A Neural Network Model for Prognostic Prediction

W. Nick Street
Computer Science Department
Oklahoma State University
Stillwater, OK 74078
`nstreet@cs.okstate.edu`

Abstract

An important and difficult prediction task in many domains, particularly medical decision making, is that of prognosis. Prognosis presents a unique set of problems to a learning system when some of the outputs are unknown. This paper presents a new approach to prognostic prediction, using ideas from nonparametric statistics to fully utilize all of the available information in a neural architecture. The technique is applied to breast cancer prognosis, resulting in flexible, accurate models that may play a role in preventing unnecessary surgeries.

1 Introduction

This paper applies artificial neural network classification to the analysis of *survival* or *lifetime data* (Lee, 1992), in which the objective can be broadly defined as predicting the future time of a particular event. In this work we are concerned specifically with prognosis, that is, predicting the course of a disease. These methods are applied to breast cancer prognosis, predicting how long after surgery we can expect the disease to recur. This problem has significant clinical importance. Decisions regarding chemotherapy its intensity are based on the anticipated course of the cancer. For example, patients with favorable outlooks may forego chemotherapy entirely. Those with less favorable outlooks may undergo varying intensities of chemotherapy, or even bone marrow transplantation.

Prognostic prediction does not fit comfortably into either of the classic learning paradigms of function approximation or classification. While a patient can be classified "recur" if the disease is observed, there is no real cutoff point at which the patient can be considered a non-recurrent case. The data are therefore *censored* in that we know a time to recur for only a subset of patients. For the others, we know only the time of their last check-up, or disease-free survival time (DFS). In particular, recurrence or survival data is *right censored*, i.e., the right endpoint (recurrence time) is sometimes unknown, since some patients will inevitably move away, change doctors, or die of unrelated causes. Therefore, in many cases, the training signal for the learning method is not well-defined. Prognosis is not viewed here as a time-series prediction problem, since the predictive features are gathered only once, at the time of diagnosis and/or surgery.

Problems involving censored data are common to several fields. In engineering, one might be interested in the survival characteristics of electronic components, while sociologists might consider what factors lead to long-lasting marriages. These problems have traditionally been approached using statistical techniques such as Cox proportional-hazards regression (Cox, 1972). In recent years, there has been an increased interest in the application of machine learning methods to prediction using censored data. Several groups have approached prognosis as a separation problem using different learning architectures, including backpropagation artificial neural networks (ANNs) (Burke, 1994; Burke et al., 1997), entropy maximization networks (Choong et al., 1996) and decision trees (Wolberg et al., 1992; Wolberg et al., 1994). This is done by choosing one or more endpoints and learning a yes/no classifier on concepts such as "patients who recurred in less than two years." Cases with follow-up time less than the cutoff are discarded from the training set. Ravdin and colleagues (De Laurentiis and Ravdin, 1994; Ravdin and Clark, 1992) use ANNs to generate *survival curves*, which plot the probability of disease-free survival against time. This work uses time as an input variable and interprets the trained network's single output as an approximation of recurrence probability. The resulting formulation results in biases in the training data that must be corrected by repeating or removing some of the examples. Their computational results are verified only by demonstrating that their predicted survival rates closely approximate those of the test cases. The problem has also been approached in an unsupervised learning fashion, using clustering (Bradley et al., 1997) and self-organizing neural networks (Schenone et al., 1993). However, these techniques did not directly address the problem of prediction using censored data.

While this research also separates the cases into classes based on recurrence time, it differs from the above techniques in several respects. Censored cases are incorporated directly into the training set, not by using an artificial cutoff time, but rather by using the probability that they will recur before a certain time as the training signal. In this way we use all of the information available in the training set. Further, interpreting the outputs as probabilities lets us not only separate the cases into "good" and "bad" prognoses, but also to generate predicted survival curves for individual patients, making the system more useful in a clinical setting.

2 Neural Architecture

The ANNs used in this work were standard feedforward networks with one hidden layer, trained with backpropagation (Rumelhart et al., 1986). The hyperbolic tangent activation function was used for hidden and output nodes. The output layer consisted of ten units; the first represented the class of examples with recurrences at one year or less following surgery, the second those with recurrences between one and two years, etc., up to ten years[1]. This approach implies the existence of an extra (in our case, eleventh) class. These are the patients with expected disease-free survival of time greater than the length of the study (10 years). The activations of the output units were trained with and interpreted as the probability that the patient would have disease-free survival up to that time. These probabilities were scaled to the range of the hyperbolic tangent function, i.e., *activation = 2 * probability - 1*.

In order to maintain the interpretation of the outputs as probabilities, the *relative entropy* error function (Baum and Wilczek, 1988; Solla et al., 1988) was used for all non-input units. For a given example i, this error function is defined as

$$E_i = \sum_k \left[\frac{1}{2}(1 + T_i^k) \log \frac{1 + T_i^k}{1 + O_i^k} + \frac{1}{2}(1 - T_i^k) \log \frac{1 - T_i^k}{1 - O_i^k} \right],$$

where T_i^k is the target value for output unit k and O_i^k is its output value. Outputs of +1 and -1 correspond to definitely true and definitely false, respectively, with intermediate values again being scaled into the appropriate range.

For recurrent cases, the network was trained with values of +1 for all outputs up to the observed recurrence time, and -1 thereafter. For instance, a recurrence at 32 months would have a training vector $T = \{1, 1, -1, -1, -1, -1, -1, -1, -1, -1\}$. The value of the probability formulation is seen in the censored cases. They were similarly trained with values of +1 up to the observed disease-free survival time. The probabilities of DFS for later times were computed using a variation of the standard Kaplan-Meier maximum likelihood approximation to the true population survival rate (Kaplan and Meier, 1958). We define the *risk* of recurrence at time $t > 0$ as the conditional probability that a patient will recur at time t, given that they have not recurred up to time $t-1$. As an example, consider a study containing a total of 20 patients. If two recurrences were observed in the first time interval, we would have $risk_1 = 0.1$. Further suppose that the study has two censored cases in the first time interval, and two more recurrences in the second interval. There are 16 patients at risk for recurrence during interval two, with two recurrences, so $risk_2 = 0.125$. The Kaplan-Meier estimator of the *disease-free survival curve*, S, tracks the cumulative probability of DFS for any time in the study, using the risks in the following fashion:

$$S_t = \begin{cases} 1, & t = 0 \\ S_{t-1}(1 - risk_t), & t > 0. \end{cases}$$

Continuing the above example, $S_0 = 1.0$, $S_1 = 0.9$, and $S_2 = 0.7875$. To compute appropriate training probabilities, we simply use the DFS time of the censored case as the starting time, rather than time 0:

$$S_t = \begin{cases} 1, & 0 \le t \le DFS(i) \\ S_{t-1}(1 - risk_t), & t > DFS(i). \end{cases}$$

For an individual output node k, this training signal represents the example's probability of membership in

[1] The available prognostic studies are approximately ten years in duration.

the class being recognized by that node, i.e., the set of cases that recur before the end of year k. Collectively, the activation values of the output units represent an expected survival curve for the individual case.

If we view the network as learning a survival curve, the task becomes one of function approximation using incomplete data. The training signal is then a modified thermometer encoding (McCullagh and Nelder, 1989), a relatively common encoding for ordered categorical outputs, with the added complication of the survival probabilities for censored cases. Since the effects of some of the input features are thought to be nonlinear over time, it is also instructive to view the problem as a sequence of highly related but distinct classification problems, all learned using the same internal representation (i.e., hidden nodes). The representation generated in learning one group (say, those cases that are likely to recur before one year) contributes to the learning of other groups (say, those cases recurring between 5 and 6 years). This is a form of functional knowledge transfer, similar to the MTL network (Baxter, 1995; Caruana, 1995). The learning of multiple classes in parallel contributes to faster learning and more reliable predictive models.

The above architecture facilitates three different uses of the resulting predictive model:

1. The output units can be divided into groups *a posteriori* to separate good from poor prognoses. For a particular application, any prediction of recurrence at a time greater than five years might be considered favorable, and indicate less aggressive treatment. The actual outcomes of those patients in the good group should be significantly better than those in the poor group.

2. An individualized disease-free survival curve can easily be generated for a particular patient by plotting the probabilities predicted by the various output units. In order for this curve to be reliable, the activations should be monotonically decreasing, or very nearly so.

3. The expected time of recurrence can be obtained merely by noting the first output unit that predicts a probability of disease-free survival of less than 0.5. This provides a convenient method of rank-ordering the cases according to expected outcome.

A significant methodological issue is that of evaluating the learned model. As discussed earlier, this is neither a function approximation nor a classification problem, since in many cases we do not know the correct answer. Still, there is a well-defined goal: the accurate prediction of individual prognosis. While our training method seeks to minimize the relative entropy error at each output unit, the reporting of this error on testing sets would be relatively uninformative. We therefore evaluate the models on two criteria: the accuracy of the predicted recurrence rates (see Section 3.4) and the ability of the models to separate cases with favorable and unfavorable prognoses (see Section 3.3).

3 Experimental Results

Computational experiments were performed on two very different breast cancer data sets. The first is known as Wisconsin Prognostic Breast Cancer (WPBC) and is characterized by a small number of cases, relatively high dimensionality, very precise values and almost no missing data. The second data set is from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute. It contains a large number of cases, with relatively few, coarsely-measured features, and a high percentage of missing values. Details on these data sets are given below.

In both cases, the prognosis data used in this study consists of those malignant patients for which follow-up data was available, after eliminating those cases with distant metastasis (cancer has already spread; prognosis is poor) and carcinoma *in situ* (cancer has not yet invaded breast tissue; prognosis is good). We therefore maximize the clinical relevance of the study by focusing on those cases that present the most difficult prognosis.

Experiments reported in this section are test set results using either tenfold cross-validation (WPBC data) or a single randomized holdout test (SEER data). The ANNs used had three hidden units, and training was terminated after 1000 on-line epochs.

3.1 Wisconsin Prognostic Breast Cancer Data

In previous work (Mangasarian et al., 1995; Wolberg et al., 1994) the author contributed to the development of an image-processing software package for breast cancer diagnosis, known as Xcyt, which analyzes digital images of cells taken from breast lumps. This program computes 10 different features of each cellular nuclei in the image: radius, perimeter, area, compactness, smoothness, size and number of concavities, symmetry, fractal dimension, and texture. The mean, standard area, and extreme values of each feature are computed for each image. The current application uses the 30 nuclear features computed by Xcyt together with two traditional prognostic predictors: tumor size and number of involved lymph nodes. This data set contains 227 cases, 61 of which have recurred. An earlier version of this data set is available at the UCI machine learning repository (Merz and Murphy, 1996).

3.2 SEER Data

The SEER (Carter et al., 1989) data set consists of data on cancer survival (rather than recurrence) for over 38,000 women newly diagnosed with breast cancer between 1977 and 1982. Each case contains the following information: histological grade (four discrete values), tumor size, tumor extent (5 discrete values), number of positive lymph nodes, and number of nodes examined. Many of these feature values are missing. For instance, only about 20% of the cases contain a value for histological grade; over 1200 of the cases contain no feature information at all. Each of the SEER features was encoded as a sequence of binary variables, with an additional binary variable representing a missing value.

3.3 Good vs. poor prognoses

To be used as a clinical tool, the predictive model should reliably separate cases with a good prognosis from those with a poor prognosis. Since treatment options are limited, this sort of stratification could be most helpful to the physician and the patient in determining a post-operative treatment plan. Figure 1 stratifies the WPBC test cases into those predicted to recur in the first five years and those predicted to recur at some time greater than five years (including the implicit 11th class). The difference in these two groups is statistically significant ($p < 0.001$, generalized Wilcoxon test). Of course, the output units could be grouped differently to define the relevant prognostic categories for a particular problem. Further, the implicit final group could also be subdivided based on the activation level of the last node.

Similarly, Figure 2 shows survival probabilities for those cases with good and poor prognosis, in this case, predicted survival less than or equal to ten years and predicted survival greater than ten years. Again the difference in the two groups is statistically significant ($p < 0.001$). The difference in dividing points between

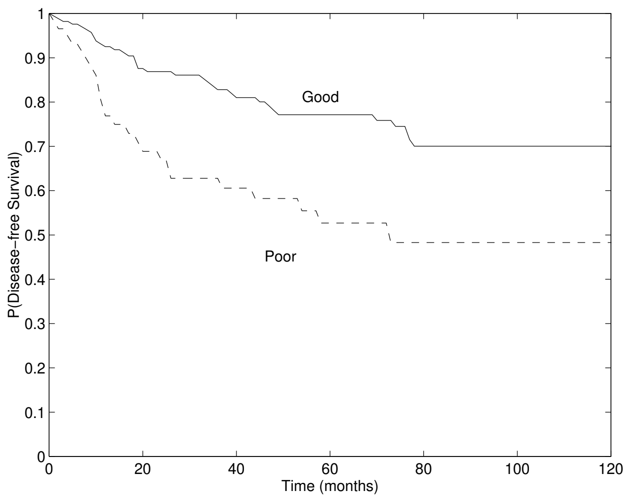

Figure 1: WPBC Data: Disease-free survival probabilities for those cases predicted to recur in the first five years (Poor, 58 cases) compared to those predicted to recur at some time greater than five years (Good, 169 cases).

the two tests is due to the difference between the measured endpoints (recurrence in the WPBC data, death in the SEER data). The ratios of good to bad prognoses were held nearly constant.

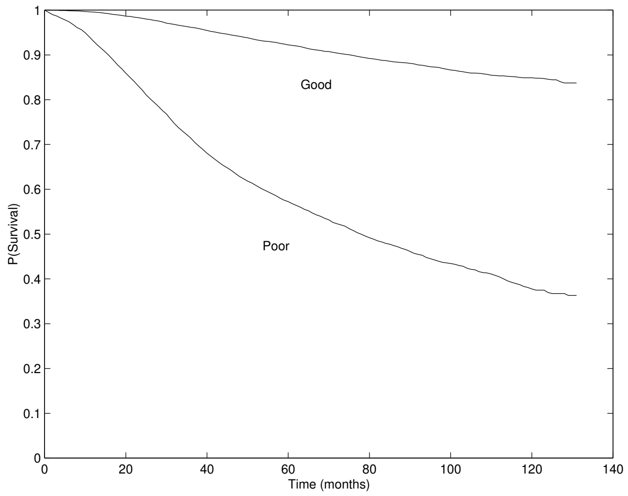

Figure 2: SEER Data: Survival probabilities for those cases predicted to die from breast cancer in the first ten years (Poor, 8,353 cases) compared to those predicted to die at some time greater than ten years (Good, 26,192 cases).

In traditional breast cancer staging, postoperative treatment decisions are based largely or even entirely on whether or not the cancer has spread to the patient's axillary lymph nodes. However, removing the nodes for examination leaves the arm subject to infection and possible lymphedema (Aitken et al., 1989), and does not affect overall survival (Abe et al., 1995). In both of our test cases, the separation with this method represents an improvement over that achieved by the lymph node status feature. Further, statistically significant separation was achieved in both data sets without using the lymph feature (WPBC, $p = 0.02$; SEER, $p < 0.001$). This is further confirmation of a previous finding (using other analytic techniques) that breast cancer prognosis can be achieved without lymph node dissection (Wolberg et al., 1997; Wolberg et al., 1998).

3.4 Predicted vs. actual group survival

Another criterion for the validity of the learned model is whether the predicted recurrence rate follows that of the actual data. Figure 3 shows the Kaplan-Meier estimate of disease-free survival curve for the entire WPBC training set, compared with the predicted DFS rates accumulated from the test folds. Again, a test case is predicted to recur at time t if the activation of output node t is the first one indicating a DFS probability of less than 0.5. The two curves are very similar and show no significant statistical difference ($p = 0.2818$, generalized Wilcoxon test (Gehan, 1965)).

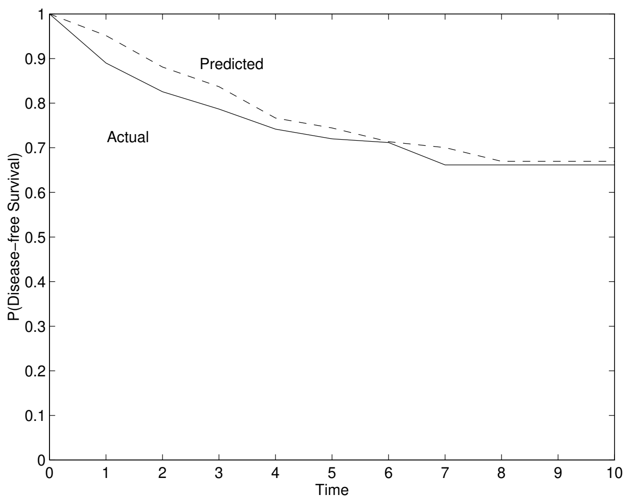

Figure 3: WPBC Data: Kaplan-Meier estimate of true disease-free survival curve compared to predicted DFS curve.

The predicted group survival for the SEER data did not closely match the actual survival curve. This is consistent with previous research (Street et al., 1996) using a variation of the RSA prognostic technique (Street et al., 1995) which also showed that the SEER data was unable to replicate group survival characteristics. This is attributable to the coarse encoding of the SEER variables and the large percentage of missing values.

3.5 Individual prognostic prediction

As mentioned, the activations of the output units can be combined to form a predicted DFS curve for an individual patient. Figure 4 shows an example of this usage, in a format appropriate for a clinical setting. Here the probabilities of disease-free survival for a case from the WPBC study are compared to the cumulative values of all patients in the study. The output activations were monotonically non-increasing, as was the case in 74% of the WPBC test examples and 87% of the SEER examples. The others had occasional small increases, with the maximum increase in any example corresponding to to a probability change of 0.024 (WPBC) and 1.0 (SEER). The expected time of recurrence can be computed by noting where the DFS curve crosses a probability of 50%, in this case, between three and four years. In fact, this patient did experience disease recurrence in the 44th month following surgery.

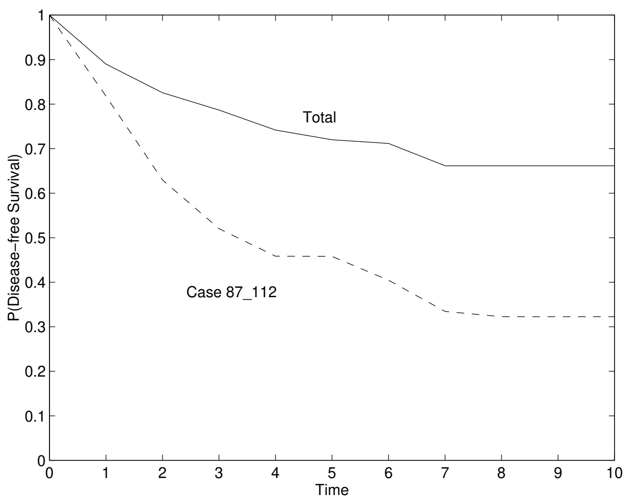

Figure 4: Predicted DFS curve of a single case (87_112) from the WPBC study compared to the overall group DFS curve of the training set.

4 Conclusions

This paper develops a novel encoding of censored data in an artificial neural network architecture to provide a framework for prognostic prediction. In applying the method to breast cancer prognosis, the resulting models are shown to be at least as accurate as current methods, while providing significantly more precision and flexibility. Among the future directions for this research is a sensitivity analysis to investigate the importance of the prognostic features at different follow-up times. To evaluate the role of knowledge transfer, predictive accuracy will be compared to classification models that predict recurrence at a chosen cut point. Most importantly from a clinical perspective, our work in the breast cancer domain continues to focus on generating accurate prognostic models without knowledge of lymph node status, in order to spare new patients an extra and potentially debilitating surgery.

Acknowledgments

The author wishes to thank Dr. William H. Wolberg of the University of Wisconsin Department of Surgery for contributing both data and direction to this research. The SEER data was made available by Dr. Don Henson of the National Cancer Institute. This work was partially supported by NSF grant IRI-9701992.

References

Abe, O., Abe, R., Asaishi, K., Enomoto, K., Hattori, T., and Iino, Y. (1995). Effects of radiotherapy and surgery in early breast cancer: An overview of the randomized trials. *New England Journal of Medicine*, 333:1444–1455.

Aitken, R. J., Gaze, M. N., Rodger, A., Chetty, U., and Forrest, A. P. M. (1989). Arm morbidity within a trial of mastectomy and either nodal sample with selective radiotherapy or axillary clearance. *British Journal of Surgery*, 76:568–571.

Baum, E. B. and Wilczek, F. (1988). Supervised learning of probability distributions by neural networks. In Anderson, D. Z., editor, *Neural Information Processing Systems*, pages 52–61, New York. American Institute of Physics.

Baxter, J. (1995). Learning internal representations. In *Proceedings of the Eighth International Conference on Computational Learning Theory*.

Bradley, P. S., Mangasarian, O. L., and Street, W. N. (1997). Clustering via concave minimization. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 368–374, Cambridge, MA. MIT Press.

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

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E., Marks, J. R., Winchester, D. P., and Bostwick, D. G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79:857–862.

Carter, C. L., Allen, C., and Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63:181–187.

Caruana, R. (1995). Learning many related tasks at the same time with backpropagation. In *Advances in Neural Information Processing Systems*, volume 7, pages 657–664.

Choong, P. L., deSilva, C. J. S., Dawkins, H. J. S., and Sterrett, G. F. (1996). Entropy maximization networks: An application to breast cancer prognosis. *IEEE Transactions on Neural Networks*, 7(3):568–577.

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

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

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, 52:203–223.

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

Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. John Wiley and Sons, New York.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, chapter 6, page 137. Chapman and Hall, London, 2 edition.

Merz, C. J. and Murphy, P. M. (1996). UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. University of California, Irvine, Department of Information and Computer Sciences.

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

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing*, volume 1, chapter 8. MIT Press, Cambridge, MA.

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

Solla, S. A., Levin, E., and Fleisher, M. (1988). Accelerated learning in layered neural networks. *Complex Systems*, 2:625–639.

Street, W. N., Mangasarian, O. L., and Wolberg, W. H. (1995). An inductive learning approach to prognostic prediction. In Prieditis, A. and Russell, S., editors, *Proceedings of the Twelfth International Conference on Machine Learning*, pages 522–530, San Francisco. Morgan Kaufmann.

Street, W. N., Mangasarian, O. L., and Wolberg, W. H. (1996). Individual and collective prognostic prediction. Technical Report 96–01, Computer Sciences Department, University of Wisconsin, Madison, WI. Available from ftp://ftp.cs.wisc.edu/math-prog/tech-reports/.

Wolberg, W. H., Bennett, K. P., and Mangasarian, O. L. (1992). Breast cancer diagnosis and prognostic determination from cell analysis. Manuscript, Departments of Surgery and Human Oncology and Computer Sciences, University of Wisconsin.

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

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1997). Computer-derived nuclear features compared with axillary lymph node status for breast cancer prognosis. *Cancer Cytopathology*, 81:172–179.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1998). A comparison of computer-based nuclear analysis versus lymph node status for staging breast cancer. *Lancet*. submitted.