervical carcinoma; Quantitative pathology; Cumulative odds

Introduction

We consider a statistical problem motivated by the following medical diagnostic problem: How can one classify a tissue section (a slice of a tissue sample) as cancerous or not, given high-dimensional measurements on cells sampled from the tissue section? This problem arises in many diagnostic situations in clinical medicine. The problem considered in this report is from quantitative pathology (QP) methods that generate multidimensional data sets and use new technology and software developed

[*] Corresponding author. The University of Texas M. D. Anderson Cancer Center, Center for Biomedical Engineering, Unit 193, 1515 Holcombe Blvd., Houston, TX 77030, USA. Fax: +1 713 792 4856.
E-mail address: mfollen@mdanderson.org (M. Follen).

[1] The work described in this article was performed while LCDR Loyd A. West, MC, USNR was a fellow at M.D. Anderson Cancer Center training in Gynecologic Oncology. The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

0090-8258/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygyno.2005.07.038

to aid physicians and pathologists in diagnosing cancer [1–3].

A method of statistical analysis that is well suited to thinking about probabilities is Bayesian analysis. "Bayesian" refers to the Reverend Thomas Bayes, known for his solution to the problem of inverse probability. We are accustomed to thinking about chance and the probability that an event will occur. The odds ratio divides the probability an event will occur (in the numerator) by the probability it will not occur (in the denominator). Bayes' formula involves the concept of the *conditional probability*, meaning a probability that is updated when additional information is available. For instance, given a test, how likely is the patient to have disease or given the disease, how likely is the patient to test positive. While the mathematics may need refreshing, the concepts of conditional probability are used daily by clinicians. In the framework of diagnostic testing, Bayes formula can be used to express the post-test probability (also known as *posterior probability*) of disease in terms of the pre-test probability (also known as *prior probability*) of disease and another quantity called the *Likelihood Ratio* (LR), which depends on the outcome of the test. It is convenient to use the odds rather than the probability to express Bayes Theorem: (Post-test odds) = (Pre-test odds) × (Likelihood ratio). This equation is known as the odds ratio form of Bayes Theorem (see Eq. (4.11) of [4]). It is convenient to apply logarithms to this formula, which turn the multiplication into addition: log (Post-test odds) = log (Pre-test odds) + log (Likelihood ratio). Now the log-odds ratio is commonly called the *logit* and is precisely the quantity that is modeled in logistic regression.[2]

To further motivate these mathematical transformations, consider that probabilities must lie between 0 and 1, and a probability >1/2 means an event is more likely to occur than not, with the opposite for a probability <1/2. A probability >1/2 corresponds to an odds ratio between 1 and infinity, whereas a probability <1/2 corresponds to an odds ratio between 0 and 1. There is, thus, an asymmetry in the estimation of the strength of the odds ratio. A mathematical transformation that corrects this asymmetry is the log-odds: a probability >1/2 corresponds to log-odds between 0 and infinity, and a probability <1/2 corresponds to log-odds between negative infinity and 0. The range of the log-odds from negative infinity to positive infinity is necessary when building linear logistic regression models, as any such model will give both negative and positive numbers. We will generally transform the odds ratio by taking its *natural* logarithm [log base *e*], but it is immaterial whether one uses natural logarithms or base-10 logarithms. Finally, the *cumulative log-odds* is the sum of the log-odds, and we will find that very useful in the research reported here. In this analysis of quantitative pathologic data, we find the log-odds that each cell is cancerous using a logistic regression model. The cumulative log-odds are then used to predict if the tissue section is cancerous. Finally, we calculate the sensitivity and specificity (the assuredness) with which we make the diagnosis.

Materials and methods

Quantitative pathology measurements are made on tissue sections using the stoichiometric thionin–Feulgen stain, for which the amount of stain increases linearly with the amount of DNA in the nucleus. Quantitative pathology methods quantitatively analyze characteristics of nuclei (such as size, shape, DNA content, etc.) using digital-image-analysis systems. Specifically, quantitative pathology techniques take high-dimensional measurements on each cell within a tissue section, then seek to classify the tissue section based on the cellular level measurements [1]. These quantitative pathology measurements facilitate objective and reproducible interpretation of cell characteristics that previously have been subjectively assessed with classical visual diagnostic procedures. Multilevel data such as the data from a quantitative pathology analysis are becoming more common in the medical setting with the development of new diagnostic technologies. Information is gathered on one level, such as the cellular level, when the goal is to classify on a higher level, such as tissue sections.

In this section, we explain the data collection procedures and statistical methods. This includes data cleaning as well as why certain nuclear features were omitted from the analysis. Finally, we explain how our model uses cell level information to classify tissue sections.

Sample selection

The histopathologic database at M.D. Anderson was used to identify all cases of adenocarcinoma in situ. All cases were identified and the tissue blocks were pulled. A subset of invasive cervical adenocarcinomas and normal cases were selected and matched by age of the specimen. Only one block was selected from each patient to avoid repeated measures.

[2] Log-odds or logit: They contain exactly the same information as odds ratios. Log-odds are measures of the strength of relationship between variables. Because they are symmetrical, as explained above, log-odds can be compared more easily. A positive log-odds means the independent variable has the effect of increasing the odds that the dependent variable equals a given value (e.g., usually 1 for binary dependents). A negative log-odds means the independent variable has the effect of decreasing the odds that the dependent variable equals the given value. Log-odds do not make immediate intuitive sense because we are not used to thinking in terms of odds, odds ratios, and natural logs. If we want to know what a log-odds of a relationship actually means, we have to re-transform the log-odds back into an odds ratio, which is a measure of effect size, as discussed above. That is, we raise the natural log *e* to an exponent equal to the log-odds, and this equals the odds ratio.

Image analysis

Image analysis was performed using the CytoSavant, an image analysis system that has been commercialized and licensed by Oncometrics (Vancouver, BC, Canada). This system includes a 12-bit double correlated sampling Micro-Imager 1400 digital camera (pixels 6.8 μm^2). This software was specially designed for semi-automatic analysis tissue sections [1–3]. Thionin–Feulgen stained nuclei were measured with a monochromatic light at a wavelength of 600 nm, using a 20× *0.75 N.A Plan Apo objective lens. With a printout of the diagnostic area on hand, a cytotechnologist locates the exact same area on the Feulgen-stained slide as on the Hematoxylin and Eosin-stained slide. This stain is used clinically on all pathology sections; whereas the Feulgen stain is used only for quantitative analysis. The pathologist and cytotechnologist outline, with a mouse, the basal membrane and the superficial membrane or the top and the bottom of the epithelium (Fig. 1A). These two membranes define the Region of Interest (ROI), or Sampling Window. Automatic detection of the nuclei is then performed for further architectural analyses (Fig. 1B). This procedure is fully automated and requires only some minor manual changes. At a high magnification (20×), the nucleus is segmented within the ROI. A nuclear segmentation algorithm has been described in detail elsewhere [2]. Briefly, a thresholding algorithm is used to separate the objects (nuclei) from the background, based on pixel intensity (Fig. 1C). A manual correction of the nuclear segmentation is made to touching objects (Fig. 1D). Auto-focusing and edge-relocation algorithms are finally applied to the nuclei to precisely and automatically place the edge of the object at the region of highest local gray-level gradient. The digital gray-level images of these nuclei are stored in a gallery (Fig. 1D) [1–3].

Lymphocyte collection

Lymphocytes are used as a way of normalizing the amount of DNA collected from the cells in the tissue sections. Between 10 and 100 lymphocyte nuclei are collected from the underlying stromal compartment in the vicinity of the region of interest or in the adjacent fields if required. The same steps described above in the image analysis section for the epithelial nuclei are applied to the lymphocyte collection.

Quality control

The cytotechnologist manually reviews each object in an image gallery of all the selected cells (Fig. 1D) and removes any object which does not fulfill the minimum requirements. Examples of the reason for removing objects include bad mask, out of focus, pale nucleus, pycnotic nucleus, etc. Special attention is given to the lymphocytes in order to obtain a homogeneous population: only dark, dense, round objects are accepted. Reviews of the quality control process,

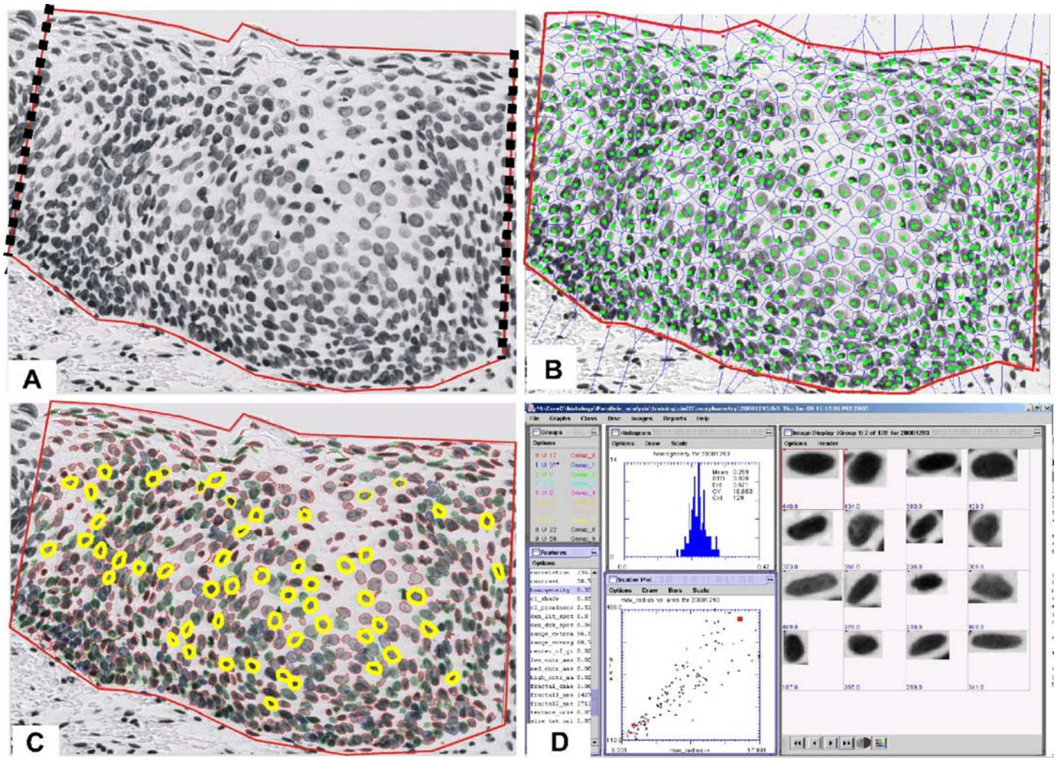

Fig. 1. Steps of the semi-automated analysis of cervical lesions.

results, and a description of the software used, are published [5–7].

Feature calculation

Nuclear features are extracted from the digitized nuclear images of each selected cell. Table 1 gives the list of the features organized into different categories; approximately 120 features are calculated [7]. Morphological features describe the nuclear size, shape, and boundary irregularities. The eight photometric features estimate the absolute intensity, optical density levels of the nucleus, and the intensity distribution characteristics. DNA amount is the raw measurement of the Integrated Optical Density (IOD) from which all the photometric features are derived. The IOD norm is the mean value of the DNA amount of the reference population. The DNA Index is the normalized measure of the integrated optical density of the object, i.e., DNA amount divided by IOD norm, meaning divided by the amount of DNA in the lymphocytes.

Discrete texture features are based on thresholded segmentation of the object into regions of low, medium, and high optical density. The thresholds are scaled to the sample staining intensity as represented by the IOD norm value determined from the reference population. Details of algorithms are described elsewhere [5–7]. Discrete texture features are by definition dependent on the normalization. Markovian texture features characterize gray-level correlation between adjacent pixels in the image. Non-Markovian texture features describe the features in terms of the maxima and minima of gray-level differences in the object. Fractal texture features describe the texture using local differences integrated over the object at multiple dimensions. Run-length texture features describe chromatin distribution in terms of the length of consecutive pixels with the same compressed gray level value along different orientations (0°, 45°, 90°, and 135°). In order to make the run-length features rotationally invariant, we used only the mean and standard deviation of the collection of angles run.

Table 1
Quantitative pathology categories and features

Category	Features
	Cytometric features
Discrete texture (20)	Low, medium, and high DNA amount Low, medium, and high DNA area Low, medium, high, and medium–high DNA compactness Low, medium, high, and medium–high DNA average distance Low, medium, and high density object Low, medium, and high center mass Low vs. medium, low vs. high, and low vs. medium–high DNA
Markovian texture (7)	Entropy, energy, contrast, correlation, homogeneity, cl-shade, and cl-prominence
Non-Markovian texture(5)	Density light spots, density dark spots, center of gravity, range-extreme, and range-average
Fractal texture (3)	Fractal area1, fractal area2, fractal dimension
Run length texture (20)	Short runs-mean, Short run-stdv, Short run-min, and Short run-max Long runs-mean, Long run-stdv, Long run-min, and Long run-max Gray-level-mean, Gray-level-stdv, Gray-level-min, and Gray-level-max Run-length-mean, Run-length-stdv, Run-length-min, and Run-length-max Run-Percent-mean, Run-Percent-stdv, Run-Percent-min, and Run-Percent-max

Statistical analysis

In this analysis of quantitative pathologic data, we find the log-odds that each cell is cancerous from the cells in each tissue section. The cumulative log-odds are then summed and used to predict if the tissue section is cancerous. Finally, we calculate the sensitivity and specificity (the assuredness) with which we make the diagnosis.

Our goal is to classify diseased tissue sections from non-diseased tissue sections given the cellular measurements. In order to classify tissue sections, we will first model the log-odds of disease at the tissue-section level using the information at the cellular level. This model requires estimates of the log-odds of disease at the cellular level, which are obtained by applying logistic regression (at the cellular level). We develop the method assuming disease is binary, and we combine the AdCa and ACIS groups to make up one group of diseased cases, which are then compared to the normal cases. All the analyses were carried out in SAS version 8.2.

Modeling the log-odds of disease for tissue sections

Let D (disease) be a binary indicator of the true state of disease for the tissue section where $D = 0$ or 1 according to whether the tissue section is normal or diseased, respectively. Let $x_1, x_2, \ldots x_n$ be the feature vectors from each of the n cells from the tissue section. Let $f(x_i|D)$ = conditional probability density function given D evaluated at x_i, $1 \leq i \leq n$. We assume for this discussion that $f(x_i|D)$ is known, but we will later show how it is possible to estimate the likelihood ratios $f(x_i|D = 1)/f(x_i|D = 0)$ from available data.

We also assume that the cell feature vectors are conditionally independent and identically distributed given D. Then the likelihood ratio for a tissue section is the product of the likelihood ratios for the individual cells. Using Bayes' rule, we can write the posterior (or post-test) odds of disease as the product of the prior (pre-test) odds of disease and the likelihood ratio for the tissue section. If we take the log of both sides and use some algebraic manipulation, we can see that the post-test log-odds of disease for the tissue section can be calculated from the log-odds of disease for each of the cells sampled from the tissue section:

$$
\begin{aligned}
&\log\left(\frac{P(D=1|x_1,x_2,\ldots x_n)}{P(D=0|x_1,x_2,\ldots x_n)}\right)\\
&\quad=\log\left(\frac{P(D=1)}{P(D=0)}\times\frac{f(x_1,x_2,\ldots x_n|D=1)}{f(x_1,x_2,\ldots x_n|D=0)}\right)\\
&\quad=\log\left(\frac{P(D=1)}{P(D=0)}\right)\\
&\qquad+\log\left(\frac{f(x_1|D=1)\times f(x_2|D=1)\times\cdots\times f(x_n|D=1)}{f(x_1|D=0)\times f(x_2|D=0)\times\cdots\times f(x_n|D=0)}\right)\\
&\quad=\log\left(\frac{P(D=1)}{P(D=0)}\right)+\sum_{i=1}^{n_j}\log\left(\frac{f(x_i|D=1)}{f(x_i|D=0)}\right)\\
&\quad=\log\left(\frac{P(D=1)}{P(D=0)}\right)+\sum_{i=1}^{n_j}\left[\log\left(\frac{P(D=1|x_i)}{P(D=0|x_i)}\right)\right]\\
&\qquad-n\log\left(\frac{P(D=1)}{P(D=0)}\right),
\end{aligned}
\tag{1}
$$

where $\log\left(\frac{P(D\,=\,1)}{P(D\,=\,0)}\right)$ is the prior log-odds of disease and $\log\left(\frac{P(D\,=\,1|x_i)}{P(D\,=\,0|x_i)}\right)$ is the posterior log-odds of disease based on the feature measurement for the ith cell. Thus, the log-odds of disease at the tissue-section level is modeled as the sum of the posterior log-odds of disease for the cells within the tissue section with a correction term. The log-odds is useful for two reasons: it gives the relatively simple formula above and we can directly model the posterior log-odds of disease for each cell using logistic regression.

Modeling the prior log-odds of disease at the cellular level

The posterior log-odds of disease on the cellular level given the high-dimensional feature vectors is modeled using a logistic regression model. The posterior log-odds of an individual cell being from a diseased tissue section given the feature vector is expressed as $\log\left(\frac{P(D\,=\,1|x)}{P(D\,=\,0|x)}\right)=x^t b$ where x is a $p+1$ vector of feature measurements with a 1 concatenated for the intercept term (with p being number of features) and b is the $p+1$ by 1 column vector of coefficients.

This logistic regression model was fit to the available data. The stepwise variable selection option employed to reduce the number of quantitative pathology features used in the model is straightforward. The variable not already in the model with the highest chi-squared score statistic is entered into the model if it is significant at the 0.05 level. Once a variable is added, all variables in the model are tested using the Wald chi-square test. The variable with the smallest statistic that is not significant at the 0.05 level is removed. This is done until no other variables can be removed. This process of adding and removing variables continues until no other variables meet the entry criterion or the last variable added is the first removed. This procedure is available in SAS V8.2 using PROC LOGISTIC with the SELECTION = STEPWISE option defaults.

Logistic regression was employed for two reasons. First, it is straightforward to use commercial computer packages to obtain logistic regression estimates. Second, we could use a stepwise variable selection technique to further reduce the dimensionality of the problem. Also, since there were approximately 2.5 times as many diseased tissue sections as there were normal tissue sections, we balanced this by giving normal cells a weight of 2.5 and diseased cells (AdCa and ACIS) were given a weight of 1.

The last term $n\log\left(\frac{P(D\,=\,1)}{P(D\,=\,0)}\right)$ in Eq. (1) involving the prior log-odds of abnormality at the cell level is a correction factor to transform the cumulative posterior log-odds of the cells being diseased to the sum of the log of the likelihood ratios for each of the cells. Since we are using logistic regression, this prior log-odds is implicitly the log-odds based on the sample proportion of abnormal cells. Therefore, the correction factor is calculated from the proportion of cells in the sample from diseased tissue sections. Letting N denote the total number of cells in the sample used to fit the logistic regression, and N_D the number of cells from diseased tissue sections, our final expression for the estimated posterior log-odds of disease for a tissue section is

$$
\begin{aligned}
&\log\left(\frac{P(D=1|x_1,x_2,\ldots x_n)}{P(D=0|x_1,x_2,\ldots x_n)}\right)=\log\left(\frac{P(D=1)}{P(D=0)}\right)\\
&\quad+\sum_{i=1}^{n_j}x_i^t b-n\log\left(\frac{N_D}{N}\right).
\end{aligned}
\tag{2}
$$

The prior log-odds of disease at the tissue level

The priors for the tissue-section log-odds for disease can either be estimated from the data, or taken from values in the literature. However, the prior log-odds for disease at the tissue-section level enters the model as a single additive term. Omitting this term simply shifts all the scores by a constant. Since we use a threshold classifier based on the log-odds, it does not matter what we assign to the prior log-odds of disease at the tissue level. However, if one wants to change the log-odds to probabilities, then one would want the prior log-odds in the model, and we would suggest either prevalence estimates from the literature or eliciting the prior.

In summary, we use the estimates of the log-odds of disease at the cellular level to estimate the log-odds of disease at the tissue-section level. The logistic regression allows us to reduce the feature set further by using only features that are significantly related to the log-odds of disease at the cell level. We will use the log-odds of disease at the tissue-section level to classify the tissue section as diseased or not.

Application of the model algorithm to generate a sensitivity and specificity

We used a simple threshold model to classify a tissue section based on its estimated log-odds of disease. We optimized the threshold to maximize the sum of the sensitivity and specificity using ROC curve analysis as explained by Metz [8]. Basically, we generated an empirical ROC curve and picked the point on this curve that maximized the sum of the sensitivity and specificity. In order to reduce the bias of the error-rate estimates, we used a leave-one-out cross-validation procedure considering a tissue section as the unit of observation. We trained the logistic regression using $S-1$ tissue sections, where S is the total number of tissue sections, and then used the trained logistic regression model to estimate the log-odds of disease for the cells from the omitted section. This was done for each tissue section. To increase stability of the algorithm, we performed the stepwise variable selection process for each run of the leave-one-out cross-validation, and only those variables present in at least 75% of the runs were kept. Then, the process was repeated until all features left in the logistic regression model were selected over 75% of the time.

Results

Samples

The histopathological samples consist of 68 independent tissue sections of the uterine cervix, each collected from separate patients. Two pathologists (I.B. and A.M.) specializing in gynecology confirmed the diagnosis in blinded fashion. Thirteen biopsies had normal histology, 37 biopsies had adenocarcinoma in situ of the cervix (AIS), and 18 had adenocarcinoma of the cervix (AdCa).

Quality control of specimens

Since these were archival samples, the samples were not collected at the same time and there was concern that the DNA might have degraded in the older samples. However, this issue was previously examined by our group and we found no significant correlation between the age of the sample and DNA content [8,9].

Image analysis results

Epithelial cells were selected from within a region on the tissue section identified by the pathologist as the ROI, ensuring that the selected cells match the correct diagnosis. The CytoSavant computer-assisted image analysis system using software described in Kamalov et al. [7] collected the quantitative pathology measurements which were stored for analysis using feature set outlined in Table 1. Certain features are normalized to internal standards (lymphocytes) on each slide to adjust for stain intensity and reduce inter-patient variability. Only non-overlapping nuclei with clearly discernible borders were selected to be used in the final analysis. The final analysis had 149 ± 57 (mean ± standard deviation) epithelial nuclei and 106 ± 42 lymphocyte nuclei on average from each section.

Creation of the posterior odds by using Cytosavant variables

As a final note, not all the features from the CytoSavant were used to create the posterior odds for the cells. First, there are several features that typically are not used in analyses because they depend on the cells' position on the slide or other confounding factors independent of the disease state. Second, despite the normalization performed on some features, there are still several features in the feature set which are not independent of variation in stain preparation. Therefore, we limited our analysis to features that we could identify as orientation invariant and stain independent features. In an ad hoc procedure, we found that these variables could be further reduced to those features found to be significantly different across at least 2 of the groups (AdCa, AIS, and normal) [10]. Therefore, we limited the analysis to the 22 staining-dependent, orientation-invariant features found to significantly differ across the groups in a previous study by West et al. [10]. Results of the 22 features considered, 15 were retained in the logistic regression model after the variable selection procedure described in the previous section. The features and their standardized parameter estimates (including the intercept) are reported in Table 2. To summarize, the fitted logistic model is $\log\left(\frac{P(D=1|x}{P(D=0|x}\right) = -1.58 + 0.96x_1 - 1.12x_2 - 1.31x_3 + 0.62x_4 - 1.00x_5 - 0.21x_6 + 0.30x_7 + 1.74x_8 - 0.66x_9, -0.15x_{10} - 0.61x_{11} - 0.26x_{12} - 1.66x_{13} - 1.04x_{14} - 0.33x_{15}$, where the intercept is -1.58, x_1 is compactness, x_2 is eccentricity, x_3 is optical density kurtosis, x_4 is energy, x_5 is medium density objects, x_6 is sphericity, x_7 is entropy, x_8 is low vs. medium DNA, x_9 is density light spot, x_{10} is low DNA compactness, x_{11} is mean center of mass, x_{12} is density dark spot, x_{13} is low vs. high DNA, x_{14} is IODs-index, and x_{15} is high DNA average distance.

Table 2
Standardized parameter estimates

Variable estimates	
Intercept	−1.58
Compactness	0.96
Eccentricity	−1.12
Optical density kurtosis	−1.31
Energy	0.62
Medium density objects	−1
Sphericity	−0.21
Entropy	0.30
Low vs. medium DNA	1.74
Density light spot	−0.66
Low DNA compactness	−0.15
Mean center of mass	−0.61
Density dark spot	−0.26
Low vs. high DNA	−1.66
IODs-index	−1.04
High DNA average dist.	−0.33

Calculating a sensitivity and specificity

Using the formula of Metz, an ROC curve was generated and a sensitivity and specificity were optimized. The method had a sensitivity of 100% and specificity of 85%, and the ROC curve is given in Fig. 2. One can see from the ROC curve analysis that this method performs well [11,12].

Discussion

Multilevel data such as the data from a quantitative pathology analysis are becoming more common in the medical setting with the development of new diagnostic technologies. Information is gathered on one level, such as the cellular level, when the goal is to classify on a higher level, such as the tissue section. There has been work by Cadez, McLaren, Smyth, and McLachlan on a similar problem arising from technologies used in the detection and classification of anemia [13]. They consider a feature vector consisting of the cell volume and the hemoglobin amount in each cell and their goal is to automate the classification of the anemia diagnosis. Cadez et al. used a hierarchical Bayesian model with two levels. First, they model the bivariate distribution of the cell volume and hemoglobin amount for each patient, then they model the probability of falling into each diagnostic group. The patient is assigned to the group with the highest probability. Furthermore, they generalize their method so it is not limited to a bivariate feature space.

Although similar, our quantitative pathology adenocarcinoma diagnosis problem differs from the anemia diagnosis problem in an important way. The two problems are similar because both have measurements on a cellular level while classification is to be made on an individual (or tissue-section) level. However, we have more than fifty measurements on each cell. Although Cadez et al. generalized their method beyond the bivariate setting, fitting a mixture density model to a 50-dimensional feature space as they describe would be highly computer intensive and require large amounts of data. Instead, the method we propose in this study calculates the posterior log-odds of disease at the tissue-section level from the estimated posterior log-odds of disease at a cellular level, and uses the tissue-section estimates to classify the tissue section. The logistic regression methodology allows one to easily deal with the large number of variables, and eliminates the need to estimate the conditional density of the cell level feature vector given the disease state.

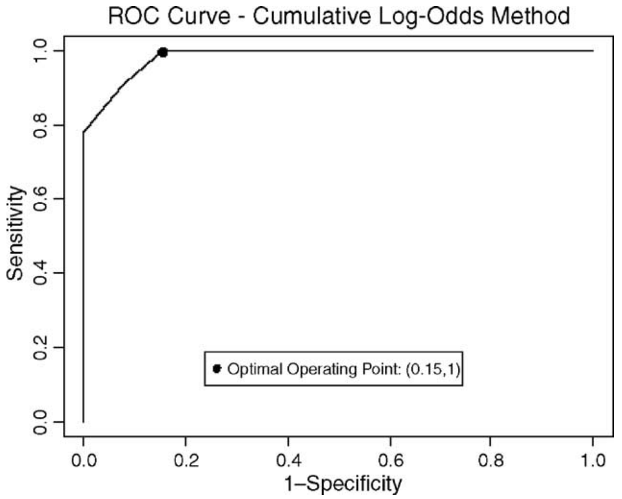

Fig. 2. This plot shows the ROC curve. The solid dot marks the point with the optimal sensitivity and specificity according to our criteria slices in this sample; it is hard to make any definitive statements concerning which method is actually better. However, the current method depends less on ad hoc procedures because the method is developed from probabilistic modeling.

Table 3
Sensitivities and specificities of the different analysis methods for adenocarcinoma diagnosis using this data set

Method	Sensitivity (%)	Specificity (%)
Cumulative log-odds method	100	85
PLR method [4]	96.4	92.3
PIV2.12 method [3]	94.5	100

Our group has previously explored various ad hoc classification methods which performed well [3,4], but the method discussed in this article provides a more theoretical approach to the problem. There are other ad hoc methods that have been tried, as described by Swartz et al. [3,4]. The best method, the Population Logistic Regression (PLR), from [4] involved an approach using logistic regression to summarize the cellular features, and then used logistic regression a second time at the tissue-section level. It also started with the reduced feature set as described in section 2.1.

The PIV2.12 (Percent of Index Values greater than 2.12) method used in [3] involved using a single, biologically motivated feature from the CytoSavant to create a tissue-section score. It involved taking the percentage of cells scoring above a threshold on the DNA Index feature. For details one is referred to the previous works. Table 3 compares this method to the other two methods mentioned here. As one can see, all the methods are comparable. Because of the small number of normal tissue slices in this sample, it is hard to make any definitive statements that one method is actually better. However, the current method depends less on ad hoc procedures because the method is developed from probabilistic modeling.

There are some limitations with this current analysis. First, our data have inherent correlation between the cells sampled from a given tissue section which is not accounted for in our model. Ideally, we would have fit a generalized linear mixed model to this data when estimating the log-odds of abnormality at the cell level to account for the correlation between cells from the same tissue section. However, we attempted to fit such a model in a couple of different statistical packages (R and SAS), and they failed to converge. One reason might be because all the cells from a given tissue section have the same diagnosis. Another reason for the convergence problems could be the dimensionality and size of our data set. However, the between-tissue section variation in this data set may not present that much of a problem, since the features used in the analysis were normalized to internal standards on each tissue slice to reduce the between-section variability.

Conclusions

Further investigation on a larger data set is required before this method can be used clinically; however, this investigation shows that this method has some clear advantages. First, the classification is developed from a statistical model involving the log-odds of disease which accounts for the multilevel structure of the data. Second, the log-odds can easily be transformed into either the odds of disease present in the tissue section or the probability of disease for the tissue section, both of which are easily understood by physicians. Third, this method is not extremely computationally intense and is easily implemented using available commercial software. Finally, the method performed well on the current data. The cumulative log-odds method provides a clear, easily implemented, interpretable model that accounts for the multilevel nature of the data and performs well.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygyno.2005.07.038.

References

[1] Doudkine A, MacAulay C, Poulin N, Palcic B. Nuclear texture measurements in image cytometry. Pathologica 1995;87:286–99.

[2] MacAulay C, Palcic B. An edge relocation segmentation algorithm. Anal Quant Cytol Histol 1990;12(3):165–71.

[3] Palcic B. Nuclear texture: can it be used as a surrogate end-point biomarker? J Cell Biochem 1994;19:40–6 (Supplement).

[4] Sox HC, Blatt MA, Higgens MC, Marton KI. Medical decision making. Butterworth-Heinemann; 1988. p. 67–80.

[5] Guillaud M, Cox D, Adler-Storthz K, Malpica A, Staerkel G, Matisic J, et al. Quantitative histopathological analysis of cervical intra-epithelial neoplasia sections: methodological issues. Cell Oncol 2004;26:31–43.

[6] Chiu D, Guillaud M, Cox D, Follen M, MacAulay C. Quality assurance system using statistical process control: an implementation for image cytometry. Cell Oncol 2004;00:1–17.

[7] Kamalov R, Guillaud M, Haskins D, Harrison A, Kemp R, Chiu D, et al. A Java application for tissue section image analysis. Comput Methods Programs Biomed 2005 (Feb);77(2):99–113.

[8] Swartz R, West L, Boiko I, Malpica A, MacAulay C, Carraro A, et al. Use of nuclear morphometry characteristics to distinguish between normal and abnormal cervical glandular histologies. Anal Cell Pathol 2003;27:193–200.

[9] Swartz RJ, West LA, Boiko IV, Malpica A, Guillaud M, Follen M, et al. Classifying populations from samples using quantitative pathology. Proceedings of the American Statistical Association, Biometrics Section [CD-ROM]. Alexandria, VA: American Statistical Association; 2002.

[10] West LA, Swartz RJ, Boiko IV, Malpica A, Guillaud M, MacAulay C, et al. DNA image cytometric measurement and nuclear morphometry characteristics of adenocarcinoma in situ and adenocarcinoma of the cervix. Am J Obstet Gynecol 2002;187(6):1566–73.

[11] Hosmer DW, Lemeshow S. Applied logistic regression, 2nd ed. New York: John Wiley and Sons, Inc.; 2000.

[12] Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8(4):283–98.

[13] Cadez IV, McLaren CE, Smyth P, McLachlan GJ. Hierarchical models for screening of iron deficiency anemia. Proceedings of the International Conference on Machine Learning. Los Gatos, CA: Morgan Kaufmann; 1999. p. 77–86.