GRANT NO: DAMD17-94-J-4383

TITLE: Developing and Implementing the AJCC Prognostic System for Breast Cancer

PRINCIPAL INVESTIGATOR(S): Dr. Philip H. Goodman
Dr. Harry Burke

CONTRACTING ORGANIZATION: University of Nevada
Reno, Nevada 89557

REPORT DATE: August 1995

TYPE OF REPORT: Annual

PREPARED FOR:
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.
The TNM staging system for breast cancer has been in existence for over 35 years in America. Within the last ten years it has become clear that: (1) the TNM staging system is not highly accurate, and (2) if new breast cancer prognostic factors are to be integrated with the TNM variables to increase outcome prediction accuracy, a new prognostic system is required.

The goal of this project is the creation of a computer-based prognostic system for breast cancer that: (1) is significantly more accurate than the TNM staging system, (2) predicts survival over time based on therapy, (3) and presents its predictions in a manner that physicians can understand.

During the first year of the project, we made substantial progress on all 3 major tasks. We believe that we will be able to successfully meet our ultimate goal of providing a computer-based prognostic system that is more accurate than the TNM staging system and that is easy to use and understand, within the four year time frame of this grant.
**GENERAL INSTRUCTIONS FOR COMPLETING SF 298**

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to stay within the lines to meet optical scanning requirements.

| Block 1. **Agency Use Only (Leave blank)** |
| Block 2. **Report Date**: Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year. |
| Block 3. **Type of Report and Dates Covered**: State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88). |
| Block 4. **Title and Subtitle**: A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses. |
| Block 5. **Funding Numbers**: To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels: |
| C - Contract | PE - Project |
| G - Grant | TA - Task |
| PE - Program | WU - Work Unit |
| Element | Accession No. |
| Block 6. **Author(s)**: Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s). |
| Block 7. **Performing Organization Name(s) and Address(es)**: Self-explanatory. |
| Block 8. **Performing Organization Report Number**: Enter the unique alphanumeric report number(s) assigned by the organization performing the report. |
| Block 9. **Sponsoring/Monitoring Agency Name(s) and Address(es)**: Self-explanatory. |
| Block 10. **Sponsoring/Monitoring Agency Report Number (If known)** |
| Block 11. **Supplementary Notes**: Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report. |
| Block 12a. **Distribution/Availability Statement**: Denotes public availability or limitations. Enter any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR). |
| DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
| DOE - See authorities.
| NTIS - Leave blank. |
| Block 12b. **Distribution Code**: |
| DOD - Leave blank.
| DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
| NASA - Leave blank.
| NTIS - Leave blank. |
| Block 13. **Abstract**: Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report. |
| Block 14. **Subject Terms**: Keywords or phrases identifying major subjects in the report. |
| Block 15. **Number of Pages**: Enter the total number of pages. |
| Block 16. **Price Code**: Enter appropriate price code (NTIS only). |
| Blocks 17.-19. **Security Classifications**: Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page. |
| Block 20. **Limitation of Abstract**: This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited. |
FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

Where copyrighted material is quoted, permission has been obtained to use such material.

Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Advising for

RMS CHAI

DV12 TAN

Unannounced

Justification

By

Distribution

Availability Codes

Mail and/or

Stat Special

Accession For

PI - Signature

Date

8-29-95

Phillip H. Jordan
TABLE OF CONTENTS

INTRODUCTION

BODY: FIRST YEAR ACCOMPLISHMENTS
   Task 1. Data analysis and prognostic factor evaluation.
   Task 2. Developing the prognostic model.
   Task 3. Implementation of a clinically useful prognostic system

CONCLUSION

REFERENCES

Appendix A: PAPERS PRESENTED
Appendix B: PHYSICIAN SURVEY
INTRODUCTION

The TNM staging system for breast cancer has been in existence for over 35 years in America. Within the last ten years it has become clear that: (1) the TNM staging system is not highly accurate, and (2) if new breast cancer prognostic factors are to be integrated with the TNM variables to increase outcome prediction accuracy, a new prognostic system is required.

The goal of this project is the creation of a computer-based prognostic system for breast cancer that: (1) is significantly more accurate than the TNM staging system, (2) predicts survival over time based on therapy, (3) and presents its predictions in a manner that physicians can understand.

BODY: FIRST YEAR ACCOMPLISHMENTS

Task 1. Data analysis and prognostic factor evaluation.

1.01) Extend analysis of binary survival endpoint to 10 year survival.

We are currently analyzing SEER 10 year survival data. Preliminary results suggest that the predictors collected at disease discovery are less accurate in predicting 10 year survival than 5 year survival. It may be that predictors be collected at regular intervals after discovery and therapy, and these predictors be used to estimate 10 year survival. Additionally, conditional probability of survival can be calculated, i.e., if a woman survives for five years, what is her probability of living another five years.

1.02) Extend the analysis to recurrence as an endpoint.

We have analyzed recurrence as an end-point, see 1.08.1 for preliminary results.

1.03) Comparison of prognostic models.

We have compared statistical method to artificial neural networks in terms of five year breast cancer-specific survival. Selected results are shown below.

NCDB/PCE 1983 Breast Cancer Data Set.

<table>
<thead>
<tr>
<th>PREDICTION MODEL</th>
<th>ACCURACY*</th>
<th>SPECIFICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>pTNM Stages</td>
<td>.720</td>
<td>Ø, I, II, IIA, IIIB, IIIA, IIIB, IV</td>
</tr>
<tr>
<td>Principal Components Analysis</td>
<td>.714</td>
<td>one scaling iteration</td>
</tr>
<tr>
<td>CART, pruned</td>
<td>.753</td>
<td>9 nodes</td>
</tr>
<tr>
<td>CART, shrunk</td>
<td>.762</td>
<td>13.7 nodes</td>
</tr>
<tr>
<td>Stepwise Logistic Regression</td>
<td>.776</td>
<td>cubic splines</td>
</tr>
<tr>
<td>Fuzzy ARTMAP NN**</td>
<td>.738</td>
<td>54-F2a, 128-1</td>
</tr>
<tr>
<td>Cascade Correlation NN</td>
<td>.761</td>
<td>54-21-1</td>
</tr>
<tr>
<td>Conjugate Gradient Descent NN</td>
<td>.774</td>
<td>54-30-1</td>
</tr>
<tr>
<td>Probabilistic NN</td>
<td>.777</td>
<td>bandwidth = 16s</td>
</tr>
<tr>
<td>Backpropagation NN</td>
<td>.784</td>
<td>54-5-1</td>
</tr>
</tbody>
</table>

* The area under the curve of the receiver operating characteristic.
** NN, neural network.
These results were recently published.


Additional publications related to statistical methods in breast cancer are listed below.


1.04.2) Create a taxonomy of prognostic factors in breast cancer.

We have created a taxonomy of prognostic factors in breast cancer. The taxonomy was based on levels of analysis: demographic, anatomic/cellular, and molecular genetic. In addition, we collected, described, and cited the primary sources for the major breast cancer prognostic factors, of which there are over 76 at the current time (with a new putative prognostic factor reported almost every month). This work was recently published.


There are many problems and pitfalls in the discovery, analysis, and use of prognostic factors. A book, that uses breast cancer as its cancer site, will be helpful in prognostic factor research for all cancers. Although we have not yet reached an agreement with a publisher regarding a book on prognostic factors in breast cancer, we are actively publishing on the subject. (See also 1.03 for statistical publications.)


Burke HB. The future of the TNM staging system. In preparation.


Burke HB. Prognostic methods is cancer: a review. In preparation.

1.06.3) Determining minimum data set size.

We have found that for predicting five year breast cancer-specific survival, using currently collected prognostic factors and a 30% five-year breast cancer-specific mortality rate, we found that approximately 2,300 patients are required for maximum accuracy. More than 2,300 cases does not provide any improvement in prediction accuracy. We have not yet reported these results because we are working with Memorial Sloan-Kettering Cancer Center on extending it to molecular-genetic prognostic factors.

1.08.1) Recurrence analysis.

There are two types of recurrence analyses; predicting recurrence based on data at discovery, and predicting survival based on there having been a recurrence. We are primarily interested in predicting recurrence at discovery. In other words, at this time, we are interested in using recurrence as an endpoint rather than as a prognostic factor for survival.

The accuracy (area under the receiver operating characteristic curve) of the probability of recurrence predictions at three, four and five years, for those women who are alive at each time period, is .731, .714, and .701, respectively. We can make two observations regarding these results. (1) Based on our analysis of five year breast cancer-specific survival, predicting recurrence from data collected at the discovery of disease is less accurate than predicting survival from the same data. (2) Predictive accuracy declines as the prediction extends further into the future.

1.11) Patient information and physician credibility.

We have performed a small, preliminary survey of oncologists' assessment of five year breast cancer specific survival. The survey is preliminary because
we want to survey more oncologists, and we want to survey oncologic surgeons, pathologists, and radiation oncologists. Oncologists were asked to estimate ten patient's five year breast cancer specific survival. (Survey instrument is presented in Appendix B). The mean of the oncologist's predictions for each patient was compared with the patient's actual survival. (see Figure below)

Oncologists tended to be pessimistic regarding breast cancer patient prognosis. Since the therapy for patients with poor prognoses is different therapy from that of patients with a good prognosis, this finding (currently unpublished) has important implications for patient care.

As shown in 1.05 above, the TNM staging system's accuracy is .720. This is only approximately 44% better than chance in predicting whether a woman with breast cancer will survive five years. Our prognostic system is significantly more accurate than the TNM staging system. This result has been submitted for publication.


Task 2. Developing the prognostic model.

2.1) Survival curves for individual patients and for groups of patients:

2.1.1) Generate survival curves for 10 year data.

Our work on 5 year survival is directly applicable to 10 year survival.

2.1.2) Generating survival curves.
We have compared the accuracy of the major statistical methods to artificial neural networks in predicting cancer-specific five-year survival for breast cancer (1.03). We extend this work by creating an artificial neural network that predicts individual patient survival over time (survival curves).

Ideally, a model that estimates the probability of survival over time should:
1. Generate an accurate estimate of an individual patient's probability of survival over time, i.e., an accurate survival curve.
2. Accommodate censored cases (with a minimum of assumptions).
3. Capture the predictive power of nonlinear and interacting prognostic factors.
4. Allow prognostic factors to have different effects on the probability of survival over time, i.e., proportional hazards is not assumed.

The methods discussed below accommodate censored cases, some more successfully than others.

The Kaplan-Meier method (Kaplan, 1958) is a descriptive method for prediction over time based on covariate "bins". Bins can range from one, all patients, to a bin for each covariate or level of covariate. The Kaplan-Meier can accommodate censored cases, and, like most methods that accommodate censoring, its accuracy can suffer as censoring increases because there are fewer cases to base prediction upon. The Kaplan-Meier can be less accurate than inferential models because it assumes independence, whereas most inferential models only assume conditional independence (any dependence is explained by the covariates). The Kaplan-Meier's problems are those of a bin model, including: an exponential increase in the number of bins as the number of covariates increase, it loses information by requiring that continuous variables be cut into ranges, and there is no optimization strategy for finding the most accurate combination of bins.

The Cox proportional hazards model (Cox, 1972) is a linear effects model. It estimates the importance of each covariate, and it handles censored cases. It assumes proportional hazards and it does not provide a survival curve without the imputation of a baseline survival curve.

Faraggi and Simon (1994) nest an artificial neural network in the Cox proportional hazards model, replacing the linear combination of covariates with an artificial neural network. This solves the problem of capturing nonlinear and interactional covariates, while handling censored cases. As an artificial neural network generalization of the Cox proportional hazards model, it retains the assumption of proportional hazards and it does not provide a survival curve unless a baseline survival curve is imputed.

The simplest approach to a full artificial neural network implementation of a probability of survival over time model is to create a artificial neural network for each time interval. Data would be time interval specific; the censored cases would be dropped from the analysis, i.e., not included in the subsequent time interval artificial neural networks, at the time of censoring. Survival probabilities can be generated by each time-interval-specific artificial neural network, and they can be multiplied in succession to provide a survival prediction for each time interval. A problem with this approach is that the information contained in variables over several time periods is lost, because each time period is a separate artificial neural network. One artificial neural network spanning all time intervals partially solves this problem. This approach, with a two layer neural network, is similar to a series of logistic regression models, one for each time interval. (Cox vs. LR comparison here).
Ravdin and Clark (1992), provide the earliest attempt to create a probability of survival artificial neural network. Employing a commercial artificial neural network, Ravdin and Clark generate a prognostic index, which is roughly proportional to the survival probability, which they stratify into four groups by predicted prognosis. They code time as an input variable, each patient's data is reproduced for each time interval, in order to represent censored outcomes. Thus, for four time intervals there are four representations of each patient, with each representation differing only in its time interval failure information, i.e., outcome status (alive/dead), and censored status. Ravdin and Clark drop censored cases from the analysis at the time interval at which censoring occurs. Since only alive or dead remain in the analysis, as time continues, the ratio of dead to alive increases dramatically, resulting in too many patients dead and too few patients alive in the later time intervals. In order to rectify this imbalance, at each time interval the authors use the Kaplan-Meier product-limit estimate to determine the overall ratio of survivor to nonsurvivor. They use this ratio, based on the independence assumption, to determine the number of dead to randomly remove from the study in later time intervals. But the Kaplan-Meier estimate is itself sensitive to censoring, and the independence assumption must be justified. When faced with this situation, a better response might be to use the predictors to determine who to remove from the study. Also, throwing out patients removes predictive information from the study.

Liestold and Anderson (1994) create an artificial neural network that estimates the probability of survival over time. Their model creates one artificial neural network, and represents each time interval as a separate output node. Each output node generates a conditional survival probability. A possible problem with generating conditional survival probabilities is that the error of each prediction (variance) may accumulate when the predictions are multiplied together to create the survival estimate over time. Further, there is the problem of equal training of the nodes resulting in unequal accuracy, as some nodes are overfitted and some underfitted. Although their model retains the proportional hazards assumption, they suggest stratifying the covariates in order to remove this assumption. The authors go on to add a penalty term to the model, to penalize for deviations from proportionality.

We have implemented a new artificial neural network, one that achieves the four objectives stated above for estimating survival over time. We have also created a Windows-based interface for online prediction of individual patient survival. (see below)

<table>
<thead>
<tr>
<th>Year</th>
<th>ANN Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td>1</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>0.992</td>
</tr>
<tr>
<td>3</td>
<td>0.981</td>
</tr>
<tr>
<td>4</td>
<td>0.962</td>
</tr>
<tr>
<td>5</td>
<td>0.911</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor:</td>
</tr>
<tr>
<td>T Size:</td>
</tr>
<tr>
<td>LN Pos:</td>
</tr>
<tr>
<td>LN Exam:</td>
</tr>
<tr>
<td>Mets:</td>
</tr>
<tr>
<td>ER:</td>
</tr>
<tr>
<td>PR:</td>
</tr>
<tr>
<td>Menopausal:</td>
</tr>
<tr>
<td>Grade:</td>
</tr>
<tr>
<td>Age:</td>
</tr>
<tr>
<td>LN pTN:</td>
</tr>
</tbody>
</table>

T NM Stage: IIA  Prediction: 0.876
The figure at the right is an individual patient's five year breast cancer-specific survival curve, based on the ten prognostic factors shown in the upper right. At the left are the numerical values associated with the curve at each year. At the bottom is a computerization of the TNM staging system. Given the TNM variables, it automatically generates the TNM stage. In addition, it also presents the probability of five year breast cancer-specific survival associated with that stage. This work has been published.


2.1.3) Determining the accuracy of the survival curves.

The accuracy of predicted survival curves, with respect to the actual times of death (or of censoring) of the patients in a data set, can be evaluated in terms of accuracies of the survival or hazard probabilities at each point in time. The accuracy of these component probability predictions should be assessed using a (strictly) proper scoring rule, such as the quadratic (e.g. Brier) or logarithmic score, whose expectation is maximized by (and only by) predicting the true probability (Winkler, 1969; Savage, 1971). Our recent work has shown that such scoring rules are in fact averages of actual decision-making loss or regret (Rosen, 1995, 1996). These averages are over the potential decision problems in which the probability predictions might be used, each such decision problem being characterized by the regret associated with a false positive vs. that associated with a false negative. This theory also suggests an ROC curve alternative whose area is a proper scoring rule.

We also seek a measure of the extent to which the predictions are in the correct relative order, regardless of their numerical values. Such indices (Somers' Dyx, c index,...) are often called measures of ordinal discrimination, or of concordance (the number of pairs of predictions in the correct order), and in the dichotomous-outcome case, can arise from the empirical ROC curve. When a proper scoring rule is used to evaluate the overall correspondence of the predictions with the outcome, we wish to know how much of this inaccuracy could be due to miscalibration, and how much is unequivocally due to mis-ordering. This question is difficult to answer using concordance or ROC-based indices without strong parametric assumptions. We have introduced (Rosen 1994; Rosen, Burke, & Goodman, 1995a) a procedure identifying an unequivocal misdiscrimination component in any proper score, including logarithmic (binomial log-likelihood or Kullback-Liebler).

The procedure calibrates the predictions on a given data set so that all proper scores are simultaneously optimized on that data subject to the constraint that the ordering of the predictions not change (though ties can be produced). This constraint is very strong; without it such a calibration could often achieve a perfect score. The resulting score of interest (log-likelihood, Brier, etc.) on these self-calibrated predictions tells how much of the original score cannot possibly be improved by any order-preserving recalibration, and is thus an index of ordinal discrimination. The method can also be applied to predictions of a continuous dependent variable's mean.
This work has recently been published.


Rosen DB, Burke HB, Goodman PH. Improving prediction accuracy using a calibration postprocessor. Submitted for publication, 1995b.


2.1.4) Comparison of artificial neural networks with Cox proportional hazards model.

We are currently performing these comparisons, using several different measures of accuracy. (see 2.1.3)

2.2) Missing data.

Most data analyses either drop cases with missing data or impute some measure of central tendency for the missing data. Dropping cases has at least two negative effects: the remaining data may be biased, and it reduces the amount of data available for analysis. It may be possible to impute a central tendency value for missing data. But there are a number of statistical problems with the imputation of a central tendency, especially when there are many cases with missing data or when the important predictor variables contain much of the missing data.

The current cancer prediction system, the TNM staging system, does not provide a stage if one of the TNM variable is missing, nor does it provide guidance regarding prediction with missing variables (Beahrs, 1992).

In cancer prognostic factor research, many large data sets, both retrospective and prospective, suffer from missing data, i.e., missing prognostic factor information (Burke, 1993, 1995b, 1995c). We estimate that 75 - 80% of cases in some national data sets contain missing data. The usual approach to missing data is to remove the entire case, but this reduction in data set size, combined with the further reduction caused by splitting the data set into training and testing subsets, can significantly reduce the accuracy of statistical models. As Little and Rubin (1987) note:

"Statistical packages typically exclude units that have missing value codes for any of the variables involved in an analysis. This strategy is generally inappropriate, since the investigator is usually interested in making inferences about the entire target population, rather than the portion of the target population that would provide responses to all relevant variables in the analysis."
Moreover, when one is predicting an individual patient's outcome in a clinical situation, there is no guarantee that values for every predictive factor will be known for that individual; clearly "removing the case" is not an option in clinical situations. The result of a missing prognostic factor in clinical practice is usually an ad hoc guess of prognosis. For example, in the TNM staging system, if one of the covariates is not available no stage can be assigned, so the clinician must guess the patient's prognosis.

The missing data problem is especially severe in small data sets, where all data is precious. Here the problem can be enough to preclude the analysis of the data set. For example, in the Duke University breast cancer data set, which contains several of the new molecular-genetic prognostic factors, of the 230 cases in the data set, only 98 cases have no missing data. Given the number of covariates and the event rate (death from breast cancer), 98 cases are not sufficient for an analysis of these data. Because the new molecular-genetic prognostic factors are not always collected, and because molecular-genetic prognostic factors can be very powerful predictors of survival, it is essential that the problem of missing data be solved so that outcome prediction in cancer can advance.

When constructing a statistical model to predict a cancer outcome, e.g. survival, missing data (incomplete feature vectors) can cause a decrease in predictive accuracy (compared to the data set which does not contain missing data) because: (1) the missing data itself reduces the amount of data available to serve as a basis for prediction, and (2) the usual practice of removing cases with missing data, which reduces sample size, and therefore accuracy, reduces the amount of usable data to a level below that required to maintain predictive accuracy. One can never predict the true values of the missing data, but unless there are a great many missing values for a particular covariate, substituting values generated by an efficient method should improve prediction accuracy, compared to removing the cases with missing data. In other words, the problem we address is what method best deals with missing data, allowing us to retain the rest of the patient's data. Best means the method that produces the least biased estimates of the missing data values. Commonly used methods for estimating the missing values, e.g., imputing the mean covariate value or zero for the missing data, create strong biases and should be avoided (Little, 1992; Little and Rubin, 1987).

To be more precise, there are two missing data problems. One involves covariate values missing in the data sets used to train and test statistical prediction methods, such as logistic regression or Cox proportional hazards. The other involves missing predictors in a clinical situation; the patient's chart does not contain all the expected prognostic factors. For missing values, we prefer a method that uses all the information in the data set to estimate the missing values. This approach contrasts with, and is more accurate than, the simple insertion of a descriptive value (usually some measure of central tendency) of the covariate (e.g., a mean or median value) (Vamplew and Adams, 1992).

We are developing an artificial neural network approach for solving the missing data problem, using Normalized Radial Basis Functions. Normalized Radial Basis Functions based on estimating the joint input-output data distribution using a network representing mixtures of many multivariate gaussians.

Normalized Radial Basis Function (NRBF) networks (Moody and Darken, 1988; 1989; Poggio, 1989; Nowlan, 1990) model the output as a weighted average of an output value associated with each hidden unit. A given hidden unit also has
an associated position in the input space, and a "width" in each input dimension specifying how fast the weighting (importance in the weighted average) falls off in that dimension. Thus each hidden unit, or term in the model, is radial (or ellipsoidal) in that its influence decreases in all directions from its center. This is in contrast to conventional sigmoidal-projective neural networks, which do not use a weighted average, and in which each hidden unit has no "center" point, but rather selects (through its input weight vector) an arbitrary direction (linear projection) in the input space, where its contribution to the final prediction is a sigmoidal function along this direction. Training of the NRBF is accomplished using any of the standard neural network algorithms based on backpropagation of errors for calculation of the gradient of the log-likelihood with respect to the parameters (weights) of the network.

An advantage of a trained NRBF (when using gaussians as the radial weighting functions) is its ability to easily handle missing inputs during performance (i.e. prediction or recall), since merely ignoring those input components that are missing in a given input vector is equivalent to the correct Bayesian marginalization over the missing components.

The nonparametric form of the NRBF is known as a kernel estimator or Probabilistic Neural Network (Rosenblatt, 1956; Parzen, 1962; Nadaraya, 1964; Watson, 1964; Specht, 1990, 1991). Here, instead of using an optimization criterion to set the parameters of the network, there is a single hidden unit corresponding to each training case, whose location in the input space, as well as output value, is taken directly as those input and output values defining the case. These methods are sometimes called memory-based or case-based, since they store all the training data but require little or no computation during training. Thus these methods are attractive where training time is expensive but storage space during performance is not limiting, and they retain the ability to handle missing data during performance.

The NRBF can be generalized to form a Gaussian Mixture Network (Tresp, et al., 1994, Gharamani and Jordan, 1994) for the joint (input-output, i.e. predictor-response) probability density. This can use basis functions with non-diagonal variance-covariance matrices, thus incorporating some of the projective aspects of conventional sigmoidal neural networks. More importantly, they can be trained using the maximum-joint-likelihood (probability of observed training data inputs and outputs given parameters) criterion, enabling training on cases with arbitrary missing data (even if every case has some missing) using the iterative Expectation and Maximization (EM) algorithm (McKendrick, 1926; Harty, 1958; Orchard and Woodbury, 1972; Dempster, Laird, and Rubin, 1977).

It has been suggested by Efron and others that, ignoring the question of missing data, maximum-joint-likelihood estimation is less efficient than conventional maximum-likelihood estimation (probability of observed training data outputs given training data inputs and the parameters). Therefore, as our first missing-data method, we will examine the use of mixture networks to perform multiple imputation of missing values, as a preprocessor to be followed by a separate conventional feedforward neural network for prediction using these imputed values. A nonparametric (memory-based) form of this method has been proposed (Tresp, et al., 1995) but has the disadvantage of requiring that a good fraction of the training cases are complete, i.e. have no missing inputs.

The figure below demonstrates that, compared to the most common approach of removing cases with missing data, where accuracy decreases as missing data
increases, the accuracy of our approach is stable across a wide range of missing data.

100 training observations from known distribution (mean results of 10 trials)

Preliminary results were presented at the National Cancer Institute.


2.3) Censored cases

Discussed in detail in 2.1.2.

Task 3. Implementation of a clinically useful prognostic system


All our work is written in either C, C++, or XLISP-STAT.

3.2) Physician interface.

It is very important that physicians find the new prognostic system easy to use and useful. To this end we have implemented the prognostic system on a DOS platform with a Windows interface. We are presenting the system to clinicians
3.2) Physician interface.

It is very important that physicians find the new prognostic system easy to use and useful. To this end we have implemented the prognostic system on a DOS platform with a Windows interface. We are presenting the system to clinicians and receiving feedback regarding what is important to them in terms of information and the graphical display of the information. (see 2.1.2)

Tasks added to the project

We have added three tasks to the project. (1) A comparison of the NCDB and the SEER data sets, (2) an examination of methods for dealing with censoring bias (cases not missing at random) and competing risks, and (3) the computerization of the TNM staging system for breast cancer.

(1) Comparison of the NCDB and the SEER data sets.

To the best of our knowledge no one has performed a comprehensive comparison of the National Cancer Data Base (NCDB) and its associated Patient Care Evaluation (PCE), and the Surveillance, Epidemiology, and End Results (SEER) data sets. It is commonly felt that the SEER, which is population based rather than hospital based, is more accurate than the NCDB. But the NCDB is five-times larger than the SEER and the SEER over represents certain minorities. We are examining these data sets in terms of missing data, censoring, and the prognostic and outcome variables. This work is very time consuming, but it is necessary to determine which is the better data set for the prognostic system. Shown below are (1) a comparison of missing data, (2) a comparison of censoring, and (3) a Kaplan-Meier comparison of the NCDB and the SEER, for five year breast cancer-specific survival.

1. So far, for breast cancer, we have not found any significant demographic variable differences between the SEER and NCDB.

2. For breast cancer, the NCDB does have significantly more missing data than the SEER.

3. The NCDB and the SEER do exhibit differences in censoring. We have not yet determined how important these differences are.
This work is being prepared for publication.

**Hoang A, Burke HB, Rosen DB.** Comparison of the two national cancer data sets: SEER and NCDB. In preparation.

(2) Censoring bias and competing risks.

It is well known that censoring (lost-to-follow-up) can be biased in cancer, that cases are not missing at random. This is important because most statistical methods in use to day assume that censored cases are missing at random. We are developing analytic and empirical methods for (i) determining if there is a bias, and (ii) adjusting for the bias.

*Hoang A, Burke HB, Rosen DB. Survival analysis with some cases nonrandomly lost-to-follow-up. In preparation.*

It is also important to recognize that other causes of death affect the probability of death from breast cancer. The prognostic system must model other causes of death, in addition to modeling breast cancer mortality and censoring.

(3) Computerization of the TNM staging system

Although there have been plans for computerizing the TNM staging system for breast cancer, our implementation (see 2.1.2) is the first PC-based program for: (i) determining the breast cancer TNM stage from the breast cancer TNM variables, and (ii) predicting five year breast cancer specific survival based on TNM stage.

CONCLUSION

We have made substantial progress during the last year. We believe that we will be able to successfully meet our goal of providing a computer-based prognostic system that is more accurate than the TNM staging system and that is easy to use and understand, within the four year time frame of this grant.
References


Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.


Little RJ A. Regression with missing X's: a review. JASA 1992;87:1227-37.


Rosen DB. Ordinal discrimination index for any proper scoring rule.
Published Abstract. Medical Decision Making 1994;14:440.


Rosen DB, Burke HB, Goodman PH. Improving prediction accuracy using a calibration postprocessor. Submitted for publication, 1995b.


APPENDIX A - PRESENTED PAPERS


Burke HB. Survival curve analysis of cancer data. Grand Rounds, Department of Radiation Oncology, Cornell Medical Center - New York Hospital, New York NY, January 26, 1995.


Burke HB. Predicting survival in cancer. Grand Rounds, Division of Hematology-Oncology, New York Medical College, Valhalla NY, June 1, 1995.


Burke HB. Outcome prediction in cancer. Grand Rounds, Division of Solid Tumor Oncology Conference, Memorial Sloan-Kettering Cancer Center, New York, NY, June 19, 1995.


APPENDIX B - PHYSICIAN SURVEY

SURVEY OF PHYSICIAN ESTIMATES OF FIVE YEAR BREAST CANCER-SPECIFIC SURVIVAL

We are interested in your estimate of the breast cancer-specific survival of women diagnosed in the United States in 1985.

You are a (check one): _____ oncologist, _____ oncolgic surgeon, _____ pathologist, _____ radiation oncologist.

You graduated from medical school: _____ years ago.

Assume that each of the patients listed below is in your office, and asks you what her chances are, from date of diagnosis, of living five years. What is your estimate (% alive) of each patient living five years (not including those patients who died from causes other than breast cancer), over all primary and adjuvant therapies? Base your estimates on 1985 patients. (Note: for purposes of TNM staging, all patients with positive lymph nodes have been classed as T1).

<table>
<thead>
<tr>
<th>PATIENT DESCRIPTION</th>
<th>% PATIENTS SURVIVING 5 YRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1. Fifty-five year old, postmenopausal, 2 cm tumor, 0 positive lymph nodes, no distant metastasis, ER and PR positive, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 2. Thirty-five year old, premenopausal, 1 cm tumor, 3 positive lymph nodes, no distant metastasis, ER and PR negative, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 3. Fifty-five year old, postmenopausal, 5 cm tumor, 3 positive lymph nodes, no distant metastasis, ER and PR negative, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 4. Fifty-five year old, postmenopausal, 6 cm tumor, 0 positive lymph nodes, no distant metastasis, ER and PR negative, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 5. Forty-five year old, premenopausal, 6 cm tumor, 3 positive lymph nodes, no distant metastasis, ER and PR positive, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 6. Sixty-five year old, postmenopausal, 6 cm tumor, 3 positive lymph nodes, no distant metastasis, ER and PR negative, Grade 1.</td>
<td></td>
</tr>
<tr>
<td>Patient 7. Forty-five year old, premenopausal, 1 cm tumor, 3 positive lymph nodes, positive distant metastasis, ER and PR positive, Grade 3.</td>
<td></td>
</tr>
<tr>
<td>Patient 8. Forty-five year old, premenopausal, 3 cm tumor, 1 positive lymph node, positive distant metastasis, ER and PR positive, Grade 3.</td>
<td></td>
</tr>
<tr>
<td>Patient 9. Sixty-five year old, postmenopausal, 3 cm tumor, 1 positive lymph node, positive distant metastasis, ER and PR positive, Grade 3.</td>
<td></td>
</tr>
<tr>
<td>Patient 10. Forty-five year old, premenopausal, 6 cm tumor, 7 positive lymph nodes, positive distant metastasis, ER and PR positive, Grade 3.</td>
<td></td>
</tr>
</tbody>
</table>