# Breast Cancer Diagnosis and Prognosis via Linear Programming *

Olvi L. Mangasarian, W. Nick Street † & William H. Wolberg ‡

Mathematical Programming Technical Report 94-10

Revised December 19, 1994

## Abstract

**Two medical applications of linear programming are described in this paper. Specifically, linear-programming-based machine learning techniques are used to increase the accuracy and objectivity of breast cancer diagnosis and prognosis. The first application to breast cancer diagnosis utilizes characteristics of individual cells, obtained from a minimally invasive fine needle aspirate, to discriminate benign from malignant breast lumps. This allows an accurate diagnosis without the need for a surgical biopsy. The diagnostic system in current operation at University of Wisconsin Hospitals was trained on samples from 569 patients and has had 100% chronological correctness in diagnosing 131 subsequent patients. The second application, recently put into clinical practice, is a method that constructs a surface that predicts when breast cancer is likely to recur in patients that have had their cancers excised. This gives the physician and the patient better information with which to plan treatment, and may eliminate the need for a prognostic surgical procedure. The novel feature of the predictive approach is the ability to handle cases for which cancer has not recurred (censored data) as well as cases for which cancer has recurred at a specific time. The prognostic system has an expected error of 13.9 to 18.3 months, which is better than prognosis correctness by other available techniques.**

*This research was supported by Air Force Office of Scientific Research Grant F–49620–94–1–0036 and National Science Foundation Grant CCR–9322479.

†Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706 USA. Email: olvi@cs.wisc.edu, street@cs.wisc.edu

‡General Surgery Department, Clinical Sciences Center, University of Wisconsin, Madison, WI 53792 USA. Email: wolberg@eagle.surgery.wisc.edu

Despite a great deal of public awareness and scientific research, breast cancer continues to be the most common cancer and the second largest cause of cancer deaths among women [19]. Approximately 12% of U.S. women will be diagnosed with breast cancer [21], and 3.5% will die of it [23]. The annual mortality rate of approximately 28 deaths per 100,000 women has remained nearly constant over the past 20 years [20]. A breast cancer victim's chances for long-term survival are improved by early detection of the disease, and early detection is in turn enhanced by an accurate diagnosis. The choice of appropriate treatments immediately following surgery is largely influenced by prognosis, that is, the expected long-term behavior of the disease.

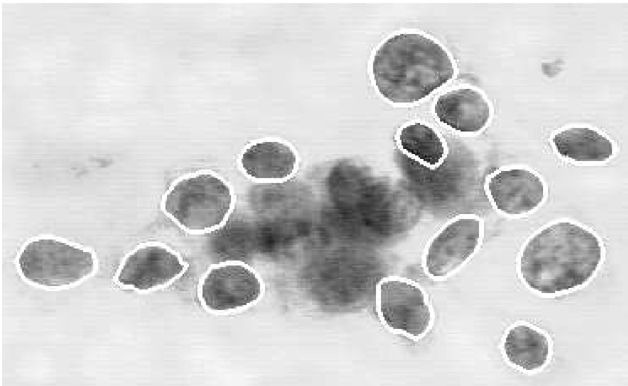

Figure 1: A magnified image of a malignant breast fine needle aspirate. Visible cell nuclei are outlined by a curve-fitting program. The **Xcyt** system also computes various features for each nucleus and accurately diagnoses the sample. Interactive diagnosis takes about 5 minutes.

We will describe two significant applications of linear programming in the field of breast cancer research, one in diagnosis and one in prognosis. Both applications, currently in clinical practice, depend on the analysis of cellular images, accomplished with a computer program called

**Xcyt**, written by one of the authors [31], that we describe in Section 1. The first application to breast cancer diagnosis has been described earlier [17, 18, 32, 34, 35] and is outlined in Section 2. The details of the diagnosis process in a clinical setting are described in Section 3. The second application to breast cancer prognosis [31] has not been published in the open literature and is described in some detail in Section 4. Computational results concerning the expected accuracy of the prognostic system are contained in Section 5. Section 6 shows the clinical importance of recurrence prediction, and Section 7 describes some possible extensions to the prognostic research. Section 8 concludes the paper and points to future work.

# 1 The Xcyt Image Analysis Program

The results of all of this research have been incorporated into an easy-to-use graphical computer program called **Xcyt**. At the time of this writing, **Xcyt** performs the following functions:

- Analysis of cytological (i.e., cellular) features based on a digital scan;
- Diagnosis of the image as benign or malignant, along with an estimated probability of malignancy;
- For cancerous samples, prediction of when the cancer is likely to recur.

First, a sample of fluid is taken from the patient's breast. This outpatient procedure involves using a small-gauge needle to take the fluid, known as a fine needle aspirate (FNA), directly from a breast lump or mass, the lump having been previously detected by self-examination and/or mammography. The fluid from the FNA is placed on a glass slide and stained to highlight the nuclei of the constituent cells. An image from the FNA is transferred to a workstation by a video camera mounted on a microscope. A portion of such an image is depicted in Figure 1. [1]

**Xcyt** uses a curve-fitting program to determine the exact boundaries of the nuclei. The boundaries are initialized by an operator using a mouse pointer. See Figure 1. For a typical image containing between 10 and 40 nuclei, the image analysis process takes approximately two to five minutes. Ten features are computed for each nucleus: area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments), and texture (variance of gray levels inside the boundary). The mean value, extreme value (i.e., largest or worst value: biggest size, most irregular shape) and standard error of each of these cellular features are computed for each image, resulting in a total of 30 real-valued features.

[1] Twenty of these images are available by anonymous ftp or via the World Wide Web from ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/images.

# 2 The Diagnostic System

Most breast cancers are detected by the patient as a lump in the breast. The majority of breast lumps are benign so it is the physician's responsibility to diagnose breast caner, that is, to distinguish benign lumps from malignant ones. There are three available methods for diagnosing breast cancer: mammography, FNA with visual interpretation, and surgical biopsy. The reported sensitivity (i.e., ability to correctly diagnose cancer when the disease is present) of mammography varies from 68% to 79% [7], of FNA with visual interpretation from 65% to 98% [8], and of surgical biopsy close to 100%. Therefore, mammography lacks sensitivity, FNA sensitivity varies widely, and surgical biopsy, although accurate, is invasive, time consuming, and costly. The goal of the diagnostic aspect of our research is to develop a relatively objective system that diagnoses FNAs with an accuracy that approaches the best achieved visually.

In contrast to previous work [17, 33] where cytological features were subjectively evaluated by the attending physician, the diagnostic system in current use at University of Wisconsin Hospitals uses the **Xcyt** system described above to generate a 30-dimensional feature vector for each patient. This analysis was performed for each of 569 patients for which the actual diagnostic outcome is known. Malignant cases were confirmed by biopsy, while benign cases were confirmed either by biopsy or by subsequent periodic medical examinations. These 569 vectors, along with the known outcomes, represent a *training set* with which a classifier can be constructed to diagnose future examples. These examples were used to train a linear programming-based diagnostic system by a variant of the multisurface method (MSM) [14, 15] called MSM-Tree (MSM-T) [1, 2] which we briefly describe now.

Let $m$ malignant n-dimensional vectors be stored in the $m \times n$ matrix $A$, and $k$ benign n-dimensional vectors be stored in the $k \times n$ matrix $B$. The points in $A$ and $B$ are strictly separable by a plane in the n-dimensional real space $\mathcal{R}^n$ represented by

$$x^T w = \gamma, \tag{1}$$

if and only if

$$Aw \geq e\gamma + e, \; Bw \leq e\gamma - e. \tag{2}$$

Here $w \in \mathcal{R}^n$ is the normal to the separating plane, $\frac{|\gamma|}{(w^T w)^{\frac{1}{2}}}$ is the distance of the plane to the origin in $\mathcal{R}^n$, and $e$ is a vector of ones of appropriate dimension. In general, the two sets will not be strictly linearly separable and the inequalities (2) will not be satisfied. Hence we attempt to satisfy them approximately by minimizing the average sum of their violations by solving the following linear program:

$$\begin{array}{lll} \underset{w,\gamma,y,z}{\text{minimize}} & \frac{e^T y}{m} + \frac{e^T z}{k} & \\ & Aw + y \;\; \geq & e\gamma + e \\ \text{subject to} & Bw - z \;\; \leq & e\gamma - e \\ & y, z \;\; \geq & 0. \end{array} \tag{3}$$

The linear program will generate a strict separating plane (1) satisfying (2) if such a plane exists, in which case $y = 0, z = 0$. Otherwise it will minimize the average sum of the violations $y$ and $z$ of the inequalities (2). This intuitively plausible linear program has significant theoretical and computational consequences [2], such as naturally eliminating the null point $w = 0$ from being a solution, a difficulty that other linear programming formulations exclude in an ad hoc manner [9, 10, 14]. Once the plane $x^T w = \gamma$ has been obtained, the same procedure can be applied recursively to one or both of the newly created halfspaces $x^T w > \gamma$ and $x^T w \leq \gamma$, if warranted by the presence of an unacceptable mixture of benign and malignant points in the halfspace. Figure 2 shows an example of the types of planes generated by MSM-T. MSM-T has been shown [1] to learn concepts as well or better than more well-known decision tree learning methods such as C4.5 [26] and CART [3]. In our implementation, the solving of the linear programs is carried out using the MINOS numerical optimization software [22].

A key issue in machine learning is to avoid "overtraining" the classifier, that is, memorizing details of the training data at the expense of good generalization to unseen data. Even a single plane can be considered an overtrained classifier if the dimensionality of the feature space is high. In our case, better generalization was achieved by reducing the number of input features considered. Taking advantage of the training speed of MSM-T, we built classifiers using all subsets of one, two, three or four features and one or two separating planes. Combinations

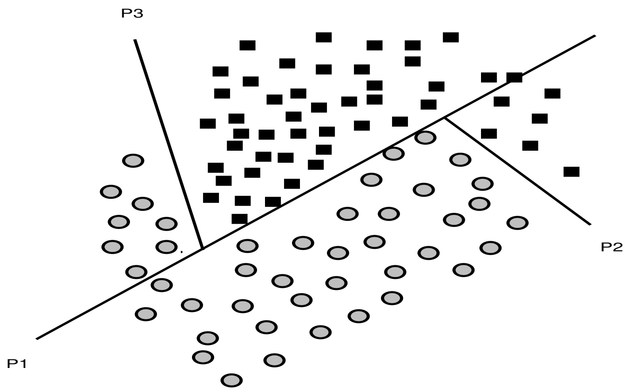

Figure 2: MSM-T separating planes.

that resulted in classifiers that separated the training set well, were evaluated using cross-validation [30] [2] to estimate their true accuracy, that is, how they would perform in actual practice. The best results were obtained with one plane and three features: Extreme area, extreme smoothness and mean texture. The predicted accuracy, estimated with cross-validation, was 97.5%. This level of accuracy is as good as the best results achieved at specialized cancer institutions.

**Xcyt** also uses the Parzen window density estimation technique [24, 25] to estimate the probability of malignancy for new patients. All of the points used to generate the separating plane $x^T w = \gamma$ in the 3-dimensional space of extreme area, extreme smoothness and mean texture, were projected on the normal $w$ to the separating plane. Using the Parzen window kernel technique, we then 'count' the number of benign and malignant points at each position along the normal, thus associating a number of malignant and a number of benign points with each point along this normal. Figure 3 depicts densities obtained in this fashion using the 357 benign points and 212 malignant projected on the normal $w$ to the plane $x^T w = \gamma$. Note that $\gamma$ is close, but not identical, to the location along the normal to the separating plane where the two density functions intersect.

The probability of malignancy for a new case can be computed with a simple Bayesian computation, taking the height of the malignant density divided by the sum of the two densities at that point. This assumes that the prior probability of each outcome is 0.5. This was thought to be preferable to using the observed prior probability

[2] In particular, ten-fold cross-validation was used. The predictive model is trained using 90% of the training examples and tested on the remaining 10%. This is done ten times, each time testing on a different 10%. The average performance on the testing sets gives an accurate, unbiased estimate of real-world performance.

(more than 60% benign) in order to avoid overly optimistic probabilities.

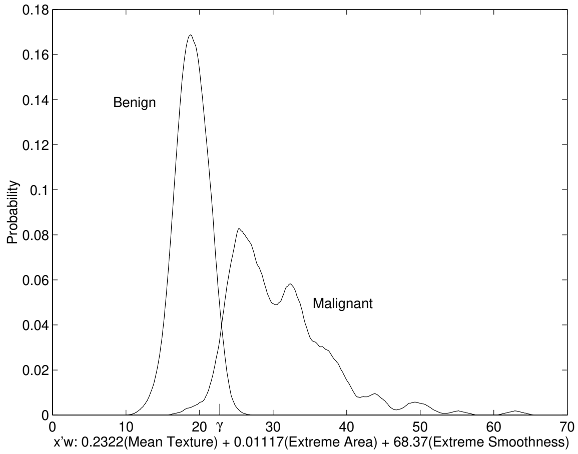


Figure 3: Densities of benign and malignant points along the normal $w$ to the separating plane $x^T w = \gamma$.

# 3 Clinical Usage of Xcyt

As mentioned earlier, the **Xcyt** system has been used by one of the authors (WHW, a surgical oncologist) in his clinical practice since 1993. In that period, the classifier has achieved 100% correctness on the 131 consecutive new cases that it has diagnosed (94 benign, 37 malignant). Analysis and diagnosis of the fine needle aspirate for a new patient can be performed in a few minutes by the attending physician using **Xcyt**. Once the FNA slide from a new patient has been analyzed, the patient is shown a density diagram similar to Figure 3 along with the value of $x^T w$ for the sample. The patient can then easily appraise the diagnosis in relation to hundreds of other cases, in much the same way that an experienced physician takes advantage of years of experience. Thus the patient has a better basis on which to base a treatment decision. For instance, a value of $x^T w$ falling in the region of Figure 3 where the densities overlap would correspond to a "suspicious" diagnosis. In particular, when the probability of malignancy (computed as described in the previous paragraph) is between 0.3 and 0.7, it is considered to be indeterminate and a biopsy is recommended. This is a rare case, as only eight of the 131 new cases have fallen into this suspicious region while 103 been assigned probabilities either less than 0.1 or greater than 0.9. Still, different patients may have very different reactions to the same readings. Masses from patients who opt for surgical biopsy have their diagnosis histologically confirmed. Patients who choose not to have the biopsy done are followed for a year at three-month intervals to check for changes in the mass.

We have successfully tested **Xcyt** on slides and images from researchers at other institutions who used the same preparation methodology. Through collaborative studies, **Xcyt** will soon be made available for clinical use at these institutions. Our research will also benefit from these collaborations by gaining new training cases, resulting in a more robust classification system. Further, there is some evidence that the classification system built into **Xcyt** may be applicable to other forms of cancer, even without modification. Twenty slides of FNA's taken from thyroid tumors at UCLA hospitals were successfully diagnosed, indicating that there may be significant structural similarities between the two types of cancers.

Our research into diagnosis from FNA has resulted in a system that delivers results (estimated 97.5% accurate, observed 100%) which are at least as reliable as any procedure for diagnosis short of a more invasive histological examination of the removed tumor. Further, FNA diagnosis can be performed quickly (about 15 minutes total) on an outpatient basis, which is not true of other types of diagnostic procedures. Our current collaborative efforts will expand the use of **Xcyt** to other specialized cancer centers; however, we foresee a time in the near future when the Internet will carry this type of medical expertise almost instantly to patients regardless of their community.

# 4 Prognosis: Recurrence Surface Approximation

Our second research area concerns the more difficult problem of prognosis, that is, the long-term outlook for the disease for patients whose cancer has been surgically removed. This problem does not fit into the usual classification paradigm of discriminating between two classes. While a patient can be classified as a 'recur' if the disease is observed at some subsequent time to tumor excision, a patient for whom cancer has not recurred and may never recur, has an unknown or censored [13] time to recur (TTR). For the latter patients, all that is known is the disease-free survival (DFS) time of their last checkup.

We approach the prediction of TTR as a function estimation problem, a mapping of an n-dimensional input of input features to a one-dimensional output which represents the expected time of recurrence. Our solution to this estimation problem is by a linear programming-based Recurrence Surface Approximation (RSA) technique. RSA determines a linear combination of the input features that approximates TTR. However, prognostic prediction is not a simple functional estimation problem, since the endpoint (time to recur) is known for only a fraction of the cases. The intuitive motivation for the RSA approach is that:

- Recurrence actually takes place at some point in time previous to its detection. However, the difference between the time a recurrence is detectable (actual TTR) and the time it is actually detected (observed TTR) is probably small.

- Observed DFS time is a *lower bound* on the recurrence time of that patient.

The RSA linear program is based on the idea of constructing a surface which bounds from above the DFS times for the non-recurring training cases and closely bounds from below the TTR times of the recurrent training cases as follows:

$$
(4)\qquad \begin{array}{ll} \underset{w,\gamma,v,y,z}{\text{minimize}} & \frac{1}{m}e^T y + \frac{1}{k}e^T z + \frac{\delta}{m}e^T v \\ \text{subject to} & \begin{array}{rcccl} -v & \leq & Mw + \gamma e - t & \leq & y \\ & & -Nw - \gamma e + r & \leq & z \\ & & v, y, z & \geq & 0 \end{array} \end{array}
$$

The purpose of this linear program is to learn the weight vector $w$ and the constant term $\gamma$. These parameters determine a recurrence surface $s = x^T w + \gamma$, where $x$ is the $n$-dimensional vector of measured features and $s$ is the surface (in this case, a plane defined on the feature space) which fits the observed recurrence times and overestimates the DFS times. Here, $M$ is an $m \times n$ matrix of the $m$ recurrent points, with recurrence times given by the $m$-dimensional vector $t$. Similarly, the $k$ non-recurrent points are collected in the $k \times n$ matrix $N$, and their last known disease-free survival times in the $k$-dimensional vector $r$. The vectors $y$ and $z$ represent the errors for recurrent and non-recurrent points, respectively; overestimating the TTR of recurrences is considered an error, while predicting a TTR which is shorter than an observed DFS is also an error. The objective averages the errors over their respective classes.

Underestimated recurrent points are not considered to be as serious of an error as overestimated ones. To reflect this, the $v$ term in the objective is weighed by an appropriately small positive parameter $\delta$, forcing underestimated recurrent points closer to the surface. Based on a perturbation theorem [16], for a sufficiently small positive $\delta$, that is $0 < \delta \leq \bar{\delta}$ for some $\bar{\delta}$, the objective minimizes the weighted term conditionally, i.e., of those possible variable values which minimize the first two terms of the objective, those values which minimize the third term are chosen. In this work, the "sufficiently small" value of $\delta$ was chosen empirically, by lowering it until further reductions had no effect on the training objective.

As in classification, it is important to choose the right subset of features to get the best generalization. We chose an appropriate feature set in the following automatic fashion. A tuning set – one tenth of the training cases – was first set aside. The RSA linear program was then solved using all of the input features, and the resulting surface was tested on the tuning set. Features were then removed, one by one, by setting the smallest element (in magnitude) of the coefficient vector $w$ to zero. [3] Each new problem was solved and the result tested on the tuning set, until only one feature remained. [4] Using the features which showed the best performance on the tuning set, we then re-optimized using all the training data (i.e., restored the tuning set). In this manner, we can use the tuning set to select the complexity of the model without paying the penalty of losing some of the training set.

# 5 Computational Results for the Prognostic System

The RSA procedure was tested with leave-one-out testing [12] to evaluate its accuracy in predicting future outcomes. The leave-one-out method is a special case of cross-validation in which the predictive model is built using all but one of the examples and tested on the left-out case; this is repeated using each example, in turn, as the test case. Of the 569 patients from the diagnosis study, the 187 malignant cases with follow-up data (44 of which have recurred) were used. The input consists of the thirty nuclear features computed by **Xcyt** together with with two traditional prognostic predictors: tumor size and number of involved lymph nodes. As with MSM-

[3] All feature values were previously scaled to be zero mean and unit standard deviation, so that the magnitude of the weight vector component correlates roughly with the relative importance of the corresponding feature.

[4] These subsequent linear programs are easily formulated by placing explicit upper and lower bounds of zero on the appropriate elements of $w$. A 'hot start' can then be used to solve the new LP starting from the solution to the previous one. These solutions are found very quickly, often in one or two orders of magnitude fewer simplex iterations than the original problem.

T, the linear program is implemented using MINOS 5.4.

Table 1 shows the mean generalization errors of the RSA formulation compared with the following prediction methods:

- **Pooled RSA:** This method is identical to RSA except that all the points are weighted equally in the objective function, rather than the recurrent and non-recurrent cases being averaged separately. The resulting objective function is

$$(5) \qquad \frac{1}{m+k}e^T y + \frac{1}{m+k}e^T z + \frac{\delta}{m+k}e^T v$$

- **Least 1-norm Error on Recurs:** An obvious method for predicting recurrence is to construct the predictive surface by a least-error fit on those cases for which the outcome is known, in our case, those with an observed recurrence time. We chose the 1-norm error, minimizing the average error on the recurrent cases but testing on all the examples. For compatibility, this test was run using the greedy feature selection method described earlier.

| | All Points | Non-recur | Recur |
|---|---|---|---|
| RSA (4) | 18.3 | 19.9 | 13.0 |
| Pooled RSA (5) | 13.9 | 6.1 | 39.3 |
| Least 1-norm Error | 21.8 | 25.8 | 9.0 |

Table 1: Average error (in months) of various prognostic formulations on Wisconsin prognostic data using leave-one-out testing.

Comparative results on all points, recurrent cases only and non-recurrent cases only are shown in Table 1 for the various prediction methods. We emphasize that these are estimates of the method's real-world performance, and measure only those cases known to be in error: overestimated recurrences (prediction was late) and underestimated non-recurrent cases (prediction was early). Both RSA approaches significantly lower the mean prediction error compared to a simple fit. We note that the original RSA formulation performs comparably on both recurrent and non-recurrent cases, while the pooled error method greatly favors the majority non-recurrent class, thereby lowering the mean error. However, since our goal is to predict *all* cases as accurately as possible, the original formulation is considered superior. As expected, the recurrent results for the least 1-norm estimation are best, while that method does worst on the non-recurs and has the worst overall mean error. This is attributable to the fact that most of the observed recurrences were at relatively short times (the mean recurrence time was 24 months), therefore a regression method which uses only the recurrent cases skews the predictions downward, matching the bias of this particular data set.

While others have applied machine learning techniques to the prognosis problem [4, 29] we are unaware of any other prognostic system that is capable of predicting cancer recurrence with an average error of 24 months. The most widespread statistical method of Cox proportional hazards regression [5] is unable to predict time of recurrence for this data. We plan to perform comparisons with the Cox method by randomly censoring other prognostic data sets.

# 6 Clinical Usage of the RSA Method

Both medical and personal decisions hinge on the projected future course of the breast cancer. Decisions whether or not chemotherapy is needed and the intensity of such therapy are based on the anticipated course of the cancer. The mental state of the patient, and personal and career plans are greatly affected by the anticipated course of the disease. Hence, improved prognostic prediction is an important goal for cancer treatment.

The best predictive model using the RSA formulation (Equation 4) has been added to the **Xcyt** program. We determined (using cross-validation) that five input features were needed for this prediction task. We then used the above feature-selection scheme to pare down the original set of 32 features to five: mean value of radius, perimeter, and fractal dimension, and extreme value of perimeter and area. Using this model, we now predict a time of recurrence for patients who have been diagnosed with a malignant tumor, to aid the choice of treatment. Further, **Xcyt** provides estimates of the patient's probability of disease-free survival in the form of a Kaplan-Meier curve [11] (Figure 4). The disease-free-survival curve for the individual patient is based on those training cases that had a similar predicted time of recurrence as determined by the RSA (4).

In current traditional medical practice, the strongest available prognostic feature is the extent to which cancer is present in the lymph nodes, which is determined by microscopic examination of lymph nodes that must be surgically removed from the patient's armpit. This procedure leaves the patient more susceptible to infection and the arm frequently develops lymphedema, a potentially severe swelling of the arm. Additionally, prognostic determinations based on lymph node involvement are in-

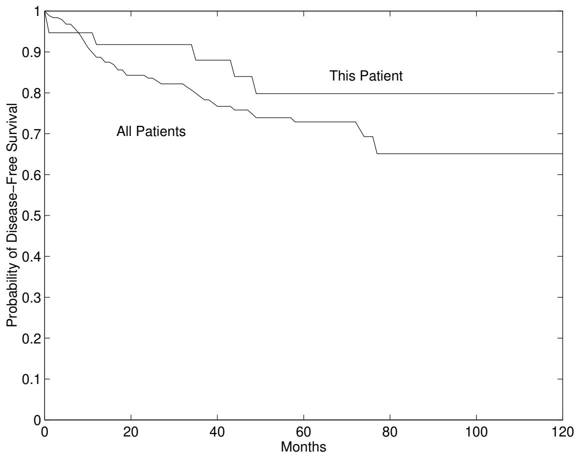


Figure 4: Estimated probability of disease-free survival, up to ten years following surgery, for all patients in the study and for the particular case shown in Figure 1.

accurate: 10% of patients in the most favorable category will die of breast cancer and 40% of those in the most unfavorable category will survive. Therefore, one of the most important findings of this research is that the best predictive models are found using just the nuclear features, and are not improved by the traditional medical prognostic factors of tumor size and lymph node status. If further studies confirm these findings, the routine and potentially hazardous removal of lymph nodes from the armpit of breast cancer patients for prognostic purposes can be avoided.

# 7 Extensions to the Recurrence Surface Approximation

Experiments have been performed with a number of extensions and variations of the RSA idea, some of that address the limitation of building a linear predictive surface. One of these, termed implicit RSA (IRSA), builds on the data preprocessing used by Ravdin and colleagues [6, 27, 28]. Time is added as an input feature, and given values along the range of follow-up times in the study. Each case thus produces a number of new training examples which are identical except for the time feature. In our implementation, a new training example is created for each six month interval, producing points with time equal to 6 months, 12 months, etc. For any particular time, the training case can be given a classification of recur (R) or non-recur (N), based on the patient's status at that point in time. A single separating plane (the "implicit" recurrence surface) is constructed between R and N cases in the $features \times time$ space. For a new case, TTR is predicted by holding all features fixed and varying time. The point at which this line meets the separating plane is the value of time for which this particular case would go from being classified "non-recur" to being classified "recur". Hence, that time can be interpreted as a predicted TTR. Specifically, for a separating plane $x^T w = \gamma$ in the $n+1$ dimensional space, we solve the equation

$$TTR = \frac{\gamma - \sum_i w_i x_i}{w_{n+1}}. \tag{6}$$

The implicit RSA procedure can be given further predictive power by adding nonlinear functions of time as new input variables. We have added $time^2$, $\frac{1}{time}$, and $\frac{1}{time^2}$, and observed predictive accuracy similar to that of RSA.

# 8 Conclusion and Future Work

We have shown how linear programming is used in actual clinical diagnosis and prognosis of breast cancer. By applying a linear-programming-based classification method, we have constructed a diagnostic system that performs at an accuracy level at or above any procedure short of surgery. The system also gives a probability of malignancy that allows the patient to compare the specific diagnosis with hundreds of previous cases. Through collaborative studies, this methodology will be employed by other researchers and applied to different types of cancer. We have also developed and applied a method for prognostic prediction, that provides accurate, patient-specific predictions of when a cancer is likely to recur. Because it used censored data to build a predictive survival model, this method is applicable to many different fields. The potential for applying these same approaches to other medical decision making, prediction and machine learning problems appears to be extremely promising and is worthy of further investigation and testing.

# References


[1] K. P. Bennett. Decision tree construction via linear programming. In *Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference*, pages 97–101, 1992.

[2] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly in-

separable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Inc., Pacific Grove, CA, 1984.

[4] H. B. Burke. Artificial neural networks for cancer research: Outcome prediction. *Seminars in Surgical Oncology*, 10:73–79, 1994.

[5] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, B 34:187–202, 1972.

[6] M. De Laurentiis and P. M. Ravdin. A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77:127–138, 1994.

[7] S. W. Fletcher, W. Black, R. Harris, B. K. Rimer, and S. Shapiro. Report of the international workshop on screening for breast cancer. *Journal of the National Cancer Institute*, 85:1644–1656, 1993.

[8] R. W. M. Giard and J. Hermans. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. *Cancer*, 69:2104–2110, 1992.

[9] F. Glover. Improved linear programming models for discriminant analysis. *Decision Sciences*, 21:771–785, 1990.

[10] R. C. Grinold. Mathematical methods for pattern classification. *Management Science*, 19:272–289, 1972.

[11] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[12] P. Lachenbruch and P. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11, 1968.

[13] E. T. Lee. *Statistical Methods for Survival Data Analysis.* John Wiley and Sons, New York, 1992.

[14] O. L. Mangasarian. Multi-surface method of pattern separation. *IEEE Transactions on Information Theory*, IT-14:801–807, 1968.

[15] O. L. Mangasarian. Mathematical programming in neural networks. *ORSA Journal on Computing*, 5:349–360, 1993.

[16] O. L. Mangasarian and R. R. Meyer. Nonlinear perturbation of linear programs. *SIAM Journal on Control and Optimization*, 17:745–752, 1979.

[17] O. L. Mangasarian, R. Setiono, and W. H. Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. In *Proceedings of the Workshop on Large-Scale Numerical Optimization*, pages 22–31, Philadelphia, Pennsylvania, 1990. SIAM.

[18] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 23:1 & 18, 1990.

[19] E. Marshall. Search for a killer: Focus shifts from fat to hormones in special report on breast cancer. *Science*, 259:618–621, 1993.

[20] B. A. Miller, E. J. Feuer, and B. F. Hankey. Recent incidence trends for breast cancer in women and relevance of early detection: An update. *CA Cancer Journal for Clinicians*, 43(1):27–41, 1993.

[21] C. Muier, J. Waterhouse, T. Mack, J. Powell, and S. Whelan. *Cancer Incidence in Five Continents*, volume 5. International Agence for Research on Cancer, Lion, France, 1987.

[22] B.A. Murtagh and M.A. Saunders. MINOS 5.0 user's guide. Technical Report SOL 83.20, Stanford University, December 1983. MINOS 5.4 Release Notes, December 1992.

[23] National Center for Health Statistics, GPO. *Vital Statistics of the United States, Mortality*, volume 2. DHHS, 1990.

[24] E. Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[25] W. L. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes.* Cambridge University Press, Cambridge, 1986.

[26] J. Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[27] P. M. Ravdin and G. M. Clark. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22:285–293, 1992.

[28] P. M. Ravdin, A. K. Tandon, D. C. Allred, G. M. Clark, S. A. W. Fuqua, S. H. Hilsenbeck, G. C. Chamness, and C. K. Osborne. Cathepsin D by western blotting and immunohistochemistry: Failure to confirm correlations with prognosis in node-negative breast cancer. *Journal of Clinical Oncology*, 12:467–474, 1994.

[29] A. Schenone, L. Andreucci, V. Sanguinetti, and P. Morasso. Neural networks for prognosis in breast cancer. *Physica Medica*, IX(Supplement 1):175–178, June 1993.

[30] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society (Series B)*, 36:111–147, 1974.

[31] W. N. Street. *Cancer Diagnosis and Prognosis via Linear-Programming-Based Machine Learning.* PhD thesis, University of Wisconsin-Madison, August 1994. Available as University of Wisconsin Mathematical Programming TR 94–14.

[32] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, San Jose, California, 1993.

[33] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A.*, 87:9193–9196, 1990.

[34] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404, December 1993.

[35] W. H. Wolberg, W. N. Street, and O. L. Mangasarian. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77:163–171, 1994.