Computer-derived Nuclear Features Distinguish Malignant from Benign Breast Cytology.

William H. Wolberg, M.D.

Departments of Surgery and Human Oncology,

University of Wisconsin, Madison, WI 53792

W. Nick Street, M.S.

Computer Sciences Department

University of Wisconsin, Madison, WI 53706

Dennis M. Heisey, Ph.D.

Department of Surgery,

University of Wisconsin, Madison, WI 53792

Olvi L. Mangasarian, Ph.D.

Computer Sciences Department.

University of Wisconsin, Madison, WI 53706

Keywords: Breast cancer, image processing, cytology, diagnosis, inductive machine learning

Corresponding author

William H. Wolberg, M.D.

Department of Surgery, University of Wisconsin Clinical Sciences Center, 600 Highland Avenue, Madison, WI 53792

FAX 608 263 7652

This study was supported in part by Air Force Office of Scientific Research grant AFOSR F49620-94-1-0036 and National Science Foundation grant CCR-9101801.

Abstract:

Computer-based analytical techniques permit accurate measurements of size, shape, and texture features. This paper describes the use of such analytical techniques to define computer-generated nuclear features. These features are then tested to distinguish between benign and malignant breast cytology.

The benign and malignant cell samples used in this study were obtained by fine needle aspiration (FNA) from a consecutive series of 569 patients: 212 with cancer and 357 with fibrocystic breast masses. Regions of FNA preparations to be analyzed were converted by a video camera to computer files that were visualized on a computer monitor. Nuclei to be analyzed were identified on the computer monitor and were roughly outlined by an operator using a mouse. Next, the computer-generated a "snake" to precisely enclose each designated nucleus. The computer calculated ten nuclear features for each nucleus. The ability to correctly classify samples as benign or malignant on the basis of these features was determined by inductive machine learning and logistic regression. Cross validation was used to test the validity of the predicted diagnosis. The logistic regression cross validated classification accuracy was 96.2% and the inductive machine learning cross validated classification accuracy was 97.5%.

The accuracy in distinguishing benign from malignant cytology based on the use of computer-generated nuclear features rivals that attained by visual diagnosis. The methods described in this paper will provide the basis for computerized systems to diagnose breast cytology.

Introduction

Progress over the past 30 years in computer analysis of microscope images has made possible highly accurate quantitative and objective feature assessment for diagnostic decision making [1]. We have identified computer-derived, quantitative digital features that accurately classify breast epithelial cells as benign or malignant. Classification is accomplished by inductive machine learning by a computer program based on the accumulated experience from previously diagnosed cases. The program can then be used to generalize in order to diagnose new cases that may differ from those previously encountered.

Materials and Methods

Patients and Aspirate

The benign and malignant cell samples used in this study were obtained by fine needle aspiration (FNA) from a consecutive series of 569 patients: 212 with cancer and 357 with fibrocystic breast masses.

To prepare an FNA, a small drop of viscous fluid is aspirated from breast masses by making multiple passes with a 23-gauge needle while negative pressure is being applied to an attached syringe. The aspirated material is expressed onto a silane-prepared glass slide and the aspirate is spread when a similar slide is applied face-to-face and the slides are separated with a horizontal motion like dealing playing cards. Preparations are immediately fixed in

95% ethanol, stained with hematoxylin and eosin, and processed.

In our analysis, aspirates were classified as cancer based on surgical biopsy and histologic confirmation. No data concerning breast histology was available in five patients who did not have definitive breast surgery because extensive distant metastases were present at diagnosis and in one patient who had definitive surgery elsewhere. Among the remaining 206 patients, 10 had in situ and 196 had infiltrating cancers: 148 ductal, 18 not otherwise specified, 12 lobular, 4 medullary 3 tubular, 3 comedo, 2 mucinous, and one each of signet ring, cribriform, sebaceous, papillary, inflammatory, and pseudosarcomatous carcinoma. Cytologically diagnosed benign breast masses were confirmed either by biopsy or by follow-up for a year. Fifty-one cytologically benign lesions were surgically excised: 35 fibroadenomas, 12 fibrocystic disease, and one each of fat necrosis, stromal fibrosis, adenolipoma, and atypical lobular hyperplasia. After a year, non-biopsied masses were considered to be benign if they had not enlarged [2].

Image Preparation

The imaged area on the aspirate slides is visually selected for minimal nuclear overlap. Areas of apocrine metaplasia are avoided. The image for digital analysis is generated by a JVC TK-1070U color video camera mounted atop an Olympus microscope and the image is projected into the camera with a 63 X objective and a 2.5 X ocular. The image is captured by a ComputerEyes/RT color framegrabber board (Digital Vision, Inc., Dedham MA) as a 640 x 400, 8-bit-per-pixel

Targa file.

The User Interface (Xcyt)

The first step in successfully analyzing the digital image is to specify the exact location of each cell nucleus. A graphical computer program called Xcyt was developed that allows the user to input the approximate location of a sufficient number of nuclei (10 to 20) to provide a representative sample. The program was developed using the X Window System and the Athena WidgetSet on a DECstation 3100.

A mouse is used to trace a rough outline of cell nuclei on the computer monitor. From this rough outline, the actual boundary of the cell nucleus is located by an active contour model known as a "snake"[3,4]. The "snake" is a deformable spline that seeks to minimize an energy function defined over the arclength of a curve. The energy function is defined in such a way that the snake, in the form of a closed curve, conforms itself to the boundary of a cell nucleus. The mathematical aspects of the snake calculations are described elsewhere [5].

Nuclear Features

Once the nuclei to be analyzed have been identified by the operator and have been enclosed by the computer-generated snakes, the computer calculates ten nuclear features for each nucleus [5]. These features are modeled such that higher values are typically associated with malignancy. Features were verified using idealized phantom cells [6]. Nuclear size is expressed by the Radius and Area features. Nuclear shape is expressed by Smoothness, Concavity, Compactness, Concave Points, Symmetry and Fractal Dimension features. Both size and shape are expressed by the Perimeter feature [6]. Nuclear Texture is measured by finding the variance of the grey scale intensities in the component pixels. The mean value, worst (mean of the three largest values), and standard error of each feature are computed for each image, resulting in a total of thirty features.

Classification procedure, cross validation

In both our inductive machine learning and logistic regression analyses, the accuracy of correctly classifying samples as benign or malignant on the basis of digitally determined nuclear features is determined by ten-fold cross-validation [7]. The resulting estimate is unbiased and accurate in cases such as ours that have a large number of training samples. Initially, the data set is divided into ten randomly selected, equal parts. One part is removed. A classifying algorithm is created from the nine remaining parts and the accuracy of the classifier is tested on the tenth part. The tenth part is then returned and the process is repeated until all parts have been tested.

Classifying algorithm developed by inductive machine learning

Image processing produces a database consisting of one 30-dimensional point for each sample. The 30 dimensions consist of the mean, standard error, and worst for the ten features. The classification procedure becomes one of pattern separation that accomplishes in, specifically, that of determining how points can best be separated into benign and malignant sets. The classification procedure is a variant on the Multisurface Method (MSM) [8,9] known as MSM-Tree (MSM-T) [10,11]. This method uses linear programming iteratively to place a series of separating planes in the feature space of the samples. If the benign and malignant sets can be separated by a single plane, the first plane will be so placed between them. If the sets are not linearly separable, MSM-T constructs a plane that minimizes an average distance of misclassified points. Depending on the separation accuracy attained, the procedure is recursively repeated on the two regions generated by each plane until satisfactory separation is achieved, i.e. each of the final regions contains mostly points of one category. The classifier thus obtained is then used as a decision tree to categorize new cases. MSM-T is similar to other decision tree methods such as CART [12] and C4.5 [13] but has been shown to be faster and more accurate on several real-world data sets [10].

Generally, simpler classifiers perform better on new data than do more complex ones. Therefore, we minimize not only the number of separating planes but also the number of features used in constructing the planes.

Classifying algorithm developed by logistic regression analysis

The classifying algorithm developed by logistic regression analysis was done with SAS [14] software. Other statistical analyses and graphics were done with Systat [15,16] software.

Results:

Feature analysis: Diagnostic features for 357 benign and 212 malignant samples are given in Table 1. Values for Area are expressed as square micra (μm^2), and for Radius and Perimeter as micra (μm). Values for remaining features are dimensionless. Independent samples t-test was not significant for differences between benign and malignant for mean of Fractal dimension, standard error of Texture, standard error of Smoothness, standard error of Symmetry, and for standard error of Fractal dimension. The differences were $p<0.001$ for all other features.

Logistic regression classification: A stepwise logistic regression selection process selected a model consisting of the variables standard error of the Radius, worst Radius, worst Texture, and worst Concave point. Standard error of the radius appeared to contribute little to the predictive ability of the model (as judged by sensitivity and specificity) and consequently was dropped. Nine benign and 12.4 malignant FNAs were misclassified when this model classified with 10-fold cross validation repeated 100 times. (Table 2).

Inductive machine learning classification: A computationally intensive search revealed that worst Area, mean Texture and worst

Smoothness gave the most accurate three-feature, single plane classification separation. The location of these points in the three dimensional space defined by the three classifying features is shown in Figure 1. The entire tenfold cross validation process was done five times and the results reported occurred three of the five times. Seven benign and seven malignant FNAs were misclassified when MSM-T was used to classify with cross validation (Table 3).

Discussion:

Nuclear feature analysis is better performed on cytologic FNA preparations than on the more commonly used histologic tissue samples. FNA cells are preserved intact whereas histologic processing cuts cells at various planes. Selection of nuclei for analysis, as is done in this paper, has been shown to be robust and operator independent [17]. A wide variety of histologic tumor types was included in this study, not only "high grade" infiltrating ductal cancers that would have made classification easy. Our pathologists do not grade cancers because of difficulties in interobserver variability [18], but the large standard error of nuclear area across samples (Table 1) indicates that an assortment of tumor grades [19] were included in the study.

We compared our feature values with those reported in the literature. Morphologic alterations result from any fixation process and features such as nuclear texture vary according to the nuclear stain used (e.g. Feulgen vrs hematoxylin). Therefore, feature values presented here pertain only to FNAs processed in the same manner as described herein. Separate algorithms will have to be developed if deviations are made in preparing or processing the FNAs (e.g. air drying). Never the less, the perimeter measurements for our benign and malignant nuclei were the same as those reported by Hutchison et. al. [20]. They found the mean nuclear perimeter ± standard deviation for 169 benign samples to be 23.99 ± 2.97 micra compared with 22.32 ± 3.38 micra for our 357 benign samples. Similarly, their values for the perimeter of 168 malignant samples was 35.29 ± 5.98 micra compared with 33.06 ± 6.28 micra for our 212 malignant samples. Their values for the nuclear area of benign samples was 36.43 ± 8.89 square micra and ours was 37.80 ± 10.96 while their malignant sample area was 76.38 ± 23.03 and ours was 80.30 ± 30.36. Therefore, the size values for both benign and malignant samples appear to be the same despite methodologic differences. However, the methodology used appears to make a difference in shape. "Shape" was defined by Hutchison et. al. [20] as perimeter^2/4π area, which we call compactness. They found the mean nuclear "shape" and standard deviation for 169 benign samples to be 1.28 ± 0.12 versus 1.071 ± 0.032 for our 357 benign samples. Similarly, their values for the "shape" of 168 malignant samples was 1.33 ± 0.15 versus 1.120 ± 0.049 for our 212 malignant samples. The difference between our Compactness and the nuclear "shape" of Hutchison et. al. is significant ($p<0.001$) for both benign and malignant samples. Although the absolute differences are small, the statistical significance arises from the small variability in the measurements; 3% and 4% respectively for benign and malignant in our series and 9% and 11% for that of Hutchison et. al.

A size, a texture, and a shape feature were selected to give the most accurate classification by both the logistic regression and by the inductive machine learning process. Worst Radius, worst Texture, and worst Concave points were used by the logistic regression model and worst Area, mean Texture, and worst Smoothness were used by the inductive machine learning model. The number of classifying features was kept to a minimum to avoid data overfitting. The Pearson correlation coefficient was so strong (0.983) between worst Radius and worst Area, and between worst Texture and mean Texture (0.912) that these feature values appear virtually interchangeable. However, the correlation was weaker (0.546) between worst Concave points and worst Smoothness. Classification based on digital feature analysis is robust; similar classification accuracy was obtained by logistic regression and inductive machine learning classification. The logistic regression cross validated classification accuracy is 96.2% and the inductive machine learning cross validated classification accuracy is 97.5%. These results are considerably better than the 89% accuracy based on individual cell analysis achieved by Hutchison et. al. [20]. Our improved accuracy apparently was achieved through the use of shape and texture features not measured by Hutchison et. al.

The logistic regression classification sensitivity is 94.1% ± 0.3% and the specificity is 97.5% ± 0.2%. The inductive machine learning classification sensitivity is 96.7% and the specificity is 98.0%. These performance parameters were determined by cross validation to test the validity of the predicted diagnosis. With our comprehensive assessment of shape features, we achieve better performance parameters than have previously been reported [20]. In fact, our performance parameters rival those obtained by visual diagnosis [21]. Giard and Hermans [21] emphasized the need for developing individual performance characteristics for persons doing fine needle aspiration of breast masses (FNA) because the accuracy achieved is very operator dependent. The reported accuracy of visually diagnosed breast fine needle aspirates (FNA) is over 90%. The overall accuracy was 94.3% in a 62-series study with a total of 23,741 satisfactory breast FNA's [6]. Individually, the mean sensitivity for these series was 0.91 +/- 0.07 and the mean specificity was 0.87 +/- 0.18. The relatively high standard deviations indicate that the accuracy achieved in individual series varies considerably and reflects the subjectivity of the procedure. Additionally, it is unlikely that such accuracy is generally achieved because publication bias is toward publishing favorable results. The variation in accuracy is largely due to the subjectivity which is inherent in visual interpretation. Visually assessed size, shape, and texture features that distinguish benign from malignant cells are now measured by computers. We believe that our methods provide a basis for highly accurate computerized diagnostic systems for breast cytology.

References:

1. Wied G, Bartels P, Bibbo M, et al: Image Analysis in Quantitative Cytopathology and Histopathology. Human Pathology 20:549-571, 1989.

2. Layfield L, Chrischilles E, Cohen M, et al: The Palpable Breast Nodule - A Cost Effective Analysis of Alternate Diagnostic Approaches. Cancer 72:1642-1651, 1993.

3. Kass M, Witkin A, Terzopoulos D: Snakes: Active contour models. Proc First Int Conf on Computer Vision 259-269, 1987.

4. Williams DJ, Shah M: A fast algorithm for active contours, in Proc. Third Int. Conf. on Computer Vision. Osaka, Japan, 1990, pp 592-595

5. Street WN, Wolberg WH, Mangasarian OL: Nuclear feature extraction for breast tumor diagnosis. Proceedings IS&T/SPIE International Symposium on Electronic Imaging 1905:861-870, 1993.

6. Wolberg WH, Street WN, Mangasarian OL: Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Letters 77:163-171, 1994.

7. Stone M: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society 36:111-147, 1974.

8. Mangasarian OL: Multi-surface method of pattern separation. IEEE Trans on information theory IT-14:801-807, 1968.

9. Mangasarian OL, Setiono R, Wolberg WH: Pattern Recognition via Linear Programming:Theory and Application to Medical Diagnosis, in Coleman TF, Li Y (eds): Large-Scale Numerical Optimization. Philadelphia, Pa, SIAM, 1990, pp 22-30

10. Bennett KP: Decision Tree Construction via Linear Programming, in Evans M (ed): Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference. 1992, pp 97-101

11. Mangasarian OL: Mathematical programming in neural networks. Technical Report, Computer Sciences, Univ Wisc 1129:1992.

12. Breiman L, Friedman J, Olshen R, et al: Classification and regression trees Pacific Grove, California, Wadsworth, Inc., 1984,

13. Quinlan JR: C4.5: Programs for Machine Learning San Mateo, CA, Morgan Kaufmann, 1993,

14. SAS/STAT User's Guide, Version 6 (ed4th). Cary, NC, SAS Institute Inc. 1989,

15. Wilkinson L, Hill MA, Welna JP, et al: SYSTAT for Windows:Statistics (ed5th). Evanston, IL, SYSTAT, Inc., 1992,

16. Wilkinson L, Hill MA, Miceli S, et al: SYSTAT for Windows:Graphics (ed5th). Evanston, IL, SYSTAT, Inc., 1992,

17. Wolberg WH, Street WN, Mangasarian OL: Breast cytology diagnosis with digital image analysis. Analyt Quant Cytol and Histol 15:396-404, 1993.

18. Gilchrist KW, Kalish L, Gould VE, et al: Interobserver reproducibility of histopathological features of stage II breast cancer. Breast Cancer Res Treatment 5:3-10, 1985.

19. VanDiest PJ, Risse EKJ, Schipper NW, et al: Comparison of light microscopic grading and morphometric features in cytological breast cancer specimens. Path Res Pract 185:612-616, 1989.

20. Hutchinson ML, Isenstein LM, Zahniser DJ: High-resolution and contextual analysis for the diagnosis of fine needle aspirates of breast. Analy Quant Cytol and Hist 13:351-355, 1991.

21. Giard RWM, Hermans J: The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer 69:2104-2110, 1992.

Table 1. Diagnostic features (357 benign, 212 malignant). Values for Area are expressed as square micra (μm^2), and for Radius and Perimeter as micra (μm). Values for remaining features are nondimensional. Independent samples t-test was not significant (indicated by * in the table) for differences between benign and malignant for mean of Fractal dimension, standard error of Texture, standard error of Smoothness, standard error of Symmetry, and for standard error of Fractal dimension. The differences were $p<0.001$ for all other features

	Benign (n=357)	Malignant (n=212)
Mean Radius (μm)	3.471 ± 0.508	5.004 ± 0.922
Mean Area (μm^2)	37.80 ± 10.96	80.30 ± 30.36
Mean Perimeter (μm)	22.32 ± 3.372	33.06 ± 6.275
Mean Texture	17.92 ± 3.993	21.60 ± 3.786
Mean Smoothness	0.093 ± 0.013	0.103 ± 0.013
Mean Compactness	0.080 ± 0.034	0.146 ± 0.053
Mean Concavity	0.046 ± 0.044	0.163 ± 0.075
Mean Concave points	0.026 ± 0.016	0.089 ± 0.034
Mean Symmetry	0.174 ± 0.025	0.194 ± 0.028
Mean Fractal dimension *	0.063 ± 0.007	0.063 ± 0.008
Std. error Radius (μm)	0.081 ± 0.032	0.175 ± 0.100

Std. error Area (μm^2)	1.730 ± 0.723	6.008 ± 5.171
Std. error Perimeter (μm)	0.570 ± 0.220	1.247 ± 0.740
Std. error Texture *	1.222 ± 0.031	1.213 ± 0.483
Std. error Smoothness *	0.007 ± 0.000	0.007 ± 0.003
Std. error Compactness	0.022 ± 0.001	0.033 ± 0.018
Std. error Concavity	0.026 ± 0.002	0.042 ± 0.022
Std. error Concave points	0.010 ± 0.000	0.015 ± 0.006
Std. error Symmetry *	0.021 ± 0.000	0.021 ± 0.011
Std. error Fractal dimension *	0.004 ± 0.000	0.004 ± 0.002
Worst Radius (μm)	3.825 ± 0.567	6.049 ± 1.245
Worst Area (μm^2)	45.67 ± 13.37	116.7 ± 50.37
Worst Perimeter (μm)	24.88 ± 3.87	40.49 ± 8.488
Worst Texture	23.53 ± 5.487	29.24 ± 5.480
Worst Smoothness	0.125 ± 0.020	0.145 ± 0.022
Worst Compactness	0.184 ± 0.094	0.375 ± 0.170
Worst Concavity	0.167 ± 0.141	0.454 ± 0.182
Worst Concave point	0.075 ± 0.036	0.183 ± 0.046
Worst Symmetry	0.271 ± 0.042	0.325 ± 0.076
Worst Fractal dimension	0.080 ± 0.014	0.092 ± 0.021

Table 1 continued

Table 2: Logistic regression classification: Predicted versus Confirmed Diagnosis.

	Malignant, Confirmed	Benign, Confirmed	Total
Malignant, Predicted	199.6 ± 0.65	9.0 ± 0.72	208
Benign, Predicted	12.4 ± 0.65	348.0 ± 0.72	361
Total	212	357	569

Table 3: Inductive machine learning classification: Predicted versus Confirmed Diagnosis.

	Malignant, Confirmed	Benign, Confirmed	Total
Malignant, Predicted	205	7	212
Benign, Predicted	7	350	357
Total	212	357	569

Legend for figure 1

The location of 212 malignant ■ and 357 benign O samples in the three dimensional space defined by the computer-generated features of mean Texture (Texture, x axis), worst Smoothness (W.Smoothness, y axis), and worst Area (W.Area, z axis). The feature values are standardized and the symbol sizes vary according to the perspective. MSM-T generates the plane that provides the best separation, in this case with 7 benign points lying on the side with 205 malignant points and 7 malignant points lying on the side with 350 benign points.