Image Analysis and Machine Learning Applied to Breast Cancer Diagnosis and Prognosis

William H. Wolberg M.D.[1], W. Nick Street M.S.[2], and Olvi L. Mangasarian Ph.D.[3]

From the Departments of Surgery, Human Oncology and Computer Sciences, University of Wisconsin, Madison, Wisconsin, U.S.A.

This study was supported in part by Air Force Office of Scientific Research grant AFOSR 89-0410 and National Science Foundation grant CCR-9101801.

Address reprint requests to:

William H. Wolberg, M.D.

Department of Surgery, University of Wisconsin Clinical

Sciences Center, 600 Highland Avenue, Madison, WI 53792

[1]Dr. Wolberg is Professor, Departments of Surgery and Human Oncology, University of Wisconsin, Madison, WI 53792

[2]Mr. Street is a Ph.D. student and Research Assistant, Computer Sciences Department, University of Wisconsin, Madison, WI 53706

[3]Dr. Mangasarian is Professor, Computer Sciences Department, University of Wisconsin, Madison, WI 53706

Running title: Breast cancer diagnosis and prognosis by computer

Keywords: Breast cancer, image processing, machine learning, diagnosis, prognosis

ABSTRACT:

Fine needle aspiration (FNA) accuracy is limited by, among other factors, the subjective interpretation of the aspirate. We have increased breast FNA accuracy by coupling digital image analysis methods with machine learning techniques. Additionally, our mathematical approach captures nuclear features ("grade") that are prognostically more accurate than are estimates based on tumor size and lymph-node status.

An interactive computer system evaluates, diagnoses, and determines prognosis based on nuclear features derived directly from a digital scan of FNA slides. A consecutive series of 569 patients provided the data for the diagnostic study. A 166 patient subset provided the data for the prognostic study. An additional 75 consecutive, new patients provided samples to test the diagnostic system. The projected prospective accuracy of the diagnostic system was estimated to be 97% by ten-fold cross validation and the actual accuracy on 75 new samples was 100%. The projected prospective accuracy of the prognostic system was estimated to be 86% by leave-one-out testing.

Introduction

We previously described a computer-based system for diagnosing breast fine needle aspirates (FNA) that is reproducible and independent of operator experience [37]. The system uses computer vision techniques to analyze size, shape and texture features of cell nuclei and classifies them using an inductive method based on linear programming. This paper describes accuracy of the system in diagnostically classifying 569 (212 malignant and 357 benign) FNAs and its prospective accuracy in testing on 75 (23 malignant, 51 benign, and 1 papilloma with atypia) newly obtained samples. Additionally, prognostic implications of the system were explored because the computer-analyzed features are very similar to those used in the visual assessment of nuclear grade.

Materials and Methods

Patients and Aspirate

The FNAs used to develop the diagnostic system were obtained from a consecutive sample of 569 patients: 212 with cancer and 357 with fibrocystic breast masses. Subsequently, 75 additional consecutive samples (23 cancerous, 51 benign, and one papilloma with atypia) were obtained and were used to test the diagnostic system. Information necessary for studying prognosis was available in 166 patients with primary invasive breast cancer of the total of 212 consecutive patients. The remaining 46 patients either had in situ cancers, or had distant metastases at the time of presentation. One hundred twenty-four patients of the 166 patients developed distant metastases sometime following surgery or were followed a minimum of 2 years without developing distant metastases.

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 by a similar slide as the slides are separated with a horizontal motion. Preparations are immediately fixed in 95% ethanol, stained with hematoxylin and eosin, and processed. Only palpable masses are aspirated and only solid masses that yield epithelial cells are analyzed. All cancers are histologically confirmed. Patients with fibrocystic masses are either biopsied or followed for a year if there is no enlargement of the previously aspirated mass.

Cancer patients were treated with either modified radical mastectomy or tylectomy, axillary dissection and radiation therapy to the breast. The maximum tumor diameter and the number of axillary lymph nodes involved with cancer were determined from the surgically excised specimens. Adjunctive chemotherapy was give to node-positive patients. Patients were followed at 3 month intervals for 2 years.

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. Non-filtered white light was used for illumination. The conversion for each pixel is grey=0.299 red + 0.587 green + 0.114 blue [10].

The User Interface

The first step in successfully analyzing the digital image is to specify the exact location of each cell nucleus. A graphical user interface was developed that allows the user to input the approximate location of enough nuclei to provide a representative sample. Eight to thirty nuclei were outlined with more being outlined when the sample consisted of visually heterogenous nuclei. The interface 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 each visible cell nucleus. Beginning with the user-defined approximate border as an initialization, the actual boundary of the cell nucleus is located by an active contour model known as a "snake"[15,35], 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 [29].

Nuclear Features

By using the computer-generated snakes, ten nuclear features are calculated for each cell [29]. These features are modeled such that higher values are typically associated with malignancy. Size and shape features were verified using idealized phantom cells. The size of the nuclei is measured by the Radius and Area features. Nuclear shape is quantified by Smoothness, Concavity, Compactness, Concave Points, Symmetry and Fractal Dimension. Both size and shape are measured by Perimeter. The Texture of the nuclei 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

Image processing produces a database consisting of one 30-dimensional point for each sample. The classification procedure becomes one of pattern separation, specifically, that of determining how points can best be separated into benign and malignant sets in the case of diagnosis, and into recurring and nonrecurring sets in the case of prognosis. The classification procedure is a variant on the Multisurface Method (MSM) [19,21] known as MSM-Tree (MSM-T) [4,20]. This method uses linear programming iteratively to place a series of separating planes in the feature space of the examples. If the two 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. The procedure is recursively repeated on the two regions generated by each plane until 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 (7) and C4.5 (25) but has been shown to be faster and more accurate on several real-world data sets (4).

Generally, simpler classifiers perform better on new data than do more complex ones. To generate a classifier that generalizes well to unseen cases, we minimize not only the number of separating planes but also the number of features used in constructing the planes. The best single-plane diagnostic classifier separates benign from malignant points based on three nuclear feature values for each case: mean texture, worst area, and worst smoothness. Multiple planes are needed for the prognostic classifier; the best results were obtained with one size feature, one or more shape features, and texture.

Estimate of Predictive Accuracy

Diagnostic predictive accuracy is estimated by ten-fold cross-validation (28). This train-and-test procedure divides the data set into ten randomly selected, equal parts and uses each in turn as a test set on a classifier created from the remaining nine sets. The estimate is unbiased and accurate in cases that have a large number of training samples. Because of the smaller number of available cases, "leave-one-out" testing (17) was used for the prognostic data.

Estimate of Probability of Malignancy

Distribution curves for malignant and benign points were determined by projecting the positions that the malignant and benign points occupy in three-dimensional space (determined by the values for mean texture, worst area, and worst smoothness) onto the normal of the separating plane. A Parzen window or kernel technique (23) was then used to approximate the probability densities of the malignant and benign points. The estimate of the probability of malignancy for a new point is determined from the ratio of the intercepts at that point with the malignant and benign distribution curves. Examples are shown in Figures 1 and 2.

Results

Reproducibility

Principal goals of computerized cytological diagnosis are higher accuracy, greater speed and decreased subjectivity. Reproducibility is a problem with visual assessments and interpretations [12]. To determine the degree of reproducibility of this computerized analysis a random group of 28 images was analyzed in duplicate and four in triplicate. Replicate assessments of symmetry, and fractal dimension varied by 1% or less; of radius, perimeter, and smoothness by 1 to 2%; and of area, compactness, concavity, and concave points by 2 to 10%.

Diagnostic Separation

Twenty-five of the 30 nuclear features measured were strongly diagnostic with t test values of p<0.001 (Table 1) [34]. Worst perimeter was the feature with the highest t value. Histograms for the benign and malignant distributions for worst perimeter are shown as Figure 3 [33]. Worst perimeter also gave the best single feature diagnostic separation with MSM-T (Table 2). Features with p<0.0001 in both backward and forward stepwise discriminant analysis as well as the logistic procedure [1] were worst radius, worst concave points, and worst texture; that is one size, one shape, and one texture feature. MSM-T provides a means to classify with more than one feature without assuming a normal distribution. Both the initial diagnostic separation and cross validation accuracy of the single-plane diagnostic classifier increased as two and three features were used for MSM-T (Table 2). The single-plane diagnostic classifier based on mean texture, the worst area, and the worst smoothness separated 97.5% of the cases successfully (Figure 4). The prospective accuracy was estimated at 97.2% with 96.7% sensitivity and 97.5% specificity using ten-fold cross validation. Using the standard error from the binomial distribution [32], we have 95% confidence that the true prospective accuracy - that is, the percentage of unseen cases that would be diagnosed correctly - lies between 95.8% and 98.6%. Seventy-five (23 malignant, 51 benign, and 1 papilloma with atypia) samples obtained subsequent to the development of the trained diagnostic classifier were used to test its accuracy. The new samples all were located in the correct diagnostic category by the classifier. The machine diagnosis was ambiguous in the case of the papilloma with atypia. The machine diagnosis based on location relative to the classification plane was benign but the estimated probability of malignancy based on the distribution curves was 0.57. The estimated probability of malignancy for all the 75 new samples and their actual diagnoses is shown in Figure 5.

Prognostic Separation

The observed median time for distance recurrence was 20 months for the 124 patients who had recurrent cancer or who had been followed for 2 years without recurrence. A breakpoint of two years was established for MSM-T analyses. Twenty-eight patients had distant recurrence of breast cancer by 2 years and 96 did not. Several of the nuclear features were strongly related to 2-year distant recurrence (Table 1). Separately, the recurrence data were analyzed by MSM-T with one, two, and three separating planes using all nuclear features or, alternatively, with the two, three, and four best prognostic features (Table 3). These data indicate that optimal separation and robustness occurred with two or three separating planes. Although better training separation was accomplished with four features using three planes, there was a marked deterioration in test accuracy, indicating overfitting of the data. Generally, nuclear features were predictive of recurrence: over 80% of those predicted to recur did so, and a similar percentage of those predicted not to recur did not. The overall accuracy is estimated at 86%, with a 95% confidence region of ± 6% [32]. The MSM-T separation based on this 2 year breakpoint and using the four best nuclear features with two separating planes accurately portrayed the patients' clinical course at times other than at 2 years. A Kaplan-Meier curve [14] shows the probability of distant disease-free survival for 166 patients; the 124 used in training the classifier plus 42 patients who have not recurred but have not yet been followed for 2 years (Figure 6).

The number of lymph nodes involved with cancer taken together with tumor size were weaker prognosticators than were the nuclear features taken alone.

Adding lymph node involvement and tumor size to the nuclear features did not increase prognostic accuracy (Table 6).

Discussion

The reported accuracy for visually diagnosing breast cancer from FNAs varies considerably. Giard and Hermans [11] reviewed the literature on FNA-performance parameters and found sensitivities from 0.65 to 0.98 and specificities from 0.82 to 1.00, with outliers of 0.34 and 0.59. They concluded that FNA diagnosis is highly operator-dependent and emphasized the need for developing individual performance characteristics for those doing this test. One goal of the present work is to improve the diagnostic accuracy of FNA by increasing its objectivity and thereby making it less operator-dependant.

Most diagnostic tests including FNA have an ambiguous gray zone between normal and abnormal. However, machine learning decisions are usually dichotomous-- in our case, either benign or malignant. To acknowledge diagnostic misclassifications and to compensate for them, we used the Parzen windows technique to estimate the probability that a specific sample is malignant. In clinical practice, after the probability of malignancy is calculated, a decision whether or not to biopsy is made in consultation with the patient.

The machine-learning techniques used in this study do not assume normal distributions so p values are not obtained. In our methodology, diagnostically or prognostically important features are identified by a computer-intensive search to find which features allow the classification algorithm to best fit the data. These features are then used to serially generate classifiers with 90% of the data; each classifier is then tested on the remaining 10% (cross validation). A similar process, leave-one-out testing, is used for smaller data sets. In leave-one-out testing, classifiers are generated with all but one of the samples and then tested on the remaining sample. Once the best set of features and the optimal number of separating planes is determined, a final classifier is generated using all the available data. The term "accuracy" is used to express correctness in machine-learning classification schemes. Accuracy is the number of true positive predictions plus the number of true negative predictions divided by the total sample size. Benign and malignant misclassifications are weighed equally.

Perimeter is the most important single feature for both diagnosis and for prognosis. This feature was developed to measure size but, by using a series of phantoms, we found that perimeter measures both size and shape. The commonality between linear regression statistics and single-plane MSM-T can be approached through Figure 3. Histograms for benign and malignant distributions cross at approximately 100, the optimal Wald-Wolfowitz [34] cut point is 106 (Z=-16.393), and 104 is the MSM-T cut point. These values are similar because the MSM-T separating plane is generated by minimizing the error distance between benign and malignant points. However, a classification method like MSM-T exploits interactions between the various features which are not obvious through the analysis of the p-values, leading to higher predictive accuracy.

In 1955, Black et al. [5] described the relationship between prognosis and nuclear atypia and in 1957 proposed a nuclear grading system [6]. A number of other investigators [9,13,18,26,30,31] subsequently confirmed the relationship between nuclear atypia and prognosis. However, visual grading systems were shown to be vulnerable to intra- and interobserver variation[27], so, calibrated oculars and projection microscopy were used to measure actual nuclear size. With these techniques, larger nuclear size was shown to be associated with a poorer prognosis[2,3,27,36]. Two studies [2,3] also found variation in nuclear size, as reflected in the standard deviation of nuclear-size features, to be prognostically unfavorable.

The advent of computerized digital image analysis made possible the measurement of nuclear size, shape, and texture features. In contrast to the methods used in other studies, our nuclear boundaries are determined directly by the computer with the "snake" program rather than manually with a digitizing tablet. Furthermore, our studies use the cellular smear-type preparations in which nuclear detail is better preserved than in the histological preparations used in previous studies. Despite these technical differences, our prognostic accuracy is almost identical to that reported by Komitowski and Janson [16]. They used projection microscopy and a digitizing tablet to determine size, shape, and texture features in 60 breast cancer patients. They achieved 85% prognostic accuracy; inclusion of tumor size increased accuracy to 92%. Pienta and Coffey [24] found that nuclear pleomorphism as measured by both nuclear area and intrasample variation increased with invasive histology and with axillary lymph node involvement with metastatic cancer.

Our observations corroborate those of others that nuclear morphometric features provide prognostic information independent of that derived from the status of metastatic disease in the axillary lymph nodes. Mittra and MacRae [22] found, in a simple meta-analysis of prognostic factors, a general interrelationship between the eight biological prognostic factors including tumor grade. These biological factors were not correlated with the clinical prognostic factors (axillary lymph node status and tumor size).

Our data indicate that nuclear features, similar to those evaluated in visual assessment of nuclear grade, are stronger predictors of recurrence than are the widely accepted prognostic features of axillary lymph node status and tumor size. Even at the extremes of tumor size and lymph node status, the accuracy is only 74% in classifying the 5-year relative survivals of patients with tumors smaller than 2 cm with no involved axillary lymph nodes and those patients with tumors equal to or larger than 5 cm with positive axillary nodes [8]. If our data are confirmed by others, many women with breast cancer who now have axillary lymph node removal for prognostic purposes will be spared the morbidity attendant that operation.

Two principal aspects of this work are the methods used and results obtained. The snake program accomplishes segmentation but other image processing methods may also be appropriate (e.g. region growing). Our results show that nuclear features, analogous to grade, can be objectively assessed and that these features are diagnostically and prognostically important. The present work is a step toward increasing the diagnostic potential of breast FNA.

We have adapted our UNIX-based system to a portable DOS based personal computer. Use of the system requires a video camera attachment for a microscope, a frame grabber board and the appropriate expert system software. Two alternatives exist for the expert system software. Either an individual expert system can be generated by the user from one's own cytology collection, or the FNA slides can be prepared in the manner described herein and our expert system based on 569 samples can be used and expanded.

Digital image analysis coupled with machine learning techniques has significant potential in making objective, accurate, and speedy cytological analysis available on a wide scale. This work is a step towards achieving this potential.

Acknowledgements

The authors gratefully acknowledge the suggestions of Kurt deVenecia about fractals, the statistical suggestions of Dennis Heisey, and the editorial assistance of Celeste Kirk. Appreciation is also expressed to Dr. Tilde Kline who, in 1983, provided technical advice on FNA preparation.

References:

1.SAS Institute Inc. editor.SAS/STAT User's Guide, Version 6. 4th ed. Cary, NC: SAS Institute Inc. 1989;

2.Baak JPA, Kurver PHJ, Snoo-Niewlaat AJE, Graef S, Makkink B. Prognostic Indicators in Breast Cancer-Morphometric Methods. Histopathology.6:327-339, 1982.

3.Baak JPA, VanDop H, Kurver PHJ, Hermans J. The Value of Morphometry to Classic Prognosticators in Breast Cancer. Cancer.56:374-382, 1985.

4.Bennett KP; Decision Tree Construction via Linear Programming. Evans M, editor.Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference. 1992; p. 97-101.

5.Black MM, Opler SR, Speer FD. Survival in breast cancer cases in relation to the structure of the primary tumor and regional lymph nodes. Surg Gynecol Obstet.100:543-551, 1955.

6.Black MM, Speer FD. Nuclear structure in cancer tissues. Surg Gynecol Obstet.105:97-102, 1957.

7.Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Pacific Grove, California: Wadsworth, Inc.; 1984;

8.Carter CL, Allen C, Henson DE. Relation of tumor size, lymph node status, and survival in 24, 740 breast cancer cases. Cancer.63:181-187, 1989.

9.Fisher ER, Redmond C, Fisher B, Bass G. Pathologic findings from the National Surgical Adjuvant Breast and Bowel Projects (NSABP). Prognostic discriminants for 8 year survival for node-negative invasive breast cancer patients. Cancer.65(supp):2121-2128, 1990.

10.Foley JD, van Dam A, Feiner SK, Hughes JF. Computer Graphics Principles and Practice.,Chapter 13, Second ed. Reading, MA: Addison-Wesley, 1990.

11.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.

12.Gilchrist KW, Kalish L, Gould VE, Hirschl S, Imbriglia JE, Levy WM, Patchefsky AS, Penner DW, Pickren J, Roth JA, Schinella RA, Schwartz IS, Wheeler JE. Interobserver reproducibility of histopathological features of stage II breast cancer. Breast Cancer Res Treatment.5:3-10, 1985.

13.Henson DE, Ries L, Freedman LS, Carriaga M. Relationship among outcome, stage of disease, and histologic grade for 22,616 cases of breast cancer. Cancer.68:2142-2149, 1991.

14.Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Statist Assoc.53:457-481, 1958.

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

16.Komitowski D, Janson C. Quantitative features of chromatin structure in the prognosis of breast cancer. Cancer.65:2725-2730, 1990.

17.Lachenbruch P, Mickey M. Estimation of error rates in discriminant analysis. Technometrics.10:1-11, 1968.

18.Le Doussal V, Tubiana-Hulin M, Friedman S, Hacene K, Spyratos F, Burnet M. Prognostic value of histologic grade nuclear components of Scraff-Bloom -Richardson (SCR): An improved score modification based on multivaraiate analysis of 1262 invasive ductal breast carcinomas. Cancer.64:1914-1921, 1989.

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

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

21.Mangasarian OL, Setiono R, Wolberg WH. Pattern Recognition via Linear Programming:Theory and Application to Medical Diagnosis. Large-Scale Numerical Optimization. Coleman TF, Li Y, editors. Philadelphia, Pa: SIAM, 1990; p. 22-30.

22.Mittra I, MacRae KD. A Meta-analysis of Reported Correlations between Prognostic Factors in Breast Cancer: Does Axillary Lymph Node Metastasis Represent Biology or Chronology? Eur J Cancer.27:1574-1583, 1991.

23.Parzen E. On estimation of a probability density and mode. Ann Mathematical Statistics.35:1065-1076, 1962.

24.Pienta KJ, Coffey DS. Correlation of nuclear morphometry with progression of breast cancer. Cancer.68:2012-2016, 1991.

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

26.Rank F, Dombernowsky P, Jespersin NCB, Pedersen BV, Keiding N. Histologic malignancy grading of invasive ductal breast carcinoma. Cancer.60:1299-1305, 1987.

27.Stenkvist B, Westman-Naeser S, Vegelius J, Holmquist J, Nordin B, Bengtsson E, Eriksson O. Analysis of reproducibility of subjective grading systems for breast carcinoma. J Clin Path.32:979-985, 1979.

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

29.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.

30.Todd JG, Dowle C, Williams MR, Elston CW, Ellis IO, Blamey RW, Haybittle JL. Confirmation of a prognostic index in primary breast cancer. Br J Cancer.56:489-492, 1987.

31.Wallgren A and Zajiecek J. The Prognostic Value of the Aspiration Biopsy Smear in Mammary Carcinoma. Acta Cytologica.20:479-485, 1976.

32.Weiss S, Kulikowski CA. Computer Systems That Learn. San Mateo, CA: Morgan Kaufmann; 1991;

33.Wilkinson L, Hill MA, Miceli S, Birkenbeuel G, Vang E.SYSTAT for Windows:Graphics. 5th ed. Evanston, IL: SYSTAT, Inc.; 1992;

34.Wilkinson L, Hill MA, Welna JP, Birkenbeuel GK.SYSTAT for Windows:Statistics. 5th ed. Evanston, IL: SYSTAT, Inc.; 1992;

35.Williams DJ, Shah M. A fast algorithm for active contours. Proc. Third

Int. Conf. on Computer Vision. Osaka, Japan: 1990; p. 592-5.

36.Wittekind C, Schulte E. Computerized morphometric image analysis of cytologic nuclear parameters in breast cancer. Analy Quant Cytol and Hist.9:480-484, 1987.

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

LEGEND FOR ILLUSTRATIONS

Figure 1: Photograph of a portion of the workstation monitor showing, at the top, a portion of the digitized image (x 157.5) from a malignant FNA with the converged "snakes". At the bottom, are the probability curves. The left curve is the projection of benign points and the right is that of the malignant ones. The vertical dashed red line is the projected MSM-T separating plane and the X along the abscissa is the value for this sample. The estimated probability of malignancy is 0.97.

Figure 2: Similar to Figure 2 but for a benign FNA. The position of the X along the abscissa indicates the estimated probability of malignancy is 0.26.

Figure 3: Histograms for the Worst Perimeter feature for the benign (left) and malignant (right) samples in the training set.

Figure 4: Diagnostic Separating Plane in Three Dimensions
In order to clarify the plot, only 10% of the correctly classified benign and malignant points are shown here. All of the misidentified points are shown.

Figure 5: Estimated probability of malignancy and the actual diagnosis for 75 new samples.

Figure 6. Kaplan Meier plot for the probability of distant disease-free survival for 166 patients classified by the MSM-T breakpoint at two years as recurring ----------or nonrecurring _______________ . The MSM-T breakpoint was established from the 124 patients who had recurred or who had been followed for two years without recurrence.

Table 1: Independent samples pooled variances t-tests on nuclear features for diagnosis (Dx) and prognosis (Px)(distant disease recurrence by 2 years) arranged by descending prognostic significance. The p values for diagnosis are listed in the second and those for prognosis are listed in the fourth column. Because of multiple comparisons (30), the reader may wish to apply a Bonferroni correction that can be accomplished by multiplying the p values by 30.

	t for Dx	p for Dx	t for Px	p for Px
W PERIMETER	29.924	<0.001	−3.955	<0.001
W AREA	25.197	<0.001	−3.929	<0.001
AREA	23.968	<0.001	−3.904	<0.001
W RADIUS	29.085	<0.001	−3.904	<0.001
PERIMETER	36.540	<0.001	−3.689	<0.001
RADIUS	25.536	<0.001	−3.615	<0.001
S AREA	15.402	<0.001	−3.297	0.001
CONCAVE POINTS	29.666	<0.001	−3.210	0.002
S RADIUS	16.340	<0.001	−2.797	0.006
S PERIMETER	16.097	<0.001	−2.620	0.010
W CONCAVE POINTS	31.216	<0.001	−2.529	0.013
FRACTAL DIMENSION	−0.180	0.857	1.760	0.081
CONCAVITY	23.455	<0.001	−1.717	0.089
S FRACTAL DIMENSION	1.925	0.055	1.667	0.098
W TEXTURE	12.016	<0.001	1.340	0.183
TEXTURE	10.850	<0.001	1.273	0.205
S TEXTURE	−0.188	0.851	1.068	0.287
S COMPACTNESS	7.380	<0.001	1.016	0.312
S SMOOTHNESS	−1.465	0.143	1.010	0.315
SYMMETRY	8.787	<0.001	0.736	0.463
W CONCAVITY	20.952	<0.001	−0.674	0.501

S CONCAVITY	6.376	<0.001	0.559	0.577
W SYMMETRY	11.022	<0.001	0.544	0.587
W FRACTAL DIMENSION	8.082	<0.001	0.393	0.695
W COMPACTNESS	17.311	<0.001	0.350	0.727
COMPACTNESS	17.908	<0.001	−0.331	0.742
SMOOTHNESS	9.292	<0.001	−0.298	0.767
S CONCAVE POINTS	10.942	<0.001	−0.289	0.773
S SYMMETRY	0.391	0.696	0.253	0.801
W SMOOTHNESS	10.932	<0.001	−0.248	0.805

Table 2: Best features (based on training set separation) and testing correctness percentages for single plane separation of diagnostic data. All possible feature combinations were tested to determine which single feature and which combinations of two, and three features most accurately separated the benign from the cancers (training). The combinations that obtained the best separation were then tested by cross validation, and the percent correctness is reported.

Number of Features	Features	Separation	Cross Validation
1	W Perimeter	91.6%	91.4%
2	W Area, W Smoothness	96.3	94.8
3	W Area, W Smoothness, M Texture	97.5	97.2

W, worst; M, mean

Table 3: Best features (based on training set separation) and testing correctness percentages for prognosis data. All possible feature combinations were tested to determine which single feature and which combinations of two, three, and four features most accurately separated the nonrecurrers from the recurrers (training). The combinations that obtained the best separation were then tested using the leave-one-out approach, and the percent correctness is reported.

	NUMBER OF PLANES		
Number of Features	1	2	3
1	SE Perimeter 71.8%		
2	SE Perimeter, SE Smoothness 74.2%	W Radius, W Fractal Dim 79.8%	
3	SE Area, SE Compactness, SE Fractal Dim 75.0%	M Area, W Concave Pts, W Fractal Dim 81.5%	M Smoothness, M Compactness, M Fractal Dim 83.9%
4	M Radius, M Area, SE Concave Pts, SE Fractal Dim 76.6%	M Texture, W Area, W Concavity, W Fractal Dim 86.3%	M Texture, M Compactness, W Area, W Fractal Dim 81.4%

SE, Standard error; W, worst; M, mean

Table 4: Separation percentages with and without Node Status and Tumor Size, using two separating planes (M=mean, SE=standard error, W=worst)

	Alone		Adding Node Status & Tumor Size	
Nuclear Features	Train	Test	Train	Test
None			77.6%	76.6%
W Radius, W Fractal Dimension	82.3%	79.8%	82.9%	80.6%
M Area, W Concave Pts, W Fractal Dimension	85.5%	81.5%	83.7%	79.0%
W Area, W Concavity, W Fractal Dimension, M Texture	88.7%	86.3%	84.5%	77.4%