AD_____

Award Number DAMD17-94-J-4383

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

PRINCIPAL INVESTIGATOR: Philip H. Goodman, M.D.,M.S.

CONTRACTING ORGANIZATION: University of Nevada
Reno, Nevada 89557

REPORT DATE: February 1999

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 1999	Final (1 Aug 94 - 31 Jan 99)

4. TITLE AND SUBTITLE
Developing and Implementing the AJCC Prognostic System for Breast Cancer

5. FUNDING NUMBERS
DAMD17-94-J-4383

6. AUTHOR(S)
Philip H. Goodman, M.D.,M.S.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Nevada
Reno, Nevada 89557

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In the past staging systems provided a simple, easily understood ordering of patient outcomes. For over thirty years breast cancer outcome prediction has been based on the TNM staging system. There are two problems with staging systems generally, and specifically with the TNM system: (1) they are not very accurate, i.e., their predictions are not close to the true outcomes), and (2) their accuracy can not be substantially improved because additional predictive factors can not be included in the system without increasing the system's complexity to the point where it is not longer useful to the clinician.

The objective of this research is to replace with current TNM stage system with a new prognostic system that is inherently more accurate than the current system and that can integrate new prognostic factors to further improve prognostic accuracy. There are three components to accomplishing this objective, which are the goals of this research project: (1) the development of the prognostic model itself, (2) the creation of the prognostic system by training the model with breast cancer outcome data, and (3) the computer-based implementation of the system for clinicians and tumor registries (clinical decision support system).

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
72

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. 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).

RG. 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.

Philip Gordman. 2/24/99

PI - Signature / Date

Table of Contents

INTRODUCTION

Purpose and Scope of the Research

Outcome predictions (natural history, therapy-specific, and post-therapy) are crucial to cancer because they estimate the natural history of the disease (natural history), are required for determining the optimal therapy (therapy-specific), evaluate the effectiveness of treatment (post-therapy), and they can be used to match patients for clinical trials, provide patient information, and perform quality assurance assessments.(Burke, 1998)

In the past staging systems provided a simple, easily understood ordering of patient outcomes (all patients were assumed to experience one of four possible outcomes). For over thirty years breast cancer outcome prediction has been based on the TNM staging system. There are two problems with staging systems generally, and specifically with the TNM system: (1) they are not very accurate, i.e., their predictions are not close to the true outcomes), and (2) their accuracy can not be substantially improved because additional predictive factors can not be included in the system without increasing the system's complexity to the point where it is not longer useful to the clinician.(Burke, 1993)

The objective of this research is to replace with current TNM stage system with a new prognostic system that is inherently more accurate than the current system and that can integrate new prognostic factors to further improve prognostic accuracy. There are three components to accomplishing this objective, which are the goals of this research project: (1) the development of the prognostic model itself, (2) the creation of the prognostic system by training the model with breast cancer outcome data, and (3) the computer-based implementation of the system for clinicians and tumor registries (clinical decision support system).

Background

In America during most of this century the treatment for breast cancer was either a radical mastectomy,(Donegan, 1979) as described by Halsted before the turn of the century,(Halsted, 1894) or a modified radical mastectomy, as described by Patey.(Preisler, 1992) More recently lumpectomy, chemotherapy, and radiation therapy have become important treatment modalities. With the rise of effective therapies has come the need for methods that accurately assess prognosis, because therapy depends on prognosis and the patient's wishes. By the 1950s there were many incompatible staging systems in existence for breast and other cancers. The TNM staging system (primary tumor, regional lymph nodes, and distant metastases) originated as a response to the need for an accurate, universal staging system.(Fleming, 1997)

Since the TNM staging system began in the 1960's many putative prognostic factors have been identified for breast cancer (Burke, 1995a). The proliferation of putative prognostic factors raises several issues regarding the identification of prognostic factors. These issues include: what are the criteria for determining what putative prognostic factors to test, in what context are the factors prognostic, do the factors retain their prognostic value in the presence of other prognostic factors or do they require other factors in order to be prognostic, and how can prognostic factors be combined to increase overall predictive accuracy? The result of the proliferation of putative prognostic factors is clinical confusion; no one knows how to integrate these factors nor how to reconcile conflicting prognostic factor predictions. Further, almost none of the putative prognostic factors have been tested in large, random sample data sets that include all important prognostic factors and all treatment modalities and have ten year follow-up.

The identification and integration of new prognostic factors is crucial to providing more accurate outcome predictions (recurrence, death, etc.). It is not possible to integrate new factors in the TNM stage model for several reasons.(Burke, 1993) First, the TNM stage model is based on a bin model with 40 bins (5x4x2), and it has all the characteristics of a bin model. One characteristic of a bin model is that the number of bins increases rapidly with the number of variables. For example, if we add the variable histologic grade, with its four types, to the TNM stage model, the result is 160 bins (5 x 4 x 2 x 4). Thus, for any set of new variables, the number of bins that would have to be added to a stage would be enormous, and the system would become too complex to be useful. Second, adding variables to the TNM stage model would demonstrate another characteristic of the model, namely that it is a post hoc system. In a post hoc system the outcomes are examined and the bins/stages are arranged in order of decreasing survival. The only way to add a variable to such a system is to collect a large data set with all the predictive variables present and create a new set of stages. With each new variable this process must be repeated. Third, since the accuracy of a bin/stage model depends on the number of patients in each bin, as the number of variables increases the number of bins increases, and the number of patients must increase exponentially to have enough patients per bin to maintain accuracy.

A new prognostic system is required; a system that can test putative prognostic factors and integrate them to increase predictive accuracy. This increase in accuracy will improve our ability to select the most effective therapy, enroll patients in clinical trials, provide information to patients, and develop quality assurance programs.

If we are to go beyond the current prognostic system three fundamental questions must be answered. How accurate is the current TNM stage system? Can another outcome prediction system increase prognostic accuracy using just the TNM variables? Can putative prognostic factors be integrated into a new prognostic system to increase prognostic accuracy?

BODY

Methods, Assumptions, and Procedures

Data sets and variables

We use the Commission on Cancer's breast cancer Patient Care Evaluation (PCE) data set, the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) breast cancer data set, and Duke University Medical Center's breast cancer data set.

In October 1992, the American College of Surgeons (ACS) requested cancer information from ACS accredited hospital tumor registries in the United States. Specifically, they requested the first 25 cases of first diagnosis breast cancer and colorectal cancer seen at each institution in 1983, as well as follow-up information, including deaths, through the date of the request. Variables from this data set used in the breast cancer analysis are: age, race, payment method, menopausal status, family history, previous biopsy, other cancer, other breast cancer, nipple discharge, mammogram, where in the breast the cancer occurred, necrosis, histologic grade, estrogen receptor status, progesterone receptor status, number of lymph nodes positive, number of lymph nodes examined, presence or absence of distant metastasis, tumor size, tumor type (in situ, extension to chest wall, inflammatory), treatment (surgery, chemotherapy, radiation therapy), and outcome (alive or dead). All variables are binary except age, tumor size, number of positive lymph nodes, and number of lymph nodes examined. The PCE data set contains up to eight years of follow-up information. The analysis endpoint is breast cancer-specific five year survival. Cases with missing data and those censored before five years were excluded. The data set was randomly divided into a training set of 5,169 cases, including training and stop-training subsets, and a testing set of 3,102 cases.

Variables from the PCE data base used in the colorectal cancer analysis are: age, race, sex, signs and symptoms (change in bowel habits, obstruction, jaundice, malaise, occult blood, abdominal pain, pelvic pain, rectal bleeding, other), diagnostic and extent of disease tests (endoscopic, radiographic, barium enema, CT scan, biopsy, CEA, x-ray, colonoscopy, flexible sigmoidoscopy, IVP, liver function tests, biopsy, other), primary site of tumor, level of tumor, histology, grade, number of lymph nodes examined, number of lymph nodes positive, distant metastases, and outcome (alive or dead). The endpoint is five year colorectal cancer specific survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 5,007 training cases, including training and stop-training subsets, and a testing set of 3,005 cases.

The National Cancer Institute's SEER breast cancer data set, for new cases collected from 1977-1982, with ten-year follow-up, is also analyzed. The SEER data set extent of disease variables are comparable to, but not always identical with, the TNM variables. The endpoint is breast cancer specific ten year survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 3,788 training cases, including training and stop-training subsets, and a testing set of 2,999 cases.

The Duke data set has been described in a previous paper.(Marks, 1994) Briefly, all patients were pathologic TNM stage I or early stage II. Early stage II included all the TNM stage II patients except those with five or more positive lymph nodes. The variables were: age, race, tumor size, nodes positive, nodal stage, nuclear grade, histologic grade, p53, c-erbB-2, estrogen receptor (ER) and progesterone receptor (PR), vascular invasion, adjuvant therapy (tamoxifen, chemotherapy), and radiation therapy. Patients who underwent a lumpectomy received radiation therapy. Patients who underwent a modified radical mastectomy did not receive radiation therapy. There are 229 cases of which 226 had complete data for all variables except ER and PR status. Because of the number of cases missing either ER or PR both were removed from the data set. The survival rate was 70%. The prediction endpoints were five and ten year overall survival.

Combining factors

It is rarely the case that one factor is sufficiently predictive, i.e., that it is able to predict the outcome of interest with 100% accuracy. The usual strategy when dealing with predictors is to combine several in a predictive model. The most useful grouping of factors is one in which all the factors are powerful and predictively orthogonal to each other, i.e., they represent independent aspects of the disease process. If they represent aspects of the disease that are not independent then their information will overlap and one will not add predictive power. The statistical method employed must be able to capture the complexity of the disease process that is represented by the factors being combined, e.g., nonlinearity and interactions.(Burke, 1998)

A predictive model is the result of using a statistical method to relate one or more predictive factors to an outcome. For example, the mathematical formula generated by the logistic regression statistical method relates the predictive factors (input variables), in terms of their ß-coefficients, to a binary disease outcome, e.g., relapse, death, etc.

It should be noted that the predictive power of a factor must always be associated with the statistical method and the other factors included in the model in any statement of the factor's accuracy because a factor's power can vary with the model. The model may or may not contain all the relevant factors and it may or may not be efficient in capturing the power of the factors.

Methods for combining factors

Many methods have been used to combine predictive factors. The main methods in cancer are: bins, stages, indexes, either as discrete endpoint or as Kaplan-Meier product-limit models; decision trees; and regression methods including logistic, proportional hazards, and artificial neural networks.

Bins are the result of the mutually exclusive and exhaustive partitioning of discrete variables. Each combination of variable values is a bin and every patient is placed in the bin corresponding to their variable value combination. An example is the TNM classification of ovarian cancer. Tumor location (T1a, T1b, T1c, T2a, T2b, T2c, T3a, T3b, T3c), regional lymph node involvement (N0, N1), and existence of metastases (M0, M1) produce thirty-six bins.

If there are enough people in each bin, it can be shown that the frequency of the outcome in the population within each bin is the best predictor of the true outcome. In other words, no prediction model can be more accurate than the bin model if the variables are discrete and the population very large. Problems with bin models include: (1) Continuous variables must be parsed into discrete variables, almost always resulting in a loss of predictive information and therefore a loss of accuracy. (2) As the number of discrete variables increase the number of bins increase exponentially. In order to maintain accuracy there most be a corresponding exponential increase in the size of the patient population. (3) The proliferation of bins reduces the ability to understand the phenomena. Since the main reason of creating a bin model is usually for ease of understanding and ease of use, bin models are rarely used in situations where there are more than two or three predictive factors.

A partial solution to the problems of a bin model is a stage model. A stage model is the grouping of bins into super-bins. The justification for the grouping is the assumption that the factors selected are indexes of the "stages" of the disease process. For example, in breast cancer, the TNM staging system combines forty TNM classification bins into six super-bins based on decreasing survival, and these super-bins are termed the TNM staging system.

A small set of stages have the potential to maintain explanatory simplicity and ease of use. Problems with stage models include: (1) The combining of bins into super-bins/stages usually substantially reduces predictive accuracy. (2) Stage systems do not overcome the exponential increase in bins and in patients associated with adding a variable to the staging system, they just delay the problem at the cost of predictive accuracy. If the stages are held constant as variables (and their associated bins) are added the staging system, the potential improvement in accuracy associated with the additional bins will be small to nonexistent. But, if the stages are expanded to accommodate additional bins, the system looses its ease of understanding and usefulness. Thus, attempts to improve predictive accuracy by adding variables to a bin/stage model are rarely successful. (3) The problems of parsing continuous variables, with the resulting loss in predictive accuracy, remains.

Indexes associate numerical scores (usually based on a bounded, linear scale) with bins or groups of bins. The scores are parsed into discrete ranges, and each range is associated with a disease stage (usually a severity of illness system). Indexes offer some flexibility in the grouping of bins, but at the cost of further degradation in predictive accuracy. The simplest example of an index is the Apgar score.

Any bin, group of bins, stages, or scores can be contrasted, in terms of outcome, with another bins, group of bins, stages or scores at the end of a single time interval or across a series of event time intervals. (In other words, comparing predictive factors.) Both the single time interval and the event interval approaches usually deal with censoring by dropping censored cases at the time interval in which they are censored. The most common descriptive approach for contrasting predictive factors across a series of event time intervals is the Kaplan-Meier product-limit method (inferential methods that can accommodate continuous variables, and that usually require a proportional hazards assumption, will be discussed later when regression methods are presented). A Kaplan-Meier plot should always include confidence intervals around each line. A significant difference is a Kaplan-Meier comparison is usually

assessed by a log-rank test (which assumes proportional hazards). It is important to note that there is currently no widely accepted method for comparing the accuracy of two Kaplan-Meier comparisons based on different stratifications of the same variables. The use of the log-rank p-value to select one stratification over another is incorrect because the log-rank test determines whether a factor stratification is likely to have occurred by chance. Extreme stratifications may result in a smaller p-value, but it may also reduce predictive accuracy over the entire population.

Univariate methods, including univariate regression methods, are not appropriate for deciding whether a factor is or is not predictive. These methods should not be used to assert that a factor is predictive because a new factor must be assessed in the context of the known factors. Univariate methods should not be used to assert that a factor is not predictive because a variable may be predictive only when it is interacting with other factors.

Decision trees split predictive factors to maximize predictive power using a loss function such as the log-likelihood and a greedy search algorithm. The most well known decision tree approach is the Classification and Regression Trees (CART) recursive partitioning method (Breiman, 1984). Empirically, we have never found it to be the most accurate statistical method, when compared to regression methods. Its problems include the selection of the correct loss function, it has difficulty dealing with continuous variables. and it can overfit when searching for the best predictors and when there are more than a two or three splits.

Logistic regression is the cumulative probability of a binary event occurring by a specific time. It uses a maximum likelihood loss function and a greedy search technique. It is a very efficient method for problems that have a binary outcome (e.g., recurrence, survival). Its limitation is that it must span a large time interval and does not differentiate when an event occurs in the time interval. Also, in order to handle censoring one must create a logistic regression model for each time interval and drop the cases that are censored in each interval.

"Proportional hazards" methods include the Weibull, exponential, and Cox. The Cox proportional hazards regression method (Cox, 1972) is the most commonly used. All three methods assume that the hazard of each patient is proportional to the hazards of all the other patients and that the degree of each patient's hazard is related to their relative risk. The Cox model does not create survival curves. For survival curves a baseline hazard must be introduced (Cox-Breslow estimates; Breslow, 1974). Some researchers incorrectly believe that the Cox is the only regression method that can deal with censoring. A multi-interval logistic regression can deal with missing data. In cancer, the proportional hazards assumption is often violated. Therefore, anyone using a Cox model must demonstrate that proportional hazards holds for their factors and outcome.

Molecular genetic factors exhibit the properties of complex systems, they are nonlinear and they are interactional, i.e., they act nonmonotonically and in concert.(Steel, 1993) Thus, capturing the factors as part of a complex system is critical to accurate prediction of the behavior of the system. Artificial neural networks are capable of complex systems. (Burke, 1996)

The idea that learning can be viewed as the modification of information by repetitively passing it through processing nodes originated in the late 1940's as a way to model the physiology of neuronal processes.(Hebb, 1949) The operationalization of this idea was called an artificial neural network. Gradually it became apparent that this information theoretic approach to learning was very powerful and very general; it was useful in, and applicable to, many learning situations. Since statistics can be viewed as learning from the data, it is not unexpected that this approach would be mathematically proved and operationalized within the domain of statistics.

Artificial neural networks are universal approximators. It has been shown that any real, continuous function can be approximated to any degree of precision by a three-layer network with x in the input layer (patient variables), a hidden layer of sigmodal units, and one layer of output

units (the outcome is what is be predicted, for example, death), as long as the hidden layer can be arbitrarily large. (Hornik, 1990, 1994)

Artificial neural networks, as a class of nonlinear regression and discrimination statistical methods, are of proven value in many areas of medicine. (Baxt, 1995; Dybowski, 1995; Westenskow, 1992; Tourassi, 1993; Leonh, 1992; Gabor, 1992, von Osdol, 1994)) They do not require a priori information regarding the phenomenon, they make no distributional assumptions, and with the appropriate method to avoid overfitting (i.e., loss of generalization by fitting the patterns to the test data too precisely), artificial neural networks are usually at least as accurate as classical statistical models and, depending on the complexity of the phenomena, can be much more accurate. Artificial neural networks have, for example, been shown to be more accurate than logistic regression, CART (pruned or shrunk), and principal components analysis at predicting five year breast cancer specific survival. (Burke, 1995b)

There are many types of neural networks. Backpropagation neural networks are the most commonly used neural networks is medical research. Examples of backpropagation neural networks include the "classical" method (described below), cascade correlation,(Fahlman, 1991) and conjugate gradient descent.(Weiss, 1991) Instead of global error reduction the cascade correlation neural network creates a hidden layer node to reduce the error. Additional nodes are created to continue to reduce the error until one is in danger of overfitting the data. Conjugate gradient descent is a type of backpropagation that requires only one setting. It is a method of optimization that assumes that the best search direction for to lowest error surface starts in the direction of steepest descent and proceeds in the direction conjugate to that taken in the previous step.

Fuzzy ARTMAP and the probabilistic neural network are not backpropagation networks and they demonstrate the diversity of possible neural networks. Fuzzy ARTMAP neural networks are based on adaptive resonance theory (ART) architectures.(Carpenter, 1991) This type of neural network uses feedback and competition to self-organize stable recognition codes in real time in response to arbitrary sequences of input patterns. The probabilistic neural network (PNN),(Specht, 1990) is a neural realization of kernel density estimation techniques. PNNs have been found to achieve performance similar to backpropagation neural networks, but with many orders of magnitude less training time. The PNN is a feed-forward network with two hidden layers. Normally each node of the first hidden layer corresponds to a single training case, and computes the similarity (in predictive variables) between that training case and the current input case. The second hidden layer and the output layer each contain one node for each possible outcome, computing first the estimated probability of the current predictive-variable values given each outcome, and then the estimated probability of each outcome given the predictive-variable values.

In medical research, the most commonly used artificial neural networks (ANN) are multilayer perceptrons that use backpropagation training. Backpropagation consists of fitting the parameters (weights) of the model by a criterion function, usually squared error or maximum likelihood, using a gradient optimization method. In backpropagation artificial neural networks, the error (the difference between the predicted outcome and the true outcome) is propagated back from the output to the connection weights in order to adjust the weights in the direction of minimum error. (For a more detailed description of artificial neural networks see: Burke, 1997a; Cross, 1995). The artificial neural network employed in this research is composed of three interconnected layers of nodes: an input layer with each input node corresponding to a patient variable, a hidden layer, and an output layer. All nodes after the input layer sum the inputs to them and use a transfer function (also known as an activation function) to send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function such as the logistic. The connections between the nodes have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the adjacent layer nodes. These weights, along with the connections among the nodes, determine the output of the network. For a one hidden layer network with input

xk, hidden units vi, output units oi, weights from input to hidden units wjk, weights from hidden units to output units wij, and transfer function g, the output of the network is given by

$$o_i = \sum_j W_{ij} g\left(\sum_k w_{jk} x_k - \phi_j \right) - \theta_i \qquad (1)$$

where ϕ and θ are bias terms at hidden unit j, and output unit i. The network implements a set of functions $o_i=F_i\{x_k\}$ for input variables x_k to output variables o_i. The weights and thus the functions F_i are to be estimated by minimizing a cost function. A common cost function is a measure of squared error given by

$$E(\underline{w}) = \frac{1}{2} \sum_{ui} [\varsigma_i^u - o_i^u]^2 \qquad (2)$$

Where ξ_k^u is the observed output for output unit i with pattern u. Other cost functions include the negative likelihood of a model. The usual learning rule for the weights of the network is gradient descent, given by

$$\Delta W_{ij} = \eta \sum_u \delta_i^u V_j^u, \qquad where \qquad \delta_i^u = g'(h_i^u)[\varsigma_i^u - o_i^u] \qquad (3)$$

The usual artificial neural network uses backpropagation training, the maximum likelihood criterion function, and a gradient descent optimization method.

All outcome analyses, except for PCA, CART, cascade correlation, and conjugate gradient descent, were performed twice. The second analysis was performed independently by a different researcher who did not know the first researcher's results. There were no significant differences between the two researcher's results. All results are on a test data set.

Assessing and comparing statistical models: measuring accuracy

In order to assess and compare models, it is necessary to distinguish between significance, accuracy, and importance. Significance is the fact that it is unlikely that either a trained statistical method (i.e., a statistical model) or a predictive factor' predictions are due to chance (e.g., the chi-square test). Significance is not necessarily accuracy. Accuracy is the association between the model's outcome predictions and the test population's known outcome. The importance of a factor or a model is based on whether the model or the factor possesses sufficient accuracy to be useful in answering a particular clinical question. Finally, the assessment of model or factor significance, accuracy, and importance must be based on test data set results, not on training data set results.

There are several approaches to assessing the accuracy of a multivariate model and for comparing multivariate models (e.g., Goodman and Kruskall's Gamma, Kendall's Tau). The best method currently in use is the area under the receiver operating characteristic curve. The area under the receiver operating characteristic curve (Az) is the measure of predictive accuracy used to assess the performance of the artificial neural networks.(Swets, 1996) It can be used to assess and compare the adequacy of statistical models. Az can be directly calculated by Somer's D (Somer, 1962) or it can be approximated by its trapezoidal area.(Bamber, 1975) The area under the curve is a nonparametric measure of discrimination. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization. Its computation requires only that the prediction method produce an ordinally-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the prediction method will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from zero to one. When the predictions are unrelated to survival, the score is .5, indicating chance accuracy. The farther the score is from .5 the better, on average, the prediction method is at predicting which of the two patients will be alive. Significant differences in the receiver operating characteristic areas

between two models can be tested following Hanley and McNeil (1982), by calculating their variances, or by the bootstrap method (Efron, 1979).

Results

The goal of this research is the creation of a clinical decision support system for women newly diagnosed with breast cancer. In order to create this system we needed to evaluate prognostic factors, develop a breast cancer prognostic model, and the implement a system that can be used clinically.

Prognostic factors and outcomes

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).(Burke, 1995a) In addition, work is continuing on the book: "Burke HB, Henson DE. Prognostic Factors and Systems in Cancer. Kluwer Academic Publishers."

Histologic grade is not a part of the TNM staging system but is widely used to assess severity of illness at diagnosis. We assessed the ability of grade to predict five year survival and found that, tumor size is a stronger predictor and that it does not increase predictive accuracy when added to tumor size.(Burke, 1997b)

Mammographic early detection of breast cancer is reducing the usefulness of the TNM staging system because most tumors detected by mammography are small and few women have involved lymph nodes or distant metastases. Providing an accurate prognosis in early-detected breast cancer is a critical problem. Can new molecular-genetic prognostic factors take over the predictive burden from the TNM in these women? We assessed the traditional prognostic factors (TNM variables, age, estrogen and progesterone receptor status, histology), and the molecular genetic factors p53 and erbB-2 in 260 women with early detected disease. We performed this analysis in conjunction with describing the proper way to assess therapy-specific prognostic factors. Although the results were somewhat complex, a fair summary would be that for women with small tumors and no metastases, p53 and erbB-2 demonstrated an improved predictive accuracy. Five and ten year survival accuracy went from chance, $Az = 0.5$, to between 0.75 and 0.85 (depending of the analysis). These results have important implications for node-negative women and for creating clinical trial populations. Although our results should be interpreted cautiously because of the number of cases, they suggest that molecular genetic prognostic factors, if properly used, will play an important role in breast cancer outcome prediction.(Burke, 1998a)

We have found that for predicting five year breast cancer-specific survival, using currently collected prognostic factors and an overall 30% five-year breast cancer-specific mortality rate, approximately 2,300 patients are required to attain maximum accuracy. Adding more cases does not provide any improvement in prediction accuracy.

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 end-point 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.

Specimen banks are an important part of prognostic factor research. Two issues are central to the evaluation of prognostic factors. The first is the time from diagnosis to the analysis of outcomes (e.g., mortality). The longer this interval the longer the prediction time interval. To provide, for example, ten year survival predictions a patient population must be followed for ten years. The ten year information is used to assess prognostic factor predictive accuracy and to provide ten year outcome predictions to future patients. The second issue is the accrual of a sufficient number of outcomes so that the assessment of the factor is statistically reliable. Reliable means that a similar result would be observed if the analysis were repeated. A human specimen bank that contains abnormal and normal tissue, white cells, serum and plasma facilitates prognostic factor research because it eliminates the waiting time problem and the outcome accrual problem by collecting specimens from a defined patient population and following the patients for a sufficient number of years. When a new putative prognostic factor is discovered the stored material can be used to immediately assess its predictive power.(Burke, 1998b)

Time plays an important role in outcome prediction. There are two kinds of time related to prognostic factors. The first is the value of the factor in predicting a future outcome. For example, the accuracy of a factor in predicting five or ten year survival (discussed above). The second type of time is the predictive value of a factor collected over time. In other words, does the predictive value of the factor change over historical time so that, for example, tumor size is less predictive for women today than it was for women 20 years ago. In collaboration with investigators in Finland we have shown the prognostic value of a factor changes over both types of time. (Lundin M, Lundin J, Burke HB, Toikkanen S, Liisa P, Heikki J. The role of time in breast cancer outcome prediction. Submitted for publication.)

Survival Models

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.

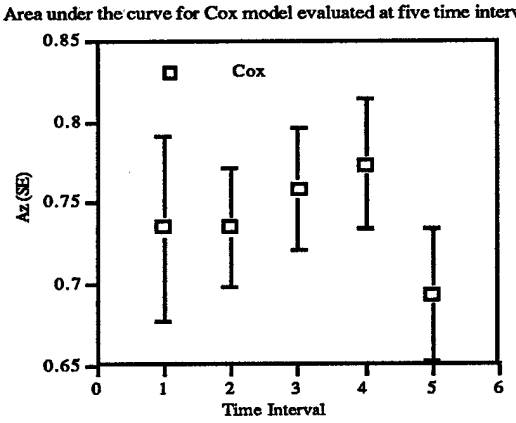
The Cox proportional hazards model (Cox, 1972) is a linear effects model. It estimates the importance of each covariate, and it handles censored cases. It assumes proportional hazards and it does not provide a survival curve without the imputation of a baseline survival curve.

We began our comparison of the Cox by examining whether breast cancer violates the proportional hazards assumption of the model. Proportional hazards methods include the Cox (1972), and less commonly the Weibull or exponential distributions (Evans, 1993). Proportional hazards methods assume that the hazard of each patient is proportional to the hazards of all the other patients and that a individual patient's hazard is related to that patient's relative risk. The Cox model does not create survival curves. For Cox-related survival curves a baseline hazard must be introduced (Breslow-Cox estimates; Breslow, 1974). Some researchers incorrectly believe that only regression methods that assume proportional hazards can deal with censoring, but a multiinterval regression model that drops patients during the interval in which they are censored is capable of dealing with censoring. It is always vital to test the proportional hazards assumption

when using a regression method that relies on it. There are several methods for assessing proportional hazards violation, including Schoenfeld's partial residuals (Schoenfeld, 1982) and the log hazard ratio as a function of time (Gore, 1986). We have created a method somewhat similar to Gore. We construct a Cox model, divide the time into sub-intervals, and assess the accuracy of the model for each sub-interval. If proportional hazards holds, accuracy should be constant across sub-intervals. Results for breast cancer are shown below.

TABLE. Area under the receiver operating characteristic (Az) for two Cox models; breast cancer (five one-year intervals) N = 1,222 and melanoma (three six-month intervals) N = 60.

Model/Interval	1	2	3	4	5
Breast (SE)	.734 (.057)	.735 (.036)	.758 (.038)	.773 (.040)	.693 (.041)



Because the Az values are not constant across the sub-intervals, proportional hazards does not hold for breast cancer.

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 compared statistical method to artificial neural networks in terms of five year breast cancer-specific survival. Principle components analysis is the linear combination of predictor variables. The logistic regression analysis is performed in a stepwise manner, without interaction terms, using the statistical program S-PLUS (S-PLUS, Seattle, WA) with cubic spline terms for age.(Smith, 1979) Two types of Classification and Regression Tree (CART) analyses are performed using S-PLUS. The first was a 9-node pruned tree (with 10-fold cross validation on the deviance), and the second was a shrunk tree with 13.7 effective nodes. The ANNs were described above.(Burke, 1997a)

FIVE YEAR SURVIVAL PREDICTION ACCURACY

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

* The area under the curve of the receiver operating characteristic.

** NN, neural network.

We extended our comparison of the TNM staging system and the ANN from 5 year survival to 10 year survival.

TEN YEAR SURVIVAL PREDICTION ACCURACY

PREDICTION MODEL	ACCURACY*	SPECIFICATIONS
TNM Stages	.692	Ø,I,IIA,IIB,IIIA,IIIB,IV
Artificial Neural Network	.730§	3-5-1

* The receiver operating characteristic area.
§ $p < .01$

We found that the predictors collected at disease discovery are less accurate in predicting 10 year survival than 5 year survival. Five year survival accuracy for the TNM staging system was 0.720 and for the artificial neural network, 0.784 and the corresponding ten year survival accuracy was 0.692 and 0.730.

Artificial neural networks are a general regression method that do not assume proportional hazards and can capture nonlinearity and complex interactions (Burke, 1994, 1995a). Multiinterval artificial neural networks can handle censoring in the same way that multiinterval logistic regression models handle censoring. It seems clear that proportional hazards is probably not appropriate for breast cancer or lung cancer (results not published).

Accuracy

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 accuracy's 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, 1995a, 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 want 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, etc.) 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,1995b) 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 re-calibration, and is thus an index of ordinal discrimination.

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 have developed 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; Hartly 1958; Orchard and Woodbury, 1972; Dempster, Laird, and Rubin, 1977).

It has been suggested by Efron (1979) 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, 1995) but has the disadvantage of requiring that a good fraction of the training cases are complete, i.e. have no missing inputs.

We wish to train a feedforward projective-sigmoidal neural network (MLP) on breast cancer outcomes data missing both binary and continuous input variable values. A Gaussian-Bernoulli mixture model is trained on the data (using EM). It then performs stochastic imputation (filling in) of the missing values, as a preprocessor to the MLP. In order to compare predictive accuracy when the training data are complete vs. incomplete/imputed, we use only complete cases from a natural data set, but artificially remove 80% of their input data values. Very little difference is observed in the comparison, suggesting that the mixture model is quite effective here, despite the fact that more than 99% of the cases/instances had some missing value(s). The mixture model can be used both for output/outcome prediction by a trained MLP and for the training process itself.

The problem of missing (incomplete) data is ubiquitous in clinical medicine, both during model development and training/fitting, and during prediction/performance/recall on new cases by the final fixed model. In the present work we employ finite mixture (Titterington, Smith & Makov, 1985) models. We will refer to these models as mixture networks, since they have a network interpretation (nodes, connections, and local computation) and Gaussian mixtures can be viewed as generalizations of normalized radial basis function (NRBF) neural networks (Moody & Darken, 1988; Moody & Darken, 1989; Poggio & Girosi, 1989; Poggio & Girosi, 1990; Nowlan, 1990; Nowlan, 1991).

Mixture networks have been successfully applied to missing-data problems (Ghahramani & Jordan, 1994; Tresp, Ahmand & Neuneier, 1994). They are very flexible models that can be used in "semi-parametric" style, i.e. making them large enough to capture the predictive complexity of a phenomenon, without necessarily determining the significance or meaning of the model terms. Thus they require relatively little operator intervention in their use. They are also able to handle several types of variables together in a multivariate problem without resorting to ad hoc combination of disparate models/methods.

The NRBF is a model for the regression (conditional expectation or mean of y given $\underline{x}$) function defined by an average of parameters $\underline{\mu^y}$ (using an underbar to indicate a vector), weighted by Gaussians (i.e. normal pdfs) and their mixture parameters$\{\gamma_j\}$ as

$$E\{y \mid \underline{x}\} = \frac{\sum_{j=1}^{m} \mu_j^y \gamma_j N(\underline{x}; \underline{\mu_j^x}, \underline{s_j^x})}{\sum_{j=1}^{m} \gamma_j N(\underline{x}; \underline{\mu_j^x}, \underline{s_j^x})}$$

$N(\underline{x}; \underline{\mu_j^x}, \underline{s_j^x})$ is a Gaussian with the specified parameter vectors of means $\underline{\mu_j^x}$ and variances $\underline{s_j^x}$ (diagonal variance-covariance matrices are assumed) for the jth term in the model. The x

superscripts indicate that these parameters correspond to x components, while the notation $\underline{\mu_j^x}$ will become clear in the next paragraph. The NRBF model can be viewed as a parametric version of the nonparametric kernel (Rosenblatt, 1956; Parzen, 1962) regression estimator as first proposed by Nadaraya (1964) and Watson (1964). Training the NRBF can be done using maximum-conditional-likelihood of the outputs given the inputs.

A Gaussian mixture network is a model for the full joint probability density of a vector $\underline{z}$ of unrestricted-real-valued variables, given by

$$p(\underline{z}) = \sum_{j=1}^{m} \gamma_j N(\underline{z}; \underline{\mu_j}, \underline{s_j})$$

where if we identify the variables as $\underline{z} = (\underline{x}, y)$, and the means and variances of the mixture terms as $\mu_j = (\underline{\mu_j^x}, \mu_j^y)$ and $s_j = (\underline{s_j^x}, s_j^y)$ respectively, we can obtain the NRBF regression formula above. Although within each mixture component (term) the variables are independent, dependencies are introduced upon summing these in the mixture. Because there are the additional parameters s_j^y to be estimated, maximum-likelihood estimation for this model must typically use the unconditional joint likelihood of all of the variables, not distinguishing between "inputs" and "outputs" during training.

A Bernoulli mixture model consists of a sum of terms, each representing a product of Bernoulli distributions (i.e. binomial distribution with a single draw) for the inputs. This model is used when the variables are binary (dichotomous). Ghahramani & Jordan (1994) studied separate Gaussian and Bernoulli mixture models for all-continuous and all-binary tasks, respectively, and mentioned that these and other distributions can be combined within each term of a mixture model when the variables are of different types (dichotomous, real-valued, and others) within the same task. A Gaussian-Bernoulli mixture model is defined by

$$p(\underline{c}, \underline{b}) = \sum_{j=1}^{m} \gamma_j N(\underline{c}; \underline{\mu_j}, \underline{s_j}) B(\underline{b}; \underline{\alpha_j})$$

where with $\underline{z} \equiv (\underline{c}, \underline{b})$ we have partitioned $\underline{z}$ into vectors of continuous and binary components, and $B(\underline{b}; \underline{\alpha_j})$ is a product $\prod_i \alpha_{j,i,b_i}$ of univariate Bernoulli distributions with Bernoulli/binomial parameter α_{ji1} (the jth model term's predicted probability that binary variable $b_i = 1$) and its complement $\alpha_{ji0} \equiv 1 - \alpha_{ji1}$.

A mixture network can predict any variable (input or output) from any combination of others. For example, Figure 1 shows how a two-variable Gaussian mixture can predict either variable from the other (just expectations in this case). Even if the same variable is always to be considered the output or response whose prediction is of direct interest, this property is closely related to the fact that the model is able to ignore any predictive variables whose values are unknown in a given observation.

In this work we estimate the maximum-likelihood parameter vector of the mixture networks using the EM algorithm (Dempster, Laird & Rubin, 1977), an iterative procedure alternating between E (expectation) and M (maximization) steps. The E-step is derived for a particular model by taking the expectation over the missing values of the log likelihood, plugging in the previous iteration's values of the parameters where needed only in the coefficients used to form the expectation. (The log likelihood for a data set for a model and parameter set is just the sum of the log $p(\underline{z})$ over for every case/instance.) The M-step is to re-estimate the parameters so as to maximize that expected-log likelihood. In some models EM is equivalent to repeatedly doing single imputation of the expectation of the missing data values themselves in the E-step. Some authors appear to believe that this is true in general; mixture models described in the present paper are among the counterexamples to that belief.

A learned/fitted mixture network can be used directly to make predictions of the output (response) variable of interest, but here we chose to use it as a preprocessing module, imputing (filling in predicted values) (Little & Rubin, 1987) the missing input (predictor) values for use in a multilayer perceptron (MLP) neural net. This decision was based in part on some preliminary comparisons we performed between the two approaches on our data set. Our imputation was stochastic: the missing value predictions were drawn from the trained mixture network's generative distribution given (conditioned on) the non-missing inputs, not for example as a mean or mode of that conditional distribution.

An advantage of an imputation approach is that one can continue to use whatever predictive model/method you have found to perform best in the data domain of interest, while using the mixture network only to allow your method to be used in the presence of missing data, even if it is not able to handle missing data well itself.

The models considered in this paper assume that the data are missing at random (MAR) (Little & Rubin, 1987), meaning (loosely) that the probability of a given value being missing does not depend on that value itself.

Method

We implemented the mixture networks in Xlisp-Stat (Tierney, 1990), a free multi-platform statistical package whose byte-compiled user code can run faster than that of some other comparable environments. We may be able to make our research code available to others.

We report experiments conducted using subsets of the Commission on Cancer Patient Care Evaluation (PCE) breast cancer data. The PCE data were collected by the American College of Surgeons, jointly sponsored with the American Cancer Society, by requesting data from individual tumor registries on the first 25 cases of first diagnosis breast cancer (among others) seen in 1985 at each American College of Surgeons-accredited hospital in the United States.

The purpose of the present work is to determine the performance of the mixture model when almost every case has at least one missing value, not to perform a complete analysis of the PCE breast cancer data.

In the first experiment, we test our Bernoulli mixture network using only three binary inputs (predictive variables) and a binary outcome. The input features were tumor size > 2 cm (1=true, 0=false), lymph node status (1=positive, 0=negative), and distant metastases (1=present, 0=absent). The binarized outcome was simply 5 year status (1=alive, 0=dead). We started with 8,000 cases not having missing values in any of the four variables. This data set was randomly split into three subsets: training data (2,000 cases), generalization/convergence

monitoring for the mixture model (2,000 cases), and a one-time final test set (4,000 cases) for the MLP.

We trained an MLP on the training set cases. The neural network software (NevProp 2) automatically split off half (1,000) of these cases and used them internally for early stopping, a method for automatic regularization/shrinkage to avoid overfitting. The architecture had three input units, two symmetric (logistic - 1/2) sigmoid hidden units, and a single asymmetric (logistic) sigmoid output unit. There was full feedforward inter-layer connectivity, including skips from input to output. Weights were initialized to small values in [-.001 , .001] in order for early stopping to be effective. The training criterion was cross-entropy, optimized by batch gradient descent with an adaptive global learning rate. Weight decay, momentum, and sigmoid-prime offset were all set to small values (.001).

After training, we noted as performance measures the quadratic (squared error per case) and logarithmic (negative log-likelihood per case) proper probability scores on the independent test set (4,000 cases). We then artificially removed 80% of the predictor values in the training set completely at random (MCAR), so that only 20% of the original data values remained. Only about a dozen of the 2,000 training cases remained complete; the others had at least one missing value.

We used the Bernoulli mixture network with only two terms in the mixture and fixed equal mixing probabilities. We used EM to find the maximum-likelihood parameter vector for this model, and then performed a single but stochastic imputation from the conditional distribution of each missing value given the nonmissing predictors in that case and the single parameter vector. It was not deemed necessary to do more than one imputation per case for this experiment. The resulting data set (80% imputed predictors) was again used to train the MLP, and the final performance on the 4,000-case complete test set was noted. Thus this is a test of the ability to train a neural network with imputed data, rather than the ability to make accurate predictions when data must be imputed during performance/recall/testing.

For the second experiment we implemented EM for the Gaussian-Bernoulli mixture model described earlier. The Gaussians' standard deviation parameter ($\sqrt{\text{var}}$) were not allowed to fall below a value of 0.001, in order to prevent the model from falling into a likelihood singularity where one of the terms is trying to fit a single data point exactly. We repeated the procedures of the first experiment, but this time using the original numerical values of the tumor size and log (1 + number of lymph nodes positive) predictors, both of which ranged from about zero to about four in the data and had been binarized only for the first experiment.

The result of using the MLP on the original complete binary data are shown in Table 1 under the column heading MLPFUL. Note that both the quadratic and logarithmic scores are always nonnegative with zero being the best possible value. FRQALIV shows the frequency of the response variable being equal to 1 (i.e. status at 5 years = ALIVE) in each data set, and PRDFRQ shows as a baseline the result of naively predicting this frequency for all cases in the data set, regardless of their input values.

After randomly removing 80% of the training set data elements, we use the mixture model to stochastically impute these for the MLP, and the results are given in the column labeled MLP - .8 . We see that even with a huge fraction of the data missing, the results are not much worse than with all the original data.

TABLE 1. Test-set results from first experiment: binary data.

(per case)	FREQALIV	PRDFRQ	MLPFUL	MLP-.8
Squared error	.808	.155	.138	.139
log likelihood	.808	.489	.436	.443

TABLE 2. Test-set results from first experiment: continuous and binary data.

(per case)	FREQALIV	PRDFRQ	MLPFUL	MLP-.8
Squared error	.808	.155	.133	.135
log likelihood	.808	.489	.427	.434

Table 2 shows the results of the second experiment, where two of the predictors are now continuous. Though the 8000 original cases are the same ones as in the first experiment, they happened to be randomly split into train and test subsets independently for the two experiments. Again, the test set performance does not worsen drastically upon removing 80% of the training set data values.

We conclude that for our data sets, the mixture networks allow us to handle well a very large proportion of missing values in the training data. Of course this could not have been the case if the remaining 20% did not contain sufficient information in comparison to the complete data set; thus our 2000 cases presumably contained a great deal of redundant information. But even given this redundancy, one could easily have lost nearly all predictive ability depending on the methods employed. For example, dropping incomplete cases from the training data was clearly not viable here. The simplest methods, such as imputing the single overall (data set-wide) mode of a binary variable where ever it is missing, could have grossly distorted the dependencies in the data distribution, so as to overwhelm the valid information in the much smaller collection of non-missing values.

The imputation approach may sometimes produce results superior to direct prediction of the outcome by the single mixture network, if the imputed data are used in a response-predictive type of model such as the original NRBF, generalized linear models (GLiM), multilayer perceptrons (MLP), etc. The latter models are estimated by maximum conditional or predictive likelihood, where the (vector) parameter θ is chosen to maximize the likelihood of the y data conditional on the $\underline{x}$ data, i.e. $p(D_y \mid D_{\underline{x}};\theta)$ According to Efron (1979) and others, these often make better predictions because they are more robust to violations of distributional assumptions, since these assumptions are less stringent than in the case of full joint probability models estimated by full maximum likelihood $p(D_y, D_{\underline{x}};\mid \theta)$.

In addition, projective models such as MLPs, GLiMs, and projection pursuit regression are often seen to perform better in practice, especially with many predictor variables, than their "local radial" and mixture cousins. There is at least one theoretical analysis (Barron, 1994) explaining how the MLP can perhaps perform well in spite of the curse of dimensionality; we are unaware of any similar results for mixture/radial models.

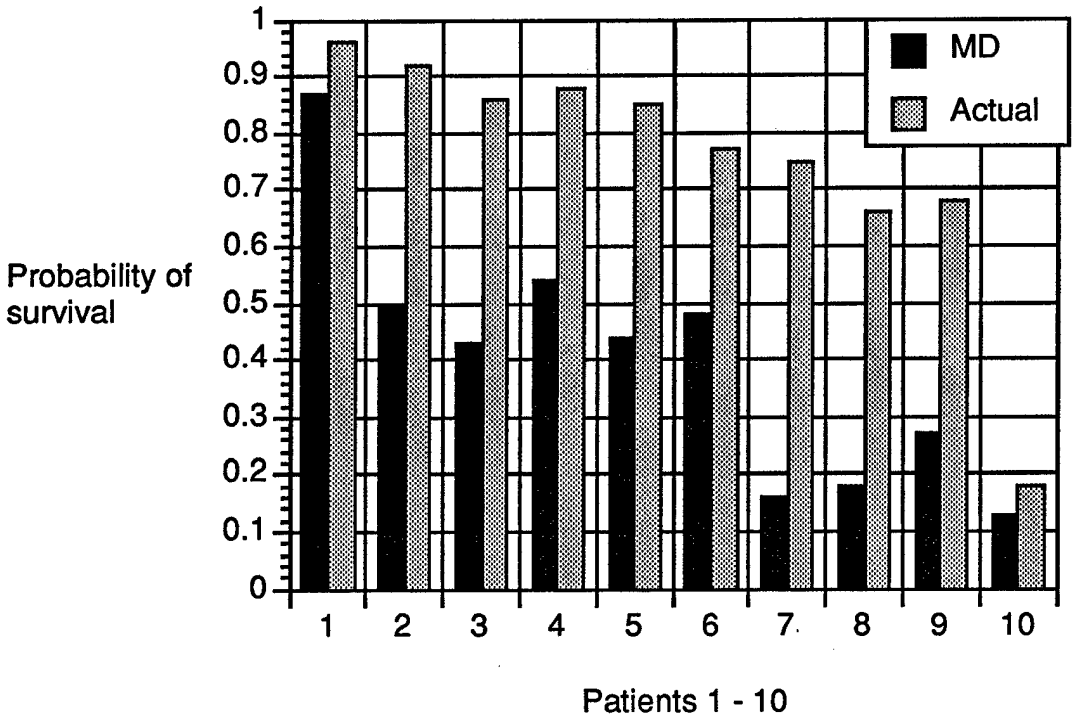
Generating stochastic imputations from the predictive distribution of the missing values given the nonmissing values and the single maximum-likelihood parameter vector is not a proper Bayesian stochastic imputation (Little & Rubin, 1987) because we do not perform the additional computation required to take into account the uncertainty in (i.e. posterior distribution of) the parameter vector itself. However, it should still be better to stochastically account for (some of) the uncertainty in each missing value itself, rather than deterministically imputing a single central

tendency (mean or mode) for the value once given the non-missing predictors, so the method we used can be viewed as a pragmatic compromise between simpler multivariate-predictive methods and a rigorous Bayesian method. The data augmentation algorithm of Schafer (1995) is an example of such a Bayesian method employing stochastic simulation, and can be applied to mixture networks as a relatively straightforward extension of the EM implementation, often initialized with the parameter vector found by EM. Neal (1991) has implemented and studied certain priors and stochastic simulation methods for discrete data.

Recent work on mixtures of factor analyses (Ghahramani & Hinton, 1996) may be one way to permit relaxation of the independence-within-term (i.e. diagonal covariance matrix) constraint on the continuous variables even in high dimension where a full covariance matrix (general multivariate Gaussian) cannot be used. This also makes the overall method partly projective, bringing some its properties closer to those of, e.g., the MLP.

Implementation of Clinical Decision Support System

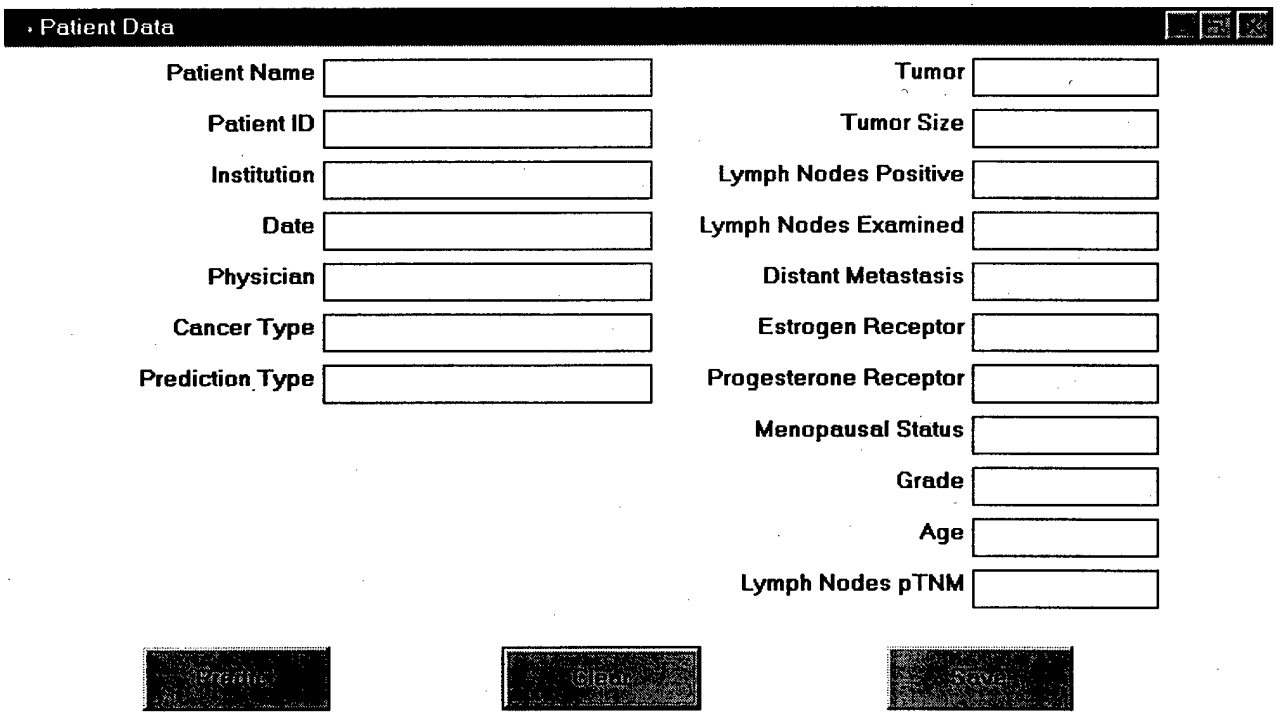
Because of all the work we had done on outcome prediction accuracy, we were interested in how good clinicians are at predicting patient outcomes. Research has suggested that physicians are not able to accurately assess breast cancer survival. (Laprinzi, 1994) We performed a small survey that assessed the oncologists' ability to predict five year breast cancer specific survival. Oncologists were asked to estimate ten patient's five year breast cancer specific survival. (Survey instrument is presented in Appendix). The mean of the oncologists' 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 (unpublished results) has important implications for patient care.

The development of the clinical decision support system on the Windows platform has been very demanding. We began by developing the artificial neural network and missing data algorithms in interpreted languages (e.g., XLISP STAT). We then rewrote the algorithms in a complied language (Borland C++ and later Visual C++). We trained the missing data algorithm to perform multiple imputation on the SEER breast cancer data set. Using these data we trained our artificial neural network to predict five year breast cancer survival. We then linked the trained artificial neural network program to the trained missing data program so that if a patient does not have all her prognostic factors the missing data algorithm will automatically fill-in the missing factor(s) and the patient will still receive a survival prediction. The prediction she receives will not be as accurate as one that contains all the prognostic factors but it is better than not providing any prediction.

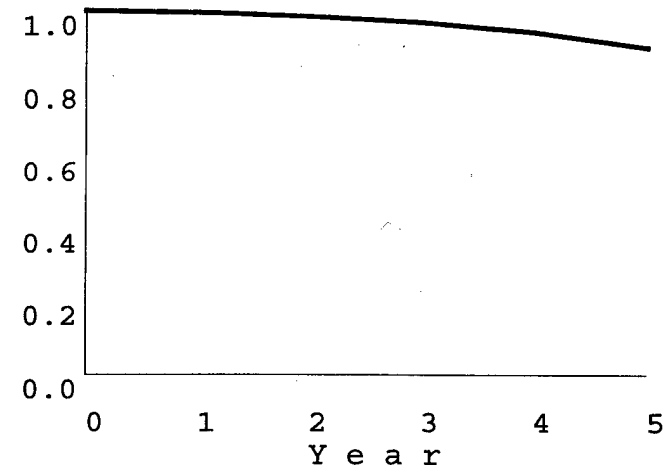
While we were developing the above algorithms and training the models we also began developing the input and output graphical user interfaces (GUI). The GUIs were first written in C++. But we found that there were problems with controlling the placement of information and the printing of the C++ screens. The problem was one of position, we could not obtain position-invariance across different computers and printer drivers. We rewrote the GUIs in Visual Basic and this resolved the printer problem. The input screen is shown below.



The input screen is linked to an Oracle data base and to both our ANN and missing data models. When patient information is entered it goes to the data base for storage, the missing data algorithm to fill-in missing data fields (this information is also sent to the ANN), and to the ANN for survival predictions. The output screen was more difficult to implement. Drawing curves in exact locations is relatively difficult to do in C++ and this took a great deal of time to implement. We are currently rewriting the input and output GUIs in Java. The output screen is shown below.

```
Patient Name:                          Variables:
Patient ID:                            Tumor:
Institution:                           T Size:      3      LN Pos:      1
Date:                                  LN Exam:     15     Mets:        n
Physician:                             ER:          y      PR:          y
Cancer Type:                           Menopausal:  y      Grade:       2
Prediction Type:                       Age:         67     LN pTNM:     2
```

Year	ANN Prediction
0	1.000
1	0.993
2	0.984
3	0.967
4	0.94
5	0.9




```
TNM Stage:   IIIA              Prediction:   0.595
```

During the development of the GUIs we consulted with clinicians (oncologists, pathologists, surgeons, radiation oncologists) to determine the optimal ergonomic approach.

We have demonstration projects with New York Medical College and Duke University Medical Center. We had hoped to have, by now, a large data set from Duke with all therapies represented and with long term follow-up. We have spent a great deal of time working with Duke and its cancer registry on this data set. The problems we have encountered included the lack of complete TNM data at the tumor registry, the need to go to the patient's charts for treatment information, and a paucity of recent follow-up information. We have used a data set with just surgery as the treatment to test our system. We are continuing to work with Duke and will create a multi-treatment clinical decision support system when the data becomes available.

With the surgical data set we have learned a great deal. The most important finding is that, generally, physicians will not directly interact with the system. There are a number of reasons for this finding. One is that many physicians are not interested in, or able to, interact with computers. A second reason is that physicians are very busy and do not want to take the time to interact with the system. This includes the physician not acquiring and entering the patient's prognostic factors.

The way we have solved this problem is to have ancillary personnel enter the factors, print out the patient's survival probability report, and place the report in the patient's chart.

Once in the chart, the report is always used. We have found that both the physician and the patient value objective clinical information that is patient-specific. In our continuing survey of physicians and patients, the report is always rated highly. Most comments are that it is valuable and that it strengthens the physician-patient relationship. We have not had the opportunity observe the benefit of the reports in decision-making since all women receive surgery. When the data set with adjuvant treatments becomes available we will be able to provide treatment comparisons. For example, which women would benefit from the combination of chemotherapy and hormonal treatment.

Tasks added to the project

(1) We computerized the TNM staging system and integrated its predictions into the prognostic system.

(2) We completed our comparison of the two national breast cancer data bases, the National Cancer Data Base (NCDB) and its associated Patient Care Evaluation (PCE), and the Surveillance, Epidemiology, and End Results (SEER) data sets. We evaluated them in terms of: (i) representativeness, is the data set an unbiased representation of the breast cancer population. (ii) Incidence/prevalence, how good is the data set in capturing the incidence and prevalence of breast cancer. (iii) Prognosis/outcome, how good is the data set in providing information that is useful for predicting outcome. An overview of the results are shown below.

	SEER	NCDB
Representativeness	good	good
Incidence/prevalence	good	adequate
Prognosis/outcome	not acceptable	adequate

Both are representative of the breast cancer population. SEER does a better job at incidence and prevalence because it ascertains all cases in a catchment area, regardless of whether the hospital belongs to the American College of Surgeons accreditation program, and it contains relatively little missing data. NCDB contains a great deal of missing data. SEER can not be used for prognosis because it does not provide therapy data due to the unreliability of their data. NCDB suffers from a lack of follow-up, resulting in high censoring rates.

CONCLUSIONS

The current breast cancer prediction system, the pTNM staging system, is only moderately accurate, yet the TNM variables are probably more accurate than any three variables in most diseases. Neural networks can increase the predictive accuracy of the TNM variables beyond that possible in the pTNM stage model and additional variables, especially molecular genetic factors, can be added to neural networks to further increase prognostic accuracy.

The current pTNM stage system is about 44% accurate in its breast cancer five year survival predictions. By using a neural network, and by adding variables that individually have little prognostic value, accuracy can be increased to 56%. This is a 27% increase in accuracy, without adding any new prognostic factors. With the addition of two new prognostic factors, p53 and HER-2/neu, accuracy increases to 70%. This increase in predictive accuracy, from 44% with the

pTNM model to 70% with neural networks integrating new prognostic factors, is a 60% improvement in our ability to predict outcome.

A 60% improvement in prognostic ability is clinically important for therapy, clinical trials, patient information, and quality assurance. (1) Therapy. It will allow the more efficient separation of node negative women who have a poor prognosis and require additional therapy, from women who have an excellent prognosis and require no adjuvant therapy. (2) Clinical trials. It will allow researchers to create more homogenous patient populations. This homogeneity will decrease interpatient variability and thus allow therapeutic trials to detect small but clinically important differences in response to therapy; responses that before would have not been statistically significant in prior clinical trials. Further, the increase in accuracy means that smaller patient populations are needed for clinical trials. (3) Patient information. Women will have a more accurate understanding of their disease, and can plan their lives accordingly. (4) Quality assurance. It will provide a better adjustment of severity of illness than is currently possible.

Cancer is primarily a genetic disease.(Rowley, 1993) Cancer genes do not act in isolation; oncogenes, suppresser genes, and genetic mutations cause cancer thorough the complex interaction of the genes and their products.(Postel, 1993; Steel, 1993) A cascade of genes is required to produce a cancer.(Lippman, 1990) Thus, we can not assume: that a gene or its product will have an independent prognostic value before it is combined with other genes and/or their products, that gene interactions are binary, or that there will only be a few simple genetic interactions. Further, we not can specify in advance of the analysis what complex genetic interactions will occur. We need to capture these complex interactions because the prognostic value of the genes and their products can depend on their interactions.(Smith, 1993) Because neural networks can approximate any continuous function to any degree of accuracy,(Hornik, 1989; Leshno, 1993) they can discover these complex interactions without the requirement of a priori specification of the important variables; the neural network will learn the variables that are important. Neural networks are able to capture the power of nonmonotonic prognostic factors and they are efficient discoverers of complex interactions. Neural networks can do everything that linear and logistic regression can do, and they can do much more.

We have proposed the following criteria for selecting an enhanced prognostic system: (1) Easy for physicians to use. (2) Provides predictions for all types of cancer. (3) Provides the most accurate relapse and survival predictions at discovery and for every year lived for each patient. (4) Provides group survival curves, where the grouping can be by any variable including outcome and therapy. (5) Accommodates missing data and censored patients, and it is tolerant of noisy and biased data. (6) Makes no a priori assumptions regarding the type of data, the distribution of the variables, or the relationships among the variables. (7) Can test putative prognostic factors for significance, independence, and clinical importance. (8) Accommodates treatment information in the evaluation of prognostic factors. (9) Accommodates new putative prognostic factors without changing the model. (10) Accommodates emerging diagnostic techniques. (11) Provides information regarding the importance of each predictive variable. (12) Is automatic. This report is the first step in implementing the prognostic system in the form of a clinical decision support system.(Burke, 1993)

The system we have created is efficient, accurate, and useful. It provides objective and important patient-specific clinical information for the patient and her physician to use in their understanding and treatment of the patient's disease.

Problems encountered in accomplishing tasks

The primary problem we encountered the paucity of data sets with molecular genetic variables, long term follow-up, and a sufficient number of cases receiving a particular treatment regimen.

REFERENCES

Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psych 1975;12:387-415.

Baxt WG. Application of artificial neural networks to clinical medicine. Lancet 1995;346:1135-38.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Pacifistic Grove, CA; Wadsworth and Brooks, 1984.

Breslow NE. Covariance analysis of censored survival data. Biometrics; 1974:80-99.

Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. Cancer 1993;72:3131-5.

Burke HB. Artificial neural networks for cancer research: outcome prediction. Sem Surg Onc 1994;10:73-79.

Burke HB, Hutter RVP, Henson DE. Breast Carcinoma. In P. Hermanek, L.H. Sobin, D.E. Henson, R.V.P. Hutter, M. Gospadoriwicz, eds. UICC Prognostic Factors in Cancer. Berlin: Springer-Verlag, 1995a, 165-176.

Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In G. Tesauro, D.S. Touretzky, T.K. Leen, eds. Advances in Neural Information Processing Systems 7. Cambridge, MA; MIT Press, 1995b, 1063-67.

Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997a;79:857-62.

Burke HB. Henson DE. Histologic grade as a prognostic factor in breast carcinoma. Cancer 1997b;80:1703-1705.

Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. Cancer 1998a;82:874-7.

Burke HB, Henson DE. Specimen banks for prognostic factor research. Arch Path Lab Med, 1998b;122:871-874.

Burke HB. Integrating multiple clinical tests to increase predictive accuracy. In M. Hanausek, Z. Walaszek (eds), Methods in Molecular Biology:, Vol. XX: Tumor Marker Protocols. Totowa, N.J., Humana Press Inc., 1998c, Chapter 1, 3 - 10.

Carpenter GA, Grossberg S, Rosen DB. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks 1991;4:759-771.

Cox DR. Regression models and life-tables (with discussion). J Royal Stat Soc B;1972:187-220.

Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995;346:1075-1079.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM Algorithm (with discussion). J. Roy. Statist. Soc. 1977;(B39):1-38.

Dempster AP, Rubin DB. Overview. In: Madow WG, Olkin I and Dubin DB (eds.) Incomplete data in sample surveys, Vol II: Theory and annotated bibliography. New York: Academic Press, 1983, 3-10.

Donegan WL. Staging methods, primary treatment options and end results. Major problems in clinical surgery 1979;5:221-301.

Dybowski R. Gant V. Artificial neural networks in pathology and medical laboratories. Lancet 1995;346:1203-1207.

Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993.

Efron B. Bootstrap methods: another look at the jackknife. Am Statist 1979;7:1-26.

Fahlman SE, Lebiere C. The Cascade-correlation learning architecture. National Science Foundation contract EET-8716324, Carnegie Mellon University, 1991.

Fleming ID, Cooper JS, Henson DE, et al. (eds.) AJCC Cancer Staging Manual, Fifth Edition. Phildelphia: Lippincott-Raven, 1997.

Gabor AJ, Seyal M. Automated interictal EEG spike detection using artificial neural networks. Electroencephalogr Clin Neurophysiol 1992;83:271-80.

Ghahramani Z, Jordan MI. Supervised learning from incomplete data via an EM approach. In: Cowan JD, Tesauro G and Alspector J (eds.) Advances in neural information processing systems. San Francisco, CA: Morgan Kaufman., 1994, 120-27.

Ghahramani Z, Hinton, GE. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Univ. of Toronto, Dept. of Computer Science, Toronto, 1996.

Halsted WS. The results of operations for the cure of the breast performed at the Johns Hopkins Hospital for June 1989 to January 1984. Ann Surg 1894-5;20:497-555.

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

Hartley HO. Maximum likelood estimation from incomplete data. Bimetrics 1958;14:174-94.

Hebb DO. The Organization of Behavior. New York: Wiley, 1949.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989;2:359-66.

Hornik K, Stinchcombe M, White H. Universal approximation of an unknown function and its derivatives using multilayer feedforward networks. Neural Networks 1990;3:551-60.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457-81.

Hornik K, Stinchcombe M, White H, Auer P. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation 1994;6:1262-75.

Laprinzi CL, Ravdin PM, DeLaurentiis M, Novotny P. Do American oncologists known how to use prognostic variables for patients with newly diagnosed primary breast cancer? J Clin Onc 1994;12:1422-26.

Leong PH, Jabri MA. MATIC - an intracardiac tachycardia classification system. PACE 1992;15:1317-31.

Leshno M, Lin VY, Pinkus A, Schocken S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 1993;6:861-67.

Liestol K, Anderson PK, Anderson U. Survival analysis and neural nets. Stat Med 1994; 13:1189-1200.

Lippman SM, Lee JS, Lotan R, Hittelman W, Wargovich MJ, Hong WK. Biomarkers as intermediate end points in chemoprevention trials. J Natl Cancer Inst 1990;82:555-60.

Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley, New York, 1987.

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

Marks JR, Humphrey PA, Wu K, Berry D, Bandarenko N, Kerns BM, Inglehart JD. Overexpression of the p53 and HER-2/neu proteins as prognostic markers in early stage breast cancer. Ann Surg 1994;219:332-41.

McKendrick AG. Applications of mathematics to medical problems. Proc. Edinburgh Math. Soc. 1926;(44):98-130.

Moody J, Darken C. Fast learning in networks of locally-tuned processing units. Neural Computation, 1989;1: 281-294.

Moody J, Darken C. Learning with localized receptive fields. In Touretzky, D., Hinton, G., & Sejnowski, T., editors, Proceedings of the 1988 Connectionist Models Summer School, 1988, 133-143, San Mateo. (Pittsburgh 1988), Morgan Kaufmann.

Nadaraya EA. On estimating regression. Theory of Probability and its Appl., 1964;10:186-190.

Neal R M. Bayesian mixture modeling by monte carlo simulation. Technical Report CRG-TR-91-2, Univ. of Toronto, Dept. of Computer Science, Toronto, 1991.

Nowlan SJ. Soft Competitive Adaptation: Neural Network Learning Algorithms, Based on Fitting Statistical Mixtures. PhD thesis, Carnegie Mellon University, 1991.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In Touretzky, D. S., editor, Advances in Neural Information Processing Systems 2 (Proc. of the 1989 Conference), 574-582. San Mateo, CA: Morgan Kaufmann.

Orchard T, Woodbury MA. A missing information principle: theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1972, 697-715.

Parzen E. On estimation of a probability density and mode. Ann Math Stat. 1962; 35: 1065-1076.

Poggio T Girosi F. A theory of networks for approximation and learning. Technical Report A.I. Memo No. 1140, Massachusetts Institute of Technology AI Laboratory, Cambridge, MA, 1989.

Postel EH, Berberich SJ, Flint SJ, Ferrone CA. Human c-myc transcription factor PuF identified as nm23-H2 nucleoside diphosphate kinase, a candidate suppresser of tumor metastasis. Science 1993;261:478-80.

Preisler HD, Raza A. The role of emerging technologies in the diagnosis and staging of neoplastic diseases. Cancer 1992; 69:1520-1526.

Ravdin PM, Clark GM. A practical application of neural network analysis for prediction outcome of individual breast cancer patients. Breast Cancer Res Treat 1992; 22:285-293.

Rosen DB. How good were those probability predictions? The expected recommendation loss (ERL) scoring rule (8 pp.). In: Maximum Entropy and Bayesian Methods: Proceedings of the Thirteenth International Workshop (August 1993), G. Heidbreder, ed., Kluwer, Dordrecht, The Netherlands, 1995a.

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

Rosen DB. Issues in selecting empirical performance measures for probabilistic classifiers. To appear in Maximum Entropy and Bayesian Methods: Proceedings of the Fifteenth International Workshop (July/August 1995), K. Hanson and R. Silver, eds., Kluwer, Dordrecht, The Netherlands, 1996.

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

Rosenblatt M. Remarks on some non-parametric estimates of a density function. Ann. Math. Stat. 1956; 27: 642-669.

Rowley JD, Aster JC, Sklar J. The clinical applications of new DNA diagnostic technology on the management of cancer patients. JAMA 1993;270:2331-37.

Savage LJ. Elicitation of personal probabilities and expectations. J Am Stat Assoc 1971;66:783-801.

Schafer, J. Analysis of Incomplete Multivariate Data by Simulation. Chapman & Hall, London, 1995.

Smith K, Houlbrook S, Greenall M, Carmichael J, Harris Al. Topoisomerase II alpha co-amplification with erbB-2 in human primary breast cancer and breast cancer sell lines: relationship to M-AMSA and mitoxantrone sensitivity. Oncogene 1993;8:933-38.

Specht DF. Probabilistic neural networks. Neural Networks 1990;3:109-18.

Specht DF. A general regression neural network. IEEE-TR-NN 1991;2(6):568-576.

Steel M. Cancer genes: complexes and complexities. Lancet 1993;342:754-5.

Swets JA. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. New Jersey: Lawrence Erlbaum, 1996.

Tierney L. LISP-STAT: An object-oriented environment for statistical computing and dynamic graphics. New York: John Wiley & Sons, 1990.

Titterington DM, Smith AFM, Makov UE. Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester, 1985.

Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. Radiology 1993;189:555-58.

Tresp V, Ahmed S, Neuneier R. Training neural networks with deficient data. In JD Cowan, G Tesauro, J Alspector (eds.), Advances in Neural Information Processing Systems, 6. San Francisco, CA: Morgan Kaufman, 1994, 128-35.

Vamplew P, Adams A. Real world problems in backpropagation: missing values and generalizability. Third Australian Conference on Neural Networks, 1992.

von Osdol W, Myers TG, Paull KD, Kohn KW, Weinstein JN. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. J Natl Cancer Inst 1994;86:1853-59.

Watson, GS. Smooth regression analysis. Sankhya 1964; A26: 359-372.

Weiss SM, Kulikowski CA. Computer systems that learn. San Mateo, CA: Morgan Kaufmann. 1991.

Westenskow DR, Orr JA, Simon FH. Intelligent alarms reduce anesthesiologist's response time to critical faults. Anesthesiology 1992;77:1074-79.

APPENDIX

APPENDIX - 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, _____ oncologic 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).

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

Predicting Response to Adjuvant and Radiation Therapy in Patients with Early Stage Breast Carcinoma

Harry B. Burke, M.D., Ph.D.[1]
Albert Hoang, Ph.D.[1]
J. Dirk Iglehart, M.D.[2]
Jeffrey R. Marks, Ph.D.[2]

[1] Department of Medicine, New York Medical College, Valhalla, New York.

[2] Departments of Surgery, Pathology, and Cell Biology, Duke University, Durham, North Carolina.

Supported in part by a research grant from the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383) and by the Duke Specialized Program of Research (SPORE) of the National Cancer Institute.

Address for reprints: Harry B. Burke, M.D., Ph.D., Department of Medicine, New York Medical College, Valhalla, NY 10595.

Received March 14, 1997; revision received July 31, 1997; accepted September 8, 1997.

BACKGROUND. Screening and surveillance is increasing the detection of early stage breast carcinoma. The ability to predict accurately the response to adjuvant therapy (chemotherapy or tamoxifen therapy) or postlumpectomy radiation therapy in these patients can be vital to their survival, because this prediction determines the best postsurgical therapy for each patient.

METHODS. This study evaluated data from 226 patients with TNM Stage I and early Stage II breast carcinoma and included the variables p53 and c-*erb*B-2 (HER-2/*neu*). The area under the receiver operating characteristic curve (Az) was the measure of predictive accuracy. The prediction endpoints were 5- and 10-year overall survival.

RESULTS. For Stage I and early Stage II patients, the 5- and 10-year predictive accuracy of the TNM staging system were at chance level, i.e., no better than flipping a coin. Both the 5- and 10-year artificial neural networks (ANNs) were very accurate—significantly more so than the TNM staging system (Az 5-year survival, TNM = 0.567, ANN = 0.758; $P < 0.001$; Az 10-year survival, TNM = 0.508, ANN = 0.894; $P < 0.0001$). For patients not receiving postsurgical therapy and for either chemotherapy or tamoxifen therapy, the ANNs containing p53 and c-*erb*B-2 and the number of positive lymph nodes were accurate predictors of survival (Az 5-year survival, 0.781, 0.789, and 0.720, respectively).

CONCLUSIONS. The molecular genetic variables p53 and c-*erb*B-2 and the number of positive lymph nodes are powerful predictors of survival, and using ANN statistical models is a powerful method for predicting responses to adjuvant therapy or radiation therapy in patients with breast carcinoma. ANNs with molecular genetic prognostic factors may improve therapy selection for women with early stage breast carcinoma. *Cancer* **1998;82:874–7.** © *1998 American Cancer Society.*

KEYWORDS: TNM staging system, artificial neural networks, prognostic factors, breast carcinoma, tamoxifen therapy, chemotherapy, radiation therapy, outcomes, c-*erb*B-2, p53.

Screening and surveillance is increasing the prevalence of early stage breast carcinoma. The ability to predict accurately the responses to adjuvant therapy (chemotherapy or tamoxifen therapy) or postlumpectomy radiation therapy in these patients can be vital to their survival, because this prediction determines the best postsurgical therapy for each patient. The pathologic TNM staging system is the current cancer prognostic system. Its predictions are based on three variables: 1) location, size, and depth of tumor; 2) existence and location of involved lymph nodes; and 3) existence of distant metastases.[1] We have shown that artificial neural networks (ANNs)

are more accurate at predicting survival than the TNM staging system for all stages of breast carcinoma.[2] It is not known how accurate the TNM staging system is in predicting the survival of patients with early stage breast carcinoma. It is also not known whether ANNs with molecular genetic prognostic factors, i.e., p53 and c-*erb*B-2 (HER-2/*neu*), can improve prognostic accuracy in early stage breast carcinoma across postsurgical therapies and for specific therapies. This article compares the survival prediction accuracy of the TNM staging system with ANN models across all postsurgical therapies. In addition, it presents a method for properly assessing putative therapy-dependent prognostic factors and examines the accuracy of ANNs in terms of specific therapies. Because the TNM staging system does not predict response to adjuvant or radiation therapy, it is not included in the individual therapy analyses.

METHODS

Data

These data were described in detail in a previous article.[3] Briefly, all patients were pathologic TNM Stage I or early Stage II. Early stage breast carcinoma includes Stage I and limited Stage II. Limited Stage II included all the TNM Stage II patients except those with five or more positive lymph nodes. The variables were age, race, tumor size, lymph nodes positive, lymph node stage, nuclear grade, histologic grade, p53, c-*erb*B-2, estrogen receptor (ER) and progesterone receptor (PR) status, vascular invasion, adjuvant therapy (tamoxifen or chemotherapy), and radiation therapy. Patients who underwent a lumpectomy received radiation therapy. Patients who underwent a modified radical mastectomy did not receive radiation therapy. There were 229 cases, of which 226 had complete data for all variables except ER and PR status. Because of the number of cases missing, both ER and PR were removed from the data set. The survival rate was 70%. The prediction endpoints were 5-and 10-year overall survival.

Accuracy

The area under the receiver operating characteristic curve (Az) is a measure of prediction accuracy.[4] It can be used to assess and compare the adequacy of statistical models. Az can be directly calculated by Somer's D,[5] or it can be approximated by its trapezoidal area.[6] The area under the curve is a nonparametric measure of discrimination. It is independent of both the prior probability of each outcome and the threshold cutoff for category. Its computation requires only that the prediction method produce an ordinally scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the prediction method will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the predictions are unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the score is from 0.5, the better, on average, the prediction method is for predicting which of the two patients will be alive.

TABLE 1
Comparison of the Accuracy of the TNM Staging System and Artificial Neural Networks in Predicting the 5- and 10-year Survival of Patients with Early Stage Breast Carcinoma

Model	5-yr survival Az (SE)[a]	10-year survival Az (SE)[b]
TNM	0.567 (0.046)	0.508 (0.053)
ANN	0.758 (0.042)	0.894 (0.034)

ANN: artificial neural network; Az: area under the receiver operating characteristic curve; SE: standard error.
[a] TNM vs. ANN 5-year survival, $P < 0.001$.
[b] TNM vs. ANN 10-year survival, $P < 0.0001$.

Statistical Models

ANN models have been described in detail elsewhere.[2] Briefly, the three-layer backpropagation ANN was composed of an input layer, a hidden layer, and an output layer. Each layer of an ANN was composed of nodes. The number of input nodes was equal to the number of variables. The hidden layer was composed of three nodes. There was one output node. All the variables were entered into the three-layer ANN model. The two-layer ANN was identical to the three-layer network, except that it did not possess a hidden layer. After a sensitivity analysis to reduce the number of input variables to the three with the highest predictive accuracy, the three selected variables, namely, the number of positive lymph nodes, p53, and c-*erb*B-2, were entered into the two-layer ANN. Both the two- and three-layer ANNs employed the maximum likelihood loss function and weight decay. Model accuracy estimates and standard errors were calculated by the bootstrap resampling method.[7]

RESULTS

The predictive accuracies of the TNM staging system and the three-layer ANN models are shown in Table 1. For Stage I and early Stage II patients, the 5- and 10-year prediction accuracy of the TNM staging sys-

condition due to receipt of an effective therapy. For example, ER status may predict response to tamoxifen. Posttherapy prognostic factors predict, after the patient has received the therapy, whether there has been a change in the course of the disease due to the intervention. For example, the number of positive lymph nodes on axillary dissection may predict whether the patient will respond to the primary surgery. Posttherapy prognostic factors are important because we do not want to wait any longer than necessary to administer a second-line therapy to patients who do not respond to the primary therapy. All three prognostic factors are relative to therapy. For each therapy in a succession of therapies (for example, if a therapy is given and the patient does not respond to that therapy and another therapy is contemplated), all three types of prognostic factors can be analyzed.

Within the context of the small sample size of this study, the molecular genetic variables p53 and c-*erb*B-2 are powerful therapy-dependent prognostic factors for early stage breast carcinoma, and ANN models are an efficient statistical method for capturing their predictive power.

REFERENCES

1. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors.. Manual for staging of cancer. American Joint Committee on Cancer. 4th edition. Philadelphia: J. B. Lippincott, 1992.
2. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79: 857–62.
3. Marks JR, Wu K, Berry D, Bandarenko N, Kerns BJ, Iglehart JD. Overexpression of p53 and HER-2/*neu* proteins as prognostic markers in early stage breast cancer. *Ann Surg* 1994; 219:332–41.
4. Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
5. Somers RH. A new asymmetric measure of association for ordinal variables. *Am Sociological Rev* 1962;27:799–811.
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic. *J Math Psy* 1975;12:387–415.
7. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993:45–57.
8. Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors. *J Cell Biochem* 1994;19S:278–82.

EDITORIAL

Counterpoint

Histologic Grade as a Prognostic Factor in Breast Carcinoma

Harry B. Burke, M.D., Ph.D.[1]
Donald Earl Henson, M.D.[2]

[1] Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, New York.

[2] Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, Maryland.

In this issue of *Cancer,* Dr. Roberti reviews the role of histologic grade in the prognosis of breast carcinoma and wonders why, because it is available, it has not been widely used in predicting outcome.[1] The position of this editorial is that there must be some fundamental reason, after 100 years of progress on histologic grade, that confusion persists regarding its prognostic value.

The systematic use of morphologic variation at the cellular level of analysis as a prognostic factor in cancer has been fraught with controversy. Currently, there is no universally agreed on set of necessary and sufficient conditions for the definition of histologic grade in breast carcinoma. There has been uncertainty regarding the identification of what variation was important, how the variation should be organized, and whether it should be integrated into a staging or index system.

An additional issue is that grading system criteria have been selected based on their ability to create subgroups of patients using histologic distinctions to produce significant differences in outcome. There are two problems with this approach. First, there are many possible criteria that can create significant differences between subgroups and there is no analytic method for finding the best criteria.[2] Second, statistical significance is not necessarily accuracy. Significance is the chance that two or more distributions of variables, as represented by their parameter estimates, for example, means and variances, are really the same. Accuracy assesses the strength of association between two or more variables.[3,4] In general, accuracy quantifies how good a variable is at predicting another variable. Specifically, we are interested in the strength of association between grade and survival, i.e., how good is grade at predicting survival.

Fundamentally, grade remains controversial because it confounds two types of time. One type is how long the tumor has been growing and the other is how rapidly it has been growing. A "high grade" tumor could be an indolent tumor that grew for a long time prior to discovery and will continue to be slow growing; alternatively, it could be an aggressive tumor of recent origin that will continue to be rapidly growing. Because one can never know when a tumor originated, it may not be possible on histologic grounds to separate

Supported in part by a research grant from the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383).

See reply to counterpoint on pages 1706–7 and referenced original article on pages 1708–16, this issue.

Address for reprints: Donald Earl Henson, M.D., Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892.

Received May 9, 1997; accepted May 15, 1997.

a slowly growing tumor from a rapidly growing tumor. In other words, one cannot always distinguish how long the tumor has been growing from how fast it has been growing. The extent to which time ambiguity exists in grade is the extent to which grade's prediction variance will increase and consequently the extent to which its prediction accuracy will decrease. This limits grade's independent prognostic value and its ability to add significant prognostic value when placed in a system that includes other time-related factors such as tumor size.

The mechanical theory of cancer, a view espoused by Halsted,[5] assumes that cancer spreads from the primary tumor to the regional lymph nodes and then to distant sites of the body. This view is the basis of the TNM staging system. For the mechanical theory, the primary purpose of a prognostic system is to capture the spread of the cancer because cancer spread is believed to be the best indicator of outcome. The three elements of the TNM staging system (local tumor, regional lymph node, and distant metastasis[6]) are believed to reflect directly the spread of cancer, i.e., the extent of disease. Grade is not one of the TNM variables because it does not fit into this mechanical epistemology; it does not directly reflect the spread of the cancer. However, even if grade could have been subsumed within the mechanical theory of cancer, it would not have replaced tumor size in breast carcinoma. Using the Surveillance, Epidemiology, and End Results data of the National Cancer Institute for 1983–1987 and the area under the receiver operating characteristic curve as the measure of accuracy (Az), we found the Az for grade alone to be .634 and the Az for tumor size alone to be .737 ($P < .05$) for 5-year survival. Furthermore, grade does not add prognostic accuracy to tumor size; the Az for tumor size and grade combined was .749, which was not significant when compared with tumor size alone. In addition, grade could not have been added to the TNM staging system because the system is a bin model comprised of five levels of tumor characteristics (T), four levels of regional lymph node involvement (N), and two levels of distant metastasis (M).[7] Adding the 4 levels of grade to the 40 bins of the TNM ($5T \times 4N \times 2M$) would have created 160 bins and made it too complex to be useful.[7]

What is the future of grade as a prognostic factor in breast carcinoma? If we no longer accept the mechanical theory of cancer spread, grade becomes a possible prognostic factor. In addition, because the TNM staging system is not very accurate[8] new computer-based prognostic systems are being developed.[7] Computer-based prognostic systems are more accurate in predicting outcome and they do not have a limitation on the number of variables that can be used. Can grade be an independent prognostic factor in a computer-based system or can grade substitute for another more difficult to assess factor such as lymph node status?

We evaluated the ability of grade to predict 5-year breast carcinoma survival using data from the National Cancer Institute's SEER program.[9] The data were collected between 1983–1987 and the patients were followed for at least 5 years. The variables were tumor size, local extent of disease, lymph node status, and histologic grade. The criteria used to determine grade were neither standardized nor explicitly reported. The data set did not include cases with metastatic disease because grade is infrequently reported in these patients. Only 14,704 of the 48,643 cases were graded (30%). All analyses without grade were performed on the full data set of 48,643 cases. An analysis using the subset of graded cases favors grade because it is almost certainly the case that the variance of grade would increase if all the cases were graded. The area under the receiver operating characteristic curve was the measure of prediction accuracy. We used the logistic regression statistical method to create our models (SAS Institute, Cary, NC) and all results were performed on the test data set.

The predictive accuracy of tumor size, local tumor extent, and lymph status was .794. Adding histologic grade slightly increased the Az to .797, but this was not significant. Therefore, in a statistical model with traditional prognostic factors, grade does not add prognostic accuracy.

Can histologic grade substitute for a factor that is becoming difficult to evaluate (e.g., lymph node status). To answer this question, we created a logistic regression model in which grade was the predictor and lymph node metastasis (detected vs. not detected) was the outcome. This addressed the issue of how well grade can take the place of lymph node status as a prognostic factor (in other words, to what extent does their prognostic information overlap?). If their predictions completely overlap, then the observed Az would be 1.0; if there was no overlap, then the observed Az would be .5. Again using the SEER data set, we found an Az of .589, which indicated that there was very little predictive overlap. Therefore, grade is not an effective surrogate for nodal status.

If grade is to be a useful prognostic factor in the future it must improve predictive accuracy for women with small tumors and few involved lymph nodes when used in predictive models that include the new molecular genetic prognostic factors. The data set from Duke University, kindly provided by Dr. Jeffrey Marks, includes patients with early stage breast carcinoma. These data were described in a previous arti-

cle.[10] Briefly, all patients were pathologic TNM Stage I or early Stage II. Early Stage II included all TNM Stage II patients except those with five or more positive lymph nodes. The variables were age, race, tumor size, positive lymph nodes, TNM lymph node status, nuclear grade, histologic grade, p53, c-*erb* B-2 (HER-2/*neu*), estrogen receptor status (ER) and progesterone receptor status (PR), vascular invasion, adjuvant therapy (tamoxifen, chemotherapy), and radiation therapy. Patients who underwent a lumpectomy received radiation therapy. Patients who underwent a modified radical mastectomy did not receive radiation therapy. There were 229 cases, 226 of which had complete data for all variables except ER and PR status. Because many individual patient ER and PR values were missing, both variables were removed from the data set. The 5-year survival rate was 70%. The logistic regression statistical method was used to create the models and a prediction endpoint of 5-year overall survival.

Neither histologic grade nor nuclear grade added any predictive power to the new molecular genetic prognostic factors in the logistic regression model. The predictive accuracy for all factors excluding histologic and nuclear grade was .733; when histologic grade was added the Az was .738 (not significant), when nuclear grade was added the Az was .736 (not significant), and when both were added the Az was .740 (not significant).

Overall, the accuracy of the Duke University logistic regression models was lower than the SEER logistic regression models because outcome prediction for early stage breast carcinoma was more difficult than outcome prediction for early and late stage breast carcinoma. In the Duke data set, the TNM staging system performed at chance level when predicting the outcome of women with early stage breast carcinoma, the Az was .567.[11]

Histologic grade alone has modest prognostic value. However, grade does not significantly increase the predictive accuracy of computer-based prognostic systems, either in data sets that represent all stages of breast carcinoma and contain traditional predictive factors or in data sets that represent early stage breast carcinoma and contain the new molecular genetic prognostic factors.

REFERENCES

1. Roberti NE. The role of histologic grading in the prognosis of patients with carcinoma of the breast: is this a neglected opportunity? *Cancer* 1997;80:1708–66.
2. Burke HB. Integrating multiple clinical tests to increase predictive accuracy. In: Hanausek M, Walaszek, Z, editors. Methods in molecular biology: tumor marker protocols. Mahwah, NJ: Lawrence Erlbaum Associates, 1997. In press.
3. Hays WL. Statistics. 5th edition. Fort Worth, TX: Harcourt Brace, 1991:335–8.
4. Burke HB. Evaluating artificial neural networks for medical applications. Proceedings of the 1997 International Congress of Neural Networks, 1997.
5. Halsted WS. The results of radical operations for the cure of carcinoma of the breast. *Ann Surg* 1907;46:1–19.
6. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors. Manual for staging of cancer. 4th edition. Philadelphia: J.B. Lippincott, 1992.
7. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–5.
8. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr., et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79:868–73.
9. Shambaugh EM, Ries LG, Young JL Jr. Extent of disease: new 4-digit schemes, codes and coding instructions. Bethesda, MD: National Institutes of Health, National Cancer Institute, Biometry Branch, 1984.
10. Marks JR, Wu K, Berry D, Bandarenko N, Kerns BJ, Iglehart JD. Overexpression of p53 and HER-2/neu proteins as prognostic markers in early stage breast cancer. *Ann Surg* 1994;219:332–41.
11. Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. *Cancer.* In press.

Specimen Banks for Cancer Prognostic Factor Research

Harry B. Burke, MD, PhD, Donald E. Henson, MD

• **Prognostic factors are necessary for determining whether a patient will require therapy, for selecting the optimal therapy, and for evaluating the effectiveness of the therapy chosen. Research in prognostic factors has been hampered by long waiting times and a paucity of outcomes. Specimen banks can solve these problems, but their implementation and use give rise to many important and complex issues. This paper presents an overview of some of the issues related to the use of specimen banks in prognostic factor research.**

(*Arch Pathol Lab Med.* 1998;122:871–874)

Prognostic factors are important for assessing the natural history of cancer, for selecting the optimal therapy, and for evaluating the effectiveness of treatment.[1] Two issues are central to the evaluation of prognostic factors. The first concerns the time from diagnosis to the analysis of outcomes (eg, mortality). The longer this interval is, the longer the prediction time interval becomes. To provide, for example, 10-year survival predictions, a patient population must be followed for 10 years. The 10-year information is used to assess the predictive accuracy of a prognostic factor and to provide 10-year outcome predictions to future patients. The second issue is the accrual of a sufficient number of outcomes so that the assessment of the factor is statistically reliable. Reliable means that a similar result would be observed if the analysis were repeated.

A human specimen bank that contains abnormal and normal tissue, white cells, serum, and plasma facilitates prognostic factor research because it eliminates the waiting time problem and the outcome accrual problem by collecting specimens from a defined patient population and following the patient population for a sufficient number of years. When a new putative prognostic factor is discovered, the stored material can be used to assess its predictive power.

Because specimen banks have all the difficulties of traditional data analysis as well as new difficulties, the development of a well-designed and useful specimen bank presents formidable challenges, especially for prognostic factor research. Furthermore, because specimen banks wait in silence for future use, an initial error may not become apparent for many years.

Although specimen banks will prove extremely useful, they do not solve the problem of how to validate a putative prognostic factor for a new therapy, that is, for a therapy that is not represented in the specimen bank population, nor do they solve the problem of the absence of agreed upon methods for validating prognostic factors and a consequent inability to replicate results.

Specimen banks are most commonly created for cancers whose initial therapy is surgical resection of the tumor. Specimen banks often collect more than the primary tumor. At surgery, blood, adjacent "normal" tissue, and metastatic tissue may also be collected. Although difficult and sometimes not feasible, blood should be collected after surgical therapy at regular intervals over many years and, if possible, tissue should be collected at recurrence, including from metastases, to assess a prognostic factor's predictive ability over time.

Clinical follow-up information and status, which are critical data for prognostic factor research, should be collected regularly, and investigators who have used the specimen bank should regularly update their data sets. A mechanism should exist for tracking patients who change physicians, move, or who for any other reason are lost to follow-up. Specimen banks for relatively slow-growing tumors (eg, breast and prostate cancer) and for rare tumors should be maintained for at least 20 years and preferably longer. Computer-based databases should be created and maintained for the life of the specimen bank.

A specimen bank must provide investigators with the relevant information regarding its data and tissue so that investigators can decide whether the tissue is appropriate for their task. Because tissue is to be distributed to many investigators for different purposes, the specimen bank's collection and reporting methods are critical. The better the specimen collection and reporting methods are, the more effectively the specimen bank can be used.

What follows is an overview of some of the important issues related to the use of specimen banks for prognostic factor research. Many of these issues are difficult to deal with and expensive to solve, and some may be insoluble, but they must be explicitly recognized by those creating and maintaining specimen banks.

MULTI-INSTITUTION COLLECTION

It has been the common practice of cancer investigators to collect and store human tissue for their own research.

Accepted for publication June 4, 1998.

From Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla (Dr Burke), and the Early Detection Branch, Division of Cancer Prevention, National Cancer Institute, Rockville (Dr Henson).

Reprints: Harry B. Burke, MD, PhD, Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, NY 10595.

The collecting investigator usually knows the characteristics of the patient population (including demographic and clinical data, treatment, and outcome) and of the tissue acquisition, storage, and retrieval process, and therefore the strengths and limitations of the specimen bank. To hasten the evaluation of molecular genetic putative prognostic factors, granting agencies have recently been providing funds for the coordinated collection and storage of human tissue from large numbers of patients.[2,3] For breast cancer alone there are currently at least 57 specimen banks.[4] Because of the number of patients required, these specimen acquisition efforts usually involve multiple investigators and multiple institutions.

An essential difference between multi-investigator, multi-institutional specimen banks and individual investigator specimen banks is that the investigators using multi-institutional specimen banks must rely on patient information and tissue supplied by the specimen bank. These investigators need detailed information regarding the patient selection criteria; demographic and clinical variable definitions; data acquisition methods; data coding system; tissue acquisition technique; and tissue preparation, storage, and retrieval methodology. Effective use of the tissue depends critically on the ability of the specimen bank to acquire, organize, and disseminate this information accurately and in a timely manner and, where possible, to standardize the patient acquisition, tissue collection, and clinical reporting process across collecting institutions to reduce interinstitutional variance.

A central issue for multi-institutional specimen banks is uniformity of methods. Surgical technique, specimen handling, tissue preparation, and quality assurance should be as uniform as possible across physicians and institutions. Some degree of standardization can be achieved by standardized procedures and training programs. Most cancer prognostic factors, including new molecular-genetic factors, are relatively weak predictors.[5] Therefore, the less uniformity that exists among institutions, the greater the loss of prognostic power and the greater the likelihood is that a strong factor will become weak and that a factor that is important for a small number of patients will be lost. It is through the combination of relatively weak factors in prognostic models that accurate predictions become possible.[1]

For multi-institutional data, the agreement among institutions in terms of clinical and laboratory variables and tissue characteristics should be assessed and reported to investigators. For continuous variables (eg, tumor size), a measure of central tendency, such as the mean (if the variable is normally distributed), and its variance should be calculated and the statistical differences assessed. For categorical variables (eg, race), the χ^2 statistic can be calculated, and for ratings, Cohen's κ statistic[6] can be calculated. If there is high interinstitutional variance, an investigator may, depending on the research issue and study design, restrict the source of tissue to one institution or to certain variables.

PATIENT SELECTION

The representativeness of a population of cancer patients is of vital importance because it determines the generalizability of the research results. For this reason, patient representativeness information should be provided to investigators prior to the use of the tissue. An unbiased patient selection process is necessary for population representativeness, since a bias in patient selection may yield results that are not generalizable. Sometimes a bias is minor and can be ignored, sometimes it is major but can be dealt with in terms of a reasonable assumption, and sometimes it is major and cannot be dealt with. In the last case, the investigator may wish to explore other more representative specimen banks. For example, tissue could be collected from a special group of women, and the frequency of *BRCA1* mutations could be determined and related to the incidence of breast cancer in that special population. It cannot be assumed, however, that the same quantitative relationship will hold true for all populations of women.

Population representativeness depends on many factors. One factor is the recruitment of patients. For example, clinical trials rarely provide a representative population of cancer patients because their entry criteria usually exclude certain patient groups. Thus, clinical trial populations are usually a biased sample because their entry criteria operate as a selection bias mechanism. This mechanism can limit the generalizability of the prognostic factor results. If an investigator expects the prognostic factor results to apply only to the patients that met the clinical trial's entry criteria, that is, to be in a position to perform only conditional prediction tasks, then the analysis may proceed. If the investigator expects the study results to apply to all the patients, however, additional assessment should be performed to determine if this generalization is reasonable given the study being contemplated. For example, for a study performed in a single institution, the investigator can determine whether the institution's patient population is representative of a more general patient population by comparing the characteristics of the institution's patient population with those of the larger patient population.

Patient populations may be biased by the method used for identifying incident cases. For example, some collection methods miss, in a nonrandom manner, approximately 18% of incident cases.[7] Patient populations may be biased by the clinical setting in which the cases are detected. For example, it is known that there are differences in the TNM stage frequencies reported by different types of hospitals.[8] Patient populations may be biased by where the patients are treated. Oncology clinic populations may differ from hospital populations. For example, in situ cancers may be more common in oncology clinics than in hospitals.[7] Patient populations may be biased by not distinguishing between incident and prevalent cases because prevalence depends on survival.[7] Patient populations may be biased by the incomplete collection of representative patient data. For example, incompleteness is an issue in data sets that have not existed for at least 40 to 50 years because the sample is not representative of the full spectrum of prevalent cases.[9] Finally, definitions may change over time. For example, changes in the TNM variable definitions[10–14] make it difficult to compare outcomes in terms of extent of disease.[15]

VARIABLE DEFINITIONS AND CODING

It is a nontrivial task to standardize the definition and coding of the clinical variables essential for prognostic factor research. Tumor registrars, for example, continue to refine the definitions and coding of the commonly collected variables. Prior to any data or tissue collection, representatives from each collecting institution should agree on a list of variables to be collected and create explicit definitions and a coding system for each variable. As the

specimens are collected and stored, an active quality assurance process should be in place at each institution. The quality assurance process will determine whether the definitions and coding system are being followed, and if not, whether to increase institutional vigilance or change the definitions and coding so that they can be adhered to.

DATA ACQUISITION METHODS

There are 3 data acquisition problems. The least difficult is the veracity of the original data source and the data entry. These errors include either incorrect entry of patient data in the patient's medical record or an error by the data collector. Data entry errors can be minimized through proper quality control. Two more difficult problems are missing data, that is, unknown variable values, and patients lost to follow-up, a type of censoring.

If a data set contains nonrandomly missing variable data,[16] then the missing data mechanism should be explicitly considered.[17] For example, in the National Cancer Data Base data set for the year 1983 with follow-up until 1990, of the 19147 cases listed, only 10357 are TNM staged (H.B.B., unpublished data, 1997). An investigator using tissue from these cases would have to ascertain whether the unstaged cases are the "same" as the staged cases. The missing values for some variables should be reacquired, while those of other variables should be filled in, depending on the importance of the variable. Basic information that should be included in the data set, such as therapy, can be easily lost,[7] and that type of data should be recaptured. If missing data are to be filled in, Little and Rubin's[17] book should be consulted.

Censoring (eg, patients lost to follow-up) is of critical importance to investigators because the primary purpose of the specimen bank is to provide outcome information that is useful for prognostic factor research. The 8-year censoring rate in the 1983 National Cancer Data Base data set is more than 25% (unpublished data). Censored cases usually have a less favorable outcome than noncensored cases.[7,18] Consequently, investigators should be informed of the specimen bank's censoring rate. If it is high (eg, greater than the 5% observed in the Surveillance, Epidemiology, and End Results [SEER] Program[19]), measures should be taken to systematically recapture these patients or the censoring mechanism should be explicitly considered. In addition, all patient variables should be compared with existing national databases. For example, the SEER Program reports an 84% 5-year survival rate,[19] and the National Cancer Data Base reports a similar 80% 5-year survival rate for patients with breast cancer (unpublished data). A survival rate in this range should be observed in breast cancer specimen banks. Using the existence of outcome data as a criterion for tissue selection by an investigator would not be appropriate if the patients who were lost to follow-up were lost because of their disease. In other words, if patients were lost because they were too sick, then selecting only those patients who were followed could bias a prognostic factor analysis.

It should be noted that there can be problems with survival estimation if the ascertainment of vital status is not random.[20] Finally, using cancer-specific survival as an outcome may be problematic because it assumes (1) that physicians accurately code death certificates and (2) that every patient dies of a single identifiable disease that can be correctly ascertained without autopsy. These assumptions are doubtful in the real world of competing risks.[21]

TISSUE ACQUISITION, PREPARATION, STORAGE, AND RETRIEVAL

In addition to clinical data collection, there are important issues related to tissue acquisition, preparation, storage, and retrieval. The manner in which the tissue has been collected, classified, stored, and retrieved may limit the types of studies that can be conducted or the "yield" of the tissue, or may affect the results of the study.[22–24] For each case, a sufficient amount of tumor should be available to investigators. If appropriate and possible, matching nonneoplastic tissue should accompany the tumor specimen. Because cellular degradation begins soon after removal, specimens should be kept as cool as possible and processed rapidly.

The 2 most common types of tissue processing are freezing and formalin/paraffin. Each has advantages and disadvantages for molecular genetic analysis. Ideally, tissue should be snap-frozen in liquid nitrogen as soon as possible after excision and should be stored at −70°C.[25] Delays in freezing or inadequate freezing can result in artifactual genetic changes. Some analyses can only be performed on frozen tissue. For example, if the concordance between abnormalities detected by immunohistochemistry and cDNA is important, then, depending on the antigen, paraffin-embedded, formalin-fixed tissue may not be useful.[26] Frozen tissue allows the polymerase chain reaction to be carried out on long stretches of DNA (1000 base pairs), since the DNA remains intact. Unfortunately, frozen tissue is difficult to handle, expensive to store, hard to provide in small amounts, and difficult to distribute to multiple investigators.[27] Procedures for processing tissue for molecular pathology have been published.[28]

For routine pathologic examination, tissues are usually fixed in formalin, dehydrated, and embedded in paraffin. Formalin forms cross-links between the reactive amine groups on adjacent proteins and between DNA and proteins, which makes the DNA rigid. Because of its rigid structure, subsequent tissue processing may cause fragmentation of the DNA. With conventional 10% formalin fixation, numerous variables are involved, including the duration of fixation, days in fixation that may result in DNA fragmentation by nucleases, size of the tissue, fixation gradients, and pH. These variables are difficult to standardize across institutions.[29]

Polymerase chain reaction amplification studies may be performed on archival formalin-fixed tissue even after years of storage.[30–32] In contrast to the situation with frozen tissue,[33] the restriction fragments will be relatively small, usually 100 to 300 base pairs in length. Cross-linking of the DNA can interfere with hybridization.[22] However, in spite of these effects of fixation, experience indicates that archived tissue is an invaluable resource for research in molecular genetics. Investigators have reported successful amplification of DNA extracted from 40-year-old specimens.[34] Fixatives that contain heavy metals, such as zinc or mercury (Zenker's fixative), will destroy DNA and are not suitable for tissue that is destined for DNA analysis. There is evidence that fixation in nonbuffered formalin may also degrade DNA.[35] The labile nature of RNA makes it difficult to recover in formalin-fixed tissue. Therefore, investigators interested in mRNA should use specimen banks that have the relevant cDNA library or that contain frozen tissue.

Highly suitable for tissue banks, immunohistochemical

methods can be used to detect the presence of specific gene products in tissue sections. For some antigens that deteriorate, the proper storage of tissues for subsequent immunohistochemistry is critical. Inappropriate fixation and tissue processing can affect the results as much as the variation in the antibody.[29] Fixed tissue may be stored as unstained cut sections mounted on glass slides or in paraffin blocks. Fixed tissues cut into sections, mounted on slides, and stored unstained may not always be suitable for immunohistochemistry. It has been shown, for example, that immunostaining intensity for p53 and other antigens will decay over time if the sections are cut from paraffin-embedded tissue and stored unstained on glass slides.[36,37] On the other hand, antigens usually do not decay if tissues are maintained in paraffin blocks. There is no decline in p53 staining intensity in tissues stored in paraffin blocks for more than 13 years.[38]

COMMENT

Prognostic factors are necessary for determining whether a patient will require therapy, for selecting the optimal therapy, and for evaluating the effectiveness of the therapy. Research in prognostic factors has been hampered by long waiting times and a paucity of outcomes. Although the implementation and use of specimen banks give rise to many important and complex issues, including the possibility of population biases and problems related to specimen handling, storage, and retrieval, specimen banks, because they solve the problems of prediction time and reliability, are a major advance in the field of prognostic factor research.

This work was supported in part by a research grant from the US Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383). We thank William E. Grizzle, MD, PhD, and Sheila Taube, PhD, for their helpful comments.

References

1. Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors. *J Cell Biochem.* 1994;19S:278–282.
2. Clausen KP, Grizzle WE, LiVolsi VA, Newton WA, Aamodt R. The cooperative human tissue network. *Cancer.* 1989;63:1452–1455.
3. LiVolsi VA, Clausen KP, Grizzle W, Newton W, Pretlow TG, Aamodt R. The cooperative human tissue network: an update. *Cancer.* 1993;71:1391–1394.
4. Anton-Culver H, Bringman DA, Taylor TH, Kim S. *National Action Plan on Breast Cancer: National Biological-Resource Banks Working Group Final Report, May, 1997.* Irvine, Calif: Epidemiology Division, Department of Medicine, University of California, Irvine; 1997.
5. Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. *Cancer.* 1998;82:874–877.
6. Cohen J. A coefficient of agreement for nominal scales. *Educ Psyol Measurement.* 1960;20:37–46.
7. McClish DK, Penberthy L, Whittemore M, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol.* 1997;145:227–233.
8. Steele GD, Winchester DP, Menck HR, Murphy GP. *National Cancer Data Base Annual Review of Patient Care 1993.* Chicago, Ill: Commission on Cancer; 1993:12.
9. Capocaccia R, DeAngelis R. Estimating the completeness of prevalence based on cancer registry data. *Stat Med.* 1997;16:425–440.
10. *Manual for Staging of Cancer, 1977.* Chicago, Ill: American Joint Committee for Cancer Staging and End-Results Reporting; 1977.
11. Beahrs OH, Myers MH, eds. *Manual for Staging of Cancer.* 2nd ed. Philadelphia, Pa: JB Lippincott; 1983.
12. Beahrs OH, Henson DE, Hutter RVP, Myers MH, eds. *Manual for Staging of Cancer.* 3rd ed. Philadelphia, Pa: JB Lippincott; 1988.
13. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, eds. *Manual for Staging of Cancer.* 4th ed. Philadelphia, Pa: JB Lippincott; 1992.
14. Fleming ID, Cooper JS, Henson DE, eds. *AJCC Cancer Staging Manual.* 5th ed. Philadelphia, Pa: JB Lippincott-Raven; 1997.
15. Henson DE, Ries L, Shambaugh EM. Survival results depend on the staging system. *Semin Surg Oncol.* 1992;8:57–61.
16. Choi SC, Lu IL. Effect of non-random missing data mechanisms in clinical trials. *Stat Med.* 1995;14:2675–2684.
17. Little RJA, Rubin DB. *Statistical Analysis With Missing Data.* New York, NY: John Wiley & Sons; 1987.
18. Elfstron J, Stubberod A, Troeng T. Patients not included in medical audit have a worse outcome than those included. *Int J Quality Health Care.* 1996;8:153–157.
19. Ries LAG, Kosary CL, Hankey BF, Miller BA, Harras A, Edwards BK, eds. *SEER Cancer Statistics Review, 1973–1994.* Bethesda, Md: National Cancer Institute; 1997.
20. Hu P, Tsiatis AA. Estimating the survival distribution when ascertainment of vital status is subject to delay. *Biometrika.* 1996;83:371–380.
21. Ebrahimi N. The effects of misclassification of the actual cause of death in competing risks analysis. *Stat Med.* 1996;15:1557–1566.
22. Moerkerk PT, Kessles HJ, ten Kate J, de Goeji AF, Bosman FT. Southern and dot blot analysis of DNA from formalin-fixed, paraffin-embedded tissue samples from colonic carcinomas. *Virchows Arch B Pathol Mol Pathol.* 1990;58:351–355.
23. Floss RD, Guha-Thakurta N, Conran RM, Gutman P. Effects of fixative and fixation time on the extraction and polymerase chain reaction amplification of RNA from paraffin-embedded tissue: comparison of two housekeeping gene mRNA controls. *Diagn Mol Pathol.* 1994;3:148–155.
24. Arnold MM, Srivastava S, Fredenburgh J, Stockard CR, Myers RB, Grizzle WE. Effects of fixation and tissue processing on immunohistochemical demonstration of specific antigens. *Biotech Histochem.* 1996;71:224–230.
25. Farkas DH, Kaul KL, Wiedbrauk DL, Kiechle FL. Specimen collection and storage for diagnostic molecular pathology investigation. *Arch Pathol Lab Med.* 1996;120:591–596.
26. Sjogren S, Inganas M, Norberg T, et al: The p53 gene in breast cancer: prognostic value of complementary DNA sequencing versus immunohistochemistry. *J Natl Cancer Inst.* 1996;88:173–182.
27. Naber SP. Continuing role of a frozen-tissue bank in molecular pathology. *Diagn Mol Pathol.* 1996;5:253–259.
28. Kiechle FL, Chambers LM, Cox RS Jr, et al: *Reference Guide for Diagnostic Molecular Pathology/Flow Cytometry.* Fascicle VII. Northfield, Ill: College of American Pathologists; 1996.
29. Grizzle WE, Myers RB, Manne U, Stockard CR, Harkins LE, Srivastava S. Factors affecting immunohistochemical evaluation of biomarker expression in neoplasia. In: Hanausek M, Walaszek Z, eds. *Methods in Molecular Medicine: Tumor Marker Protocols.* Totowa, NJ: Humana Press; 1998.
30. Pavelic J, Gall-Troselj K, Bosnar MH, Kardum MM, Pavelic K. PCR amplification of DNA from archival specimens: a methodological approach. *Neoplasma.* 1996;43:75–81.
31. Shibata D. Extraction of DNA from paraffin-embedded tissue for analysis by polymerase chain reaction: new tricks from an old friend. *Hum Pathol.* 1994;25:561–563.
32. Mies C. Molecular biological analysis of paraffin-embedded tissues. *Hum Pathol.* 1994;25:555–560.
33. Farkas DH, Drevon AM, Kiechle FL, et al. Specimen stability for DNA-based diagnostic testing. *Diagn Mol Pathol.* 1996;5:227–235.
34. Shibata D, Martin WJ, Arhneim N. Analysis of DNA sequences in forty-year-old paraffin-embedded thin-tissue sections: a bridge between molecular biology and classical histology. *Cancer Res.* 1988;48:4564–4566.
35. Goelz SE, Hamilton SR, Vogelstein B. Purification of DNA from formaldehyde fixed and paraffin embedded human tissue. *Biochem Biophys Res Commun.* 1985;130:118–126.
36. Jacobs TW, Prioleau JE, Stillman IE, Schnitt SJ. Loss tumor marker immunostaining intensity on stored paraffin slides of breast cancer. *J Natl Cancer Inst.* 1996;88:1054–1059.
37. Shin HJC, Kalapurakal SK, Lee JJ, Ro JY, Hong WK, Lee JS. Comparison of p53 immunoreactivity in fresh-cut versus stored slide with and without microwave heating. *Mod Pathol.* 1997;10:224–230.
38. Manne U, Myers RB, Srivastava S, Grizzle WE. Stability of p53 and Bcl-2 antigens in paraffin blocks. *J Natl Cancer Inst.* 1997;89:585–586.

Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction

Harry B. Burke, M.D., Ph.D.[1]
Philip H. Goodman, M.D., M.S.[2]
David B. Rosen, Ph.D.[1]
Donald E. Henson, M.D.[3]
John N. Weinstein, M.D., Ph.D.[4]
Frank E. Harrell, Jr., Ph.D.[5]
Jeffrey R. Marks, Ph.D.[6]
David P. Winchester, M.D.[7]
David G. Bostwick, M.D.[8]

[1] Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, New York.

[2] Department of Medicine, University of Nevada School of Medicine, Reno, Nevada.

[3] Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, Maryland.

[4] Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland.

[5] Department of Health Evaluation Sciences, University of Virginia School of Medicine, Charlottesville, Virginia.

[6] Department of Surgery, Duke University, Durham, North Carolina.

[7] Department of Surgery, Evanston Hospital, Evanston, Illinois; Commission on Cancer, American College of Surgeons, Chicago, Illinois.

[8] Department of Pathology, Mayo Clinic and Mayo Foundation, Rochester, Minnesota.

BACKGROUND. The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system. Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified and new methods for integrating prognostic factors have been developed. This study compares the prediction accuracy of the TNM staging system with that of artificial neural network statistical models.

METHODS. For 5-year survival of patients with breast or colorectal carcinoma, the authors compared the TNM staging system's predictive accuracy with that of artificial neural networks (ANN). The area under the receiver operating characteristic curve, as applied to an independent validation data set, was the measure of accuracy.

RESULTS. For the American College of Surgeons' Patient Care Evaluation (PCE) data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of the 5-year survival of patients with breast carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.720; ANN, 0.770; $P < 0.001$). For the National Cancer Institute's Surveillance, Epidemiology, and End Results breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM, 0.692; ANN, 0.730; $P < 0.01$). For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of the 5-year survival of patients with colorectal carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.737; ANN, 0.815; $P < 0.001$). Adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network's predictions of breast carcinoma survival (0.784) and colorectal carcinoma survival (0.869).

CONCLUSIONS. Artificial neural networks are significantly more accurate than the TNM staging system when both use the TNM prognostic factors alone. New prognostic factors can be added to artificial neural networks to increase prognostic accuracy further. These results are robust across different data sets and cancer sites. ***Cancer* 1997; 79:857–62.** © *1997 American Cancer Society.*

KEYWORDS: TNM staging system, artificial neural networks, prognostic factors, breast carcinoma, colorectal carcinoma, survival, outcomes, decision-making, clinical trials, quality assurance.

Presented at the annual meeting of the American Joint Committee on Cancer, Scottsdale, Arizona, January 14, 1995.

Supported in part by research grants from the American Cancer Society (CCG-274), the National Cancer Institute (CA 11606-17), the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383), the Agency for Health Care Policy and Research (HS 06830), and the American Joint Committee on Cancer.

The authors thank John H. Hellier, M.S., for his assistance on this project.

Address for reprints: Harry B. Burke, M.D., Ph.D., Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, NY 10595.

Received May 3, 1996; revision received October 4, 1996; accepted October 15, 1996.

The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system.[1] Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified[2,3] and new methods for integrating prognostic factors have been developed.[3] These methods may be capable of (1) providing more accurate predictions than the TNM staging system, using the TNM variables alone (primary tumor size, regional lymph node involvement, and distant metastasis), and (2) further increasing prognostic accuracy by integrating new prognostic factors with the TNM variables. This study compares the cancer specific 5-year survival prediction accuracy for breast and colorectal carcinoma of the TNM staging system with that of artificial neural network statistical models.

METHODS

Data

We used the Commission on Cancer's breast and colorectal carcinoma Patient Care Evaluation (PCE) data sets and the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) breast carcinoma data set.

In October 1992, the American College of Surgeons (ACS) requested cancer information from ACS-accredited hospital tumor registries in the United States. Specifically, they requested the first 25 cases of first-diagnosis breast and colorectal carcinoma seen at each institution in 1983, as well as follow-up information, including deaths, through the date of the request. Variables from this data set used in the breast carcinoma analysis were age, race, payment method, menopausal status, family history, previous biopsy, other cancer, other breast carcinoma, nipple discharge, mammogram, where in the breast the carcinoma occurred, necrosis, histologic grade, estrogen receptor status, progesterone receptor status, number of lymph nodes positive, number of lymph nodes examined, presence or absence of distant metastasis, tumor size, tumor type (in situ, extension to chest wall, or inflammatory), treatment (surgery, chemotherapy, or radiation therapy), and patient outcome (alive or dead). All variables were binary except age, tumor size, number of positive lymph nodes, and number of lymph nodes examined. The PCE data set contained up to 8 years of follow-up information. The analysis end point was breast carcinoma specific 5-year survival. Cases with missing data and those censored before 5 years were excluded. The data set was randomly divided into a training set of 5169 cases, including training and stop-training subsets, and a validation set of 3102 cases.

Variables from the PCE data base used in the colorectal carcinoma analysis were age, race, gender, signs and symptoms (changes in bowel habits, obstruction, jaundice, malaise, occult blood, abdominal pain, pelvic pain, rectal bleeding, or others), diagnostic and extent-of-disease tests (endoscopy, radiography, barium enema, computed tomography scan, biopsy, carcinoembryonic antigen, X-ray, colonoscopy, flexible sigmoidoscopy, intravenous pyelography, liver function tests, biopsy, or other tests), primary site of tumor, level of tumor, histology, grade, number of lymph nodes examined, number of lymph nodes positive, distant metastases, and patient outcome (alive or dead). The end point was 5-year colorectal carcinoma specific survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 5007 training cases, including training and stop-training subsets, and a validation set of 3005 cases.

The National Cancer Institute's SEER breast carcinoma data set, for new cases collected from 1977–1982, with 10-year follow-up, was also analyzed. The extent-of-disease variables for the SEER data set were comparable to, but not always identical with, the TNM variables. The end point was breast carcinoma specific 10-year survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 3788 training cases, including training and stop-training subsets, and a validation set of 2999 cases.

Models

The TNM staging system used in this analysis was the pathologic system based on the American Joint Committee on Cancer's *Manual for Staging of Cancer*.[1] The TNM staging system's predicted survival for a patient in a particular stage is the average survival of patients in that stage.

In medical research, the most commonly used artificial neural networks (ANN) are multilayer perceptrons that use backpropagation training (Figure 1). Backpropagation consists of fitting the parameters (weights) of the model by a criterion function, usually squared error or maximum likelihood, using a gradient optimization method. In backpropagation artificial neural networks, the error (the difference between the predicted outcome and the true outcome) is propagated back from the output to the connection weights in order to adjust the weights in the direction of minimum error. (For a more detailed description of artificial neural networks, see Burke[4] and Cross.[5]) The artificial neural network employed in this research was composed of three interconnected layers of nodes: an input layer, with each input node corresponding to a patient variable; a hidden layer; and an output layer. All nodes after the input layer sum the inputs to them and use a transfer function (also known as an activa-

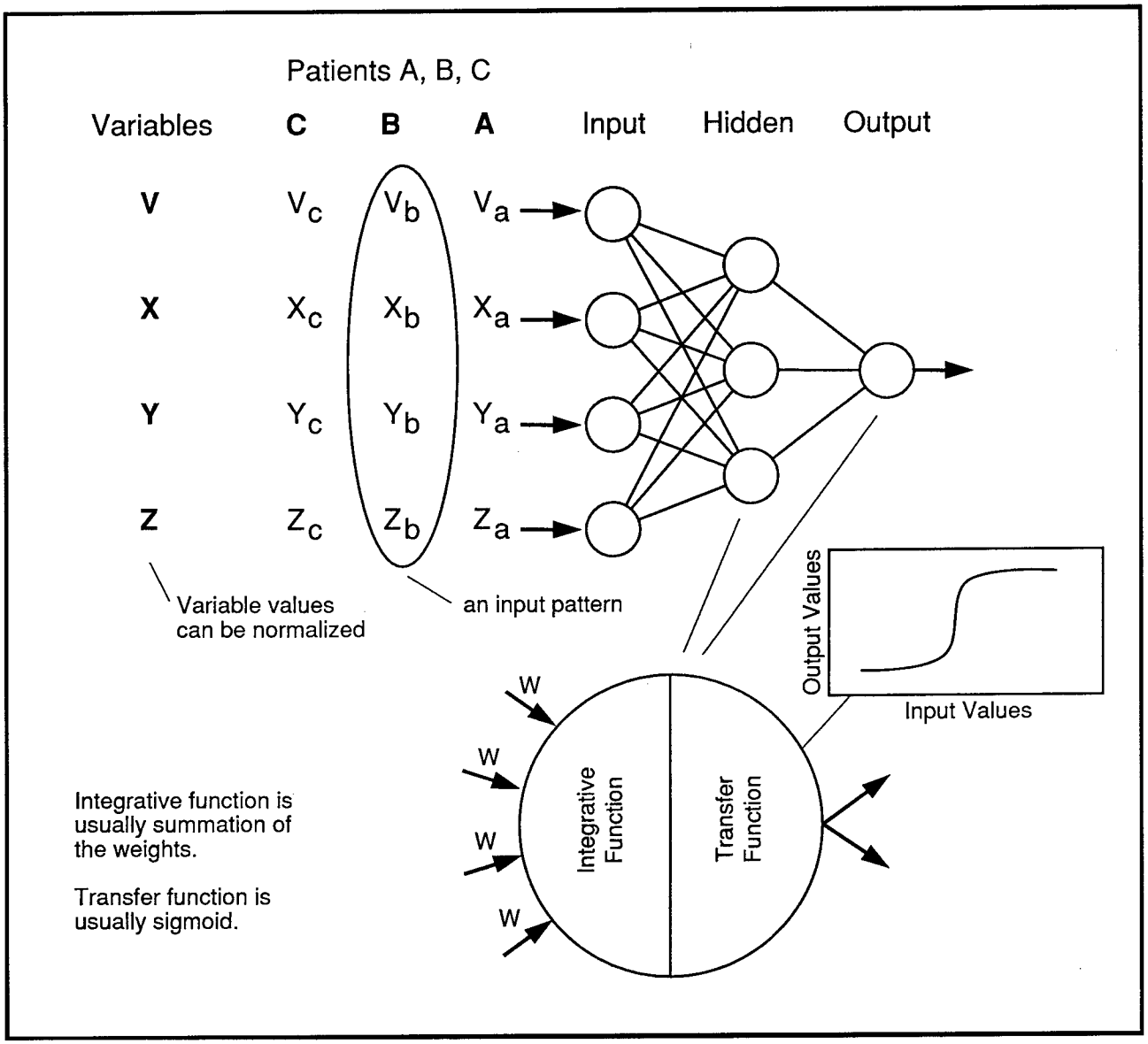


FIGURE 1. Patient A's variable values (Va–Za) are entered into the artificial neural network, followed by patient B, etc. Each variable's input value is multiplied by the weight between the input node for that variable and each hidden layer node it is connected to. All the weighted values going to a hidden layer node are summed at the hidden layer node and go through a sigmoid function before being transferred to the output node. All the weighted values coming into the output node are again summed and put through a sigmoid function. For each patient, the output is a probability from 0–1.0. In training the artificial neural network, the output of each patient is compared with each patient's true outcome. The weights are adjusted so that the next time the patient is presented to the network, the network output is closer to the true outcome.

tion function) to send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function, e.g., the logit. The connections between the nodes have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the adjacent layer nodes. These weights, along with the connections among the nodes, determine the output of the network.

The mathematical representation of an artificial neural network shown here is equivalent to the graphic model in Figure 1:

$$h_j = f(w^h_{j1}x_1 + w^h_{j2}x_2 + \cdots + w^h_{jn}x_n) \quad (1)$$

$$o_j = g(w^o_1 h_1 + w^o_2 h_2 + \cdots + w^o_n h_n) \quad (2)$$

where "h_j," in Equation 1 is the output of each of the hidden nodes j, f is a nonlinear transfer function, w^h is the weight from predictor i to hidden node j, and

x_i is an input variable. In Equation 2, o_j is the prediction of the network, g is a nonlinear transfer function, w^o is the weight to the output node, and h is the hidden node output. It should be noted that Equation 2, without the input from Equation 1, is equivalent to logistic regression, where g is the logistic function, w is the beta coefficient, and h is the x covariate.

Specifically, our artificial neural network (NevProp software implementation) used backpropagation training, the maximum likelihood criterion function, and a gradient descent optimization method. The number of input nodes correspond to the number of input variables, the number of hidden layer nodes ranged from three to five, and there was one output mode. Significant differences in the receiver operating characteristic areas between the TNM staging system and the artificial neural network were tested according to the method of Hanley and McNeil.[6] The training data set was divided into training and stop-training subsets. (Training was stopped when accuracy started to decline on the stop-training data subset.) All analyses employed the same training and validation data sets, and all results were based on the one-time use of the validation data sets.

Accuracy

There are three components to predictive accuracy: the amount and quality of the data, the predictive power of the prognostic factors, and the prognostic method's ability to capture the power of the prognostic factors. This study focused on the third component.

The measure of comparative accuracy is the trapezoidal approximation to the area under the receiver operating characteristic curve.[7] The area under this curve is a nonparametric measure of discrimination. While squared error summarizes how close each patient's prediction is to the true outcome, the receiver operating characteristic area measures the relative goodness of the set of predictions as a whole by comparing the predicted probability of each patient with that of all pairs of patients. This area is calculated using the predictive scores of each algorithm in order to compare their average accuracy in predicting outcome. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization, and its computation requires only that the algorithm produce an ordinally-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the algorithm will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the prognostic score is unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the score is from 0.5, the better, on average, the prediction model is at predicting which of the two patients will be alive.

TABLE 1
Comparison of the TNM Staging System with the Artificial Neural Network

Data sets	TNM staging system	Artificial neural network
PCE breast CA, TNM variables alone	0.720	0.770[a]
PCE breast CA, TNM and added variables	0.720	0.784[a]
SEER breast CA, TNM variables alone	0.692	0.730[b]
PCE colorectal CA, TNM variables alone	0.737	0.815[a]
PCE colorectal CA, TNM and added variables	0.737	0.869[a]

PCE: Patient Care Evaluation (Commission on Cancer); SEER: Surveillance, Epidemiology, and End Results (National Cancer Institute).
[a] $P < 0.001$.
[b] $P < 0.01$.

RESULTS

A comparison of the accuracy of the TNM staging system and the artificial neural network is shown in Table 1. For the PCE breast carcinoma data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of breast carcinoma specific 5-year survival were significantly more accurate than those of the TNM staging system (TNM 0.720; vs. ANN, 0.770, $P < 0.001$). Since the TNM staging system is, by definition, limited to the TNM variables, additional variables do not improve the TNM staging system's predictive accuracy. However, adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network (to 0.784).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across data sets. For the National Cancer Institute's 1977–1982 SEER breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM 0.692 vs. ANN 0.730, $P < 0.01$).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across cancer sites. For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of 5-year colorectal carcinoma specific survival were significantly more accurate than those of the TNM staging system (TNM 0.737 vs. ANN 0.815, $P < 0.001$). Adding commonly collected demographic and anatomic variables

to the TNM variables further increased the accuracy of the artificial neural network (0.869).

To clarify the clinical importance of the observed increases in accuracy, we changed the area under the curve (A_z) scale to a −1 to +1 scale, i.e., [2(A_z − 0.5)]. On this scale, 0 was chance and 1.0 was perfect prediction. By this measure, the TNM staging system's accuracy was 44% greater than chance for breast carcinoma specific 5-year survival predictions. Placing the TNM variables in the artificial neural network increased predictive accuracy to 54%, and adding variables that individually had little prognostic value to the artificial neural network further increased prognostic accuracy to 57% greater than chance prediction. Corresponding increases in predictive accuracy specific to colorectal carcinoma were as follows: 47% for the TNM staging system increased to 63% when the TNM variables were placed in the artificial neural network, and that increased to 74% when several commonly collected variables were added to the artificial neural network.

DISCUSSION

The TNM staging system is only moderately accurate in its breast and colorectal carcinoma specific 5-year survival predictions. The significant superiority in predictive accuracy that the artificial neural network showed when compared with the TNM staging system across data sets and cancer sites suggests that it is able to improve our ability to predict the survival of cancer patients. In addition, artificial neural networks can be expanded to include any number of prognostic factors. They can accommodate continuous variables and they can provide presurgery and postsurgery treatment predictions.

Artificial neural networks are a class of nonlinear regression and discrimination statistical methods. They are of proven value in many areas of medicine.[8–19] They do not require a priori information regarding the phenomenon, and they make no distributional assumptions. When the appropriate method is used to avoid overfitting (i.e., loss of generalization by fitting the patterns to the test data too precisely), artificial neural networks are usually at least as accurate as classical statistical models, and, depending on the complexity of the phenomena, they can be much more accurate. In predicting 5-year breast carcinoma specific survival, they have been shown to be more accurate than logistic regression, classification and regression trees (CART; pruned or shrunk), and principal components analysis.[20]

The improvement in prognostic ability made possible by artificial neural networks may be clinically important for therapy, clinical trials, patient information, and quality assurance. In decision-making regarding therapy, it may allow the efficient separation of patients with a poor prognosis (who require therapy) from patients with an excellent prognosis (who require little or no therapy), and it may predict who will respond to a particular therapy. In clinical trials, it may decrease interpatient variability. This would allow for the creation of more homogenous patient populations for clinical trials, resulting in smaller clinical trial patient populations, less expensive trials, and the ability to detect treatment effects that would be undetectable in more heterogeneous study populations. With regard to patient information, it may give patients a clearer understanding of the time course of their disease. Finally, for assessment and quality assurance, it may provide a better severity of illness adjustment.

REFERENCES

1. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors. American Joint Committee on Cancer. Manual for staging of cancer. 4th edition. Philadelphia: JB Lippincott, 1992.
2. Burke HB, Hutter RVP, Henson DE. Breast carcinoma. In: P Hermanek, MK Gospadoriwicz, DE Henson, RVP Hutter, LH Sobin, editors. UICC prognostic factors in cancer. Berlin: Springer-Verlag, 1995: 165–76.
3. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–5.
4. Burke HB. Artificial neural networks for cancer research: outcome prediction. *Semin Surg Oncol* 1994;10:1–7.
5. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet* 1995;346:1075–9.
6. Hanley JA, McNeil BJ. The meaning of the use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic. *J Math Psy* 1975;12;387–415.
8. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995;346:1135–8.
9. Dybowski R, Gant V. Artificial neural networks in pathology and medical laboratories. *Lancet* 1995;346:1203–7.
10. Westenskow DR, Orr JA, Simon FH. Intelligent alarms reduce anesthesiologist's response time to critical faults. *Anesthesiology* 1992;77:1074–9.
11. Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. *Radiology* 1993;189:555–8.
12. Leong PH, Jabri MA. MATIC: an intracardiac tachycardia classification system. *Pacing Clin Electrophysiol* 1992; 15:1317–31.
13. Gabor AJ, Seyal M. Automated interictal EEG spike detection using artificial neural networks. *Electroencephalogr Clin Neurophysiol* 1992;83:271–80.
14. Goldberg V, Manduca A, Ewert DL. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Med Phys* 1992; 19:1275–81.
15. O'Leary TJ, Mikel UV, Becker RL. Computer-assisted image interpretation: use of a neural network to differentiate tubular carcinoma from sclerosing adenosis. *Mod Pathol* 1992;5:402–5.
16. Dawson AE, Austin RE, Weinberg DS. Nuclear grading of breast carcinoma by image analysis. *J Clin Pathol* 1991; 95(Supp1):S29–S37.

17. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology* 1993;187:81–7.
18. Astin ML, Wilding P. Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clin Chem* 1992;38:34–8.
19. von Osdol W, Myers TG, Paull KD, Kohn KW, Weinstein JN. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J Natl Cancer Inst* 1994;86:1853–9.
20. Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems 7. Cambridge, MA: MIT Press, 1995: 1063–7.

Journal of Cellular Biochemistry, Supplement 19:278–282 (1994)

Increasing the Power of Surrogate Endpoint Biomarkers: The Aggregation of Predictive Factors

Harry B. Burke, MD, PhD

University of Nevada School of Medicine, Washoe Medical Center, Reno, NV 89520

Abstract A variable that predicts an outcome with sufficient accuracy is called a predictive factor. Predictive factors can be divided into three types based on the outcomes to be predicted and on the accuracy with which they can be predicted. These three types include risk factors, where the main outcome of interest is incidence and the predictive accuracy is less than 100%; diagnostic factors, where the main outcome of interest is also incidence but the predictive accuracy is almost 100%; and prognostic factors, where the main outcome of interest is death and the predictive accuracy is variable. Surrogate outcomes are predictive factors that are used for a purpose beyond the prediction of an outcome—surrogate outcomes are predictive factors that are substituted for the true outcome in order to determine the effectiveness of an intervention. Surrogate outcomes used in clinical trials are called intermediate endpoints and surrogate endpoints.

Predictive factors used as surrogate outcomes have a poor accuracy rate in predicting the true outcome; aggregating risk factors increases predictive accuracy. Artificial neural networks effectively combine predictive factors. Aggregating predictive factors increases the degree of linkage of the surrogate outcome to the true outcome. The resulting increase in predictive accuracy allows enrollment of people most likely to benefit from intervention. This increases the trial's efficiency, reducing the number of people required to assess a chemopreventive agent.

Key words: Chemoprevention, predictive factors, risk factors, surrogate endpoint biomarkers, surrogate outcomes

The risk, diagnostic, and prognostic cancer domains have their own literature and nomenclature. With the advent of molecular genetics, risk assessment, surrogate outcomes and chemoprevention, early detection, and new prognostic factors the divisions between these domains have blurred. This has led to some confusion as each domain's terminology is applied to the overlap between the domains. In this paper we propose to standardize several terms common to these three domains, and to demonstrate a method for combining predictive factors to increase prediction accuracy.

Address correspondence to Harry B. Burke, MD, PhD, University of Nevada School of Medicine, Washoe Medical Center, 77 Pringle Way, Reno, NV 89520.

PREDICTIVE FACTORS

For a predictive factor to be useful, its value must change in a predictable way when an intervention changes the outcome. An outcome is anything we are interested in predicting. In cancer, certain outcomes are important because they guide therapy. The three most common outcomes in cancer are incidence, recurrence, and death. Predictive factors can be outcome-specific; a variable may be a predictive factor for one outcome but not for another. Factors are level-of-analysis dependent; a particular factor exists only at a particular level of analysis. The terms "marker," "biomarker," "predictor," "prognosticator," and "indicator" have been used interchangeably with the term "factor," but they are not always synonymous. For example, most

predictive factors are markers of disease, but few markers of disease are predictive.

To determine whether a variable is a predictive factor, and if so, to determine its predictive accuracy, an outcome must be selected and the variable must be tested in a population. The population must be followed until a sufficient number of people in that population have achieved that outcome. If the variable predicts the outcome we are interested in with a sufficient accuracy, we call it a predictive factor. Sufficiency depends on the domain under study, and accuracy depends on the strength of the relationship between variable and outcome, the quality of data collection, and the ability of the predictive model to capture the relationship between variable and outcome. For prediction with a single factor, people with that factor are subsequently predicted to live as long as those with that factor in the original population. If the predicted outcome always occurs, we say that the predictive factor and the outcome are 100% linked, *i.e.*, that the factor has a 100% predictive accuracy.

RISK, DIAGNOSTIC, AND PROGNOSTIC FACTORS

Predictive factors can be divided into three types based on the outcomes to be predicted and the accuracy with which they can be predicted (Table I). These three types include risk factors, where the main outcome of interest is incidence and the predictive accuracy is less than 100%; diagnostic factors, where the main outcome of interest is also incidence but the predictive accuracy is almost 100%; and prognostic factors, where the main outcome of interest is death and the predictive accuracy is variable.

The term "risk" has several meanings. It can be used as a general term to denote the probability of the occurrence of an outcome, but it can also be used to denote a particular kind of predictive factor. This can be confusing, *e.g.*, the risk of disease given certain risk factors. In order to avoid this confusion, we will replace the general meaning of the term "risk" with the term "probability." Thus, we can speak of the probability of death given certain risk factors.

Risk factors are factors that either alone, or in combination with other factors, are less than 100% predictive of disease (incidence). They represent a propensity for disease at some future date. When a group of risk factors can be combined so that there is an almost 100% certainty of the disease at some future date, they become preclinical diagnostic factors (defined in the following paragraph) and are equivalent to screening for the disease. People at substantial risk require chemoprevention to prevent them from expressing the disease.

Diagnostic factors are factors that either alone, or in combination with other factors, are almost 100% predictive of disease. They can predict that disease exists at the time the factor is determined, or that it will exist at a usually unspecified time in the future. Two types of diagnoses—the existence of preclinical disease or clinical disease—can be made. In the preclinical disease state there is no evidence of invasive disease; in the clinical disease state there is evidence of invasive disease. Incidence occurs when invasive disease is detected by a diagnostic test. The preclinical disease state is almost always discovered by screening using biological and/or radiological tests, or by accident. The clinical disease state can be asymptomatic or symptomatic. Asymptomatic clinical disease is also almost always discovered by screening or accident, whereas the discovery of symptomatic disease is usually the result of a directed search. Early detection is the existence of one or more positive diagnostic factors in the preclinical or asymptomatic patient. Preclinical patients require chemoprotection, to protect them from expressing the disease.

Prognostic factors exist in patients with the disease and predict the outcome of interest. They are susceptible to change when therapy changes the future course of the disease. Prognostic factors are usually less than 100% predictive of the outcome, and are usually combined to increase their prognostic accuracy. They may be prognostic only for certain outcomes and certain times in the disease process, or they may be prognostic for all outcomes at any time in the course of the disease. For example, predicting the outcome recurrence may require different prognostic factors than predicting the outcome survival. The relationship between predictive factors, disease states, and interventions is shown in Figure 1.

Some diagnostic and prognostic factors are related; some diagnostic factors are prognostic

TABLE I. Three Types of Predictive Factors

PREDICTIVE FACTOR	ACCURACY	MAIN OUTCOME OF INTEREST
Risk	Much less than 100%	Incidence
Diagnostic	Close to 100%	Incidence
Prognostic	Variable	Death

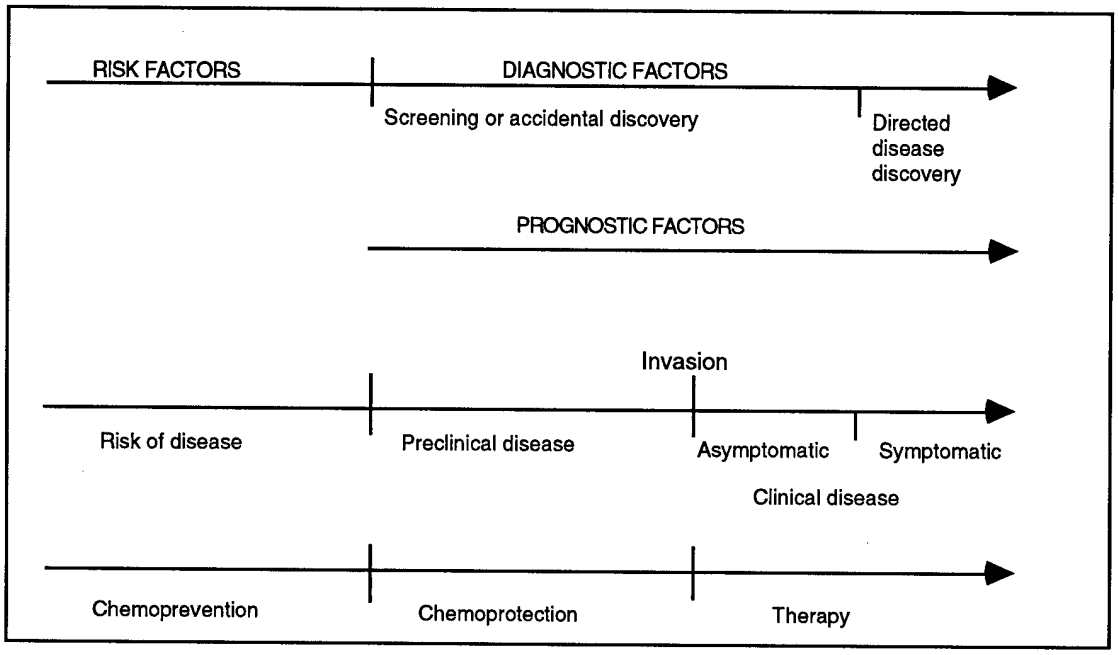


Fig. 1. Relationships between predictive factors, disease states, and interventions.

and some prognostic factors are diagnostic. Diagnostic and prognostic factors are distinguished by different purposes; diagnostic factors are used to predict the outcome existence-of-disease (incidence), and prognostic factors are used to predict outcomes related to the course of the disease. Thus, diagnostic factor analysis is similar, but not identical to, prognostic factor analysis.

Irrespective of the type of cancer, in order for an intervention to be maximally beneficial to the population at risk for the disease, to the patients with preclinical disease, and to the patients with clinical disease, three conditions must be met. First, since individual factor predictions are rarely close to 100% accurate, there must be a way to aggregate predictive factors. Second, the level at which the intervention (chemoprevention, chemoprotection, or therapy) will be instituted must be determined. Third, the intervention must be effective. Clearly the level at which the intervention will be instituted depends on accuracy of the aggregate prediction and the effectiveness of the intervention. Although at first it may seem that these three conditions do not apply to diagnosis, as we move more and more into the realm of early detection, with its greater diagnostic uncertainty, we will require

the aggregation of diagnostic factors, the setting of a level of diagnostic certainty, and the means to treat the disease we have discovered.

SURROGATE OUTCOMES

Because we do not know if an intervention is effective until the outcome of interest has occurred, and because many years can separate an intervention and the occurrence of the outcome, we would like to find something (a surrogate outcome) that changes soon after the intervention if the intervention is effective in changing the outcome. Surrogate outcomes are predictive factors that are used for a purpose beyond the prediction of an outcome; surrogate outcomes are predictive factors that are substituted for the true outcome for the purpose of determining the effectiveness of an intervention. Intermediate endpoint and surrogate endpoint refer to using surrogate endpoints in clinical trials. If there is a choice between these three terms, it is best to use the term "surrogate outcome."

Perfect Surrogate Outcomes

A perfect surrogate outcome is a factor that is 100% linked to the true outcome. We are almost always interested in a perfect surrogate outcome that precedes the true outcome. Since the perfect preceding surrogate outcome is totally linked to the true outcome, a change in the perfect preceding surrogate outcome due to an intervention will always signal a change in the true outcome. Having a perfect surrogate outcome means that we do not have to wait for the true outcome to occur to assess the effectiveness of the intervention on the true outcome. A perfect preceding surrogate outcome can be used as an index of the effectiveness of the intervention. All risk factors, diagnostic factors, and prognostic factors are potential surrogate outcomes, but few will meet the criteria for a perfect surrogate outcome.

For a factor to be a perfect surrogate outcome, two criteria must be met. First, it must be possible to discover the factor and determine its value prior to the occurrence of the true outcome. Second, there must be a 100% link between the factor and the true outcome. To determine the effect of an intervention using a surrogate outcome, one must determine the value of the surrogate outcome before and after the intervention. If the value of the surrogate outcome has changed in the desired direction, then we would expect the true outcome to change in the desired direction. A surrogate outcome may not be detected in everyone who has the disease. However, those predicted to experience the true outcome must actually do so.

Preclinical diagnostic factors can be used as a surrogate outcome for the true incidence because they accurately predict the true incidence. A clinical diagnostic factor is a lagging indicator of incidence, and therefore not a useful surrogate outcome.

Imperfect Surrogate Outcomes

Risk factors and prognostic factors are more problematic surrogate outcomes than diagnostic factors because they do not meet the second condition for a perfect surrogate outcome, namely, a tight linkage between the factor and the outcome. If the factor and the true outcome are not 100% linked, then a change in the surrogate outcome does not always reflect a change in the true outcome, and a lack of change in a surrogate outcome does not always mean that the true outcome has remained unchanged. Thus, the existence, magnitude, and direction of change in a true outcome are in doubt when the surrogate and true outcomes are not inextricably linked.

If we wish to use a factor as a surrogate outcome in spite of a weak relationship between the factor and the true outcome, there will be patients predicted to experience the true outcome who do not, and vice versa. This means that a change in the post-intervention value of a factor does not mean that we have necessarily changed the true outcome. To the degree that we can achieve close to 100% predictive accuracy, we will approach the ability to effectively use the factor as a surrogate outcome. If we allow a degree of error in the surrogate outcome's ability to predict the true outcome, we can use factors with less than 100% linkage as surrogate outcomes. In that case, we must be able to quantify the degree of linkage (the accuracy of the factor in predicting the true outcome) to determine if the prediction error is within the error tolerance. Error tolerance depends on the efficacy, side effects, and cost of the intervention, and the morbidity and mortality of the disease.

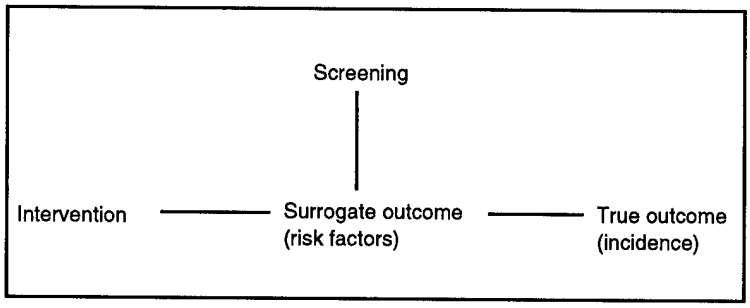


Fig. 2. Aspects of chemoprevention.

It is not clear that surrogate outcomes can reduce the time required for the initial investigation of chemopreventive agents, but it is almost certainly the case that the aggregation of risk factors can reduce the number of people that are required for the clinical trial, and that surrogate outcomes can be used in post-clinical trial chemoprevention efforts. Figure 2 shows the relationship between screening, intervention (chemoprevention), surrogate outcome (risk factor), and true outcome (incidence of disease). It is clear that the initial investigation must determine the accuracy of the screening test used to detect the surrogate outcome, the link between the surrogate outcome and the true outcome, and the efficacy of the intervention. Aggregation of risk factors into one surrogate outcome can reduce the size of the clinical trial. After the clinical trial, the surrogate outcome can allow physicians to determine whether the intervention is helping their patients, *i.e.*, they can use the surrogate outcome to monitor the efficacy of the intervention.

Cholesterol is an example of such monitoring. Cholesterol is a surrogate outcome for coronary artery disease. Patients are screened with a blood test; those with elevated cholesterol levels receive cholesterol lowering medications, and their cholesterol level is followed. Because there is a link between cholesterol levels and coronary artery disease, we believe that lowering the patient's cholesterol lowers the incidence of coronary disease. In cancer there are few risk factors strongly linked to the incidence of cancer; therefore, we must combine risk factors to increase the linkage between the surrogate outcome and the true outcome.

AGGREGATION OF PREDICTIVE FACTORS

Aggregation uses an analytic model to combine predictive factors to increase predictive accuracy. The analytic model used to combine prognostic factors for the American Joint Committee on Cancer's new prognostic system, a system that will replace the TNM staging system, is an artificial neural network.

The pTNM staging system is approximately 44% accurate in its predictions of five year survival for breast cancer. Placing the three pTNM variables in an artificial neural network increases their predictive accuracy to 52%. Combining other routinely collected variables with the pTNM variables in an artificial neural network increases predictive accuracy to 56%. Adding several of the new putative prognostic factors, *e.g.*, HER-2/*neu* and p53, to the artificial neural network further improves predictive accuracy to 70%.

Artificial neural networks are an effective method for combining predictive factors. In chemoprevention, the aggregation of predictive factors increases the degree of linkage of the surrogate outcome to the true outcome. This increased linkage increases the effectiveness of the chemopreventive agent by targeting people most likely to benefit from the intervention.

The ability to aggregate predictive factors can increase the accuracy of risk assessment, the accuracy of disease detection, and the ability to predict outcome for use in determination of therapy, patient information, quality assurance, and clinical trials.

1

Integrating Multiple Clinical Tests to Increase Predictive Power

Harry B. Burke

1. Introduction

Clinical tests provide information that can be used by statistical methods to make patient outcome predictions. Outcomes are risk of disease, existence of disease, and prognosis. In this chapter we define and describe predictive factors and clinical prediction and explain how combining predictive factors can increase predictive accuracy, describe the advantages and disadvantages of commonly used statistical methods, and recommend an approach to the reporting of predictive factor research.

2. Predictive Factors

A predictive factor predicts an outcome (risk of disease, existence of disease, or prognosis) by virtue of its relationship with the disease process that causes the outcome. For example, the prognostic factor mutant p53 is associated with breast cancer because of its role in the regulation of apoptosis. Such terms as marker, biomarker, predictor, prognosticator, indicator, surrogate factor, and intermediate biomarker have been used to identify variables that are connected to medical outcomes. Their meanings overlap, and their undifferentiated use can cause confusion. All predictive factors are markers of disease (i.e., they are in some way associated with the disease process), but not all markers of disease have sufficient predictive power to be called predictive factors. We use the term factor to identify markers of disease that either are, or have the potential to be, predictive for a given outcome in a specified model.

Determining whether a marker is a predictive factor requires that:

1. The variable is measured in a defined population;
2. The population is followed until enough outcomes have occurred (i.e., deaths); and
3. The relationship between the variable and the outcome is determined.

From: *Methods in Molecular Medicine, Vol. 14: Tumor Marker Protocols*
Edited by: M. Hanausek and Z. Walaszek

If the variable predicts the outcome with "sufficient" accuracy (where "sufficient" varies with the question being addressed) in a specified model, it is called a predictive factor. If the predicted outcome always occurs, we say that the predictive factor and the outcome are 100% linked, i.e., the factor has a 100% predictive accuracy *(1)*.

There are three types of predictive factors; risk, diagnostic, and prognostic *(1)*. They differ in their outcomes and predictive power. "Risk" is an ambiguous term. We use "risk" to refer to "risk of disease." "Risk," when used in the context of "risk of recurrence" or "risk of death," is called "probability," as in "probability of recurrence" and "probability of death." Risk factor; the main outcome of interest is incidence of disease. The factor, either alone or in combination with other factors, is much less than 100% predictive of the disease occurring by a specified time in the future. Risk can be viewed as a propensity for the disease. Diagnostic factor; the main outcome of interest is also incidence of disease. The factor, either alone or in combination with other factors, is close to 100% predictive of disease. Prognostic factor; the main outcome of interest is death. A factor is rarely a strong predictor in isolation from other prognostic factors. There is domain overlap in that risk factors can be prognostic, but they cannot be diagnostic, and diagnostic factors can be prognostic, but they cannot be risk factors.

There are three subtypes of predictive factors: natural history, therapy-dependent, and post-therapy *(1)*. Natural history predictive factors predict the future occurrence (risk), current existence (diagnosis), or course (prognostic) of a disease without an intervention. For risk and prognosis, natural history should the baseline against which all interventions are tested. Therapy-dependent predictive factors assume that there are effective therapies and predict whether the patient will respond to a particular intervention (for example, chemoprevention or chemotherapy). A natural history predictive factor may also be a therapy-dependent predictive factor. Post-therapy predictive factors require that patients respond to an intervention. They predict recurrence of the risk of disease or recurrence of the disease.

The predictive power of a factor depends on its intrinsic and extrinsic powers. The intrinsic predictive power of a factor is related to its "connectedness" to the disease process, i.e., its association to the disease process. The less connected the factor is, the less predictive it is. A direct connection means that the factor is an integral part of the disease process itself. An indirect connection means that it is not an integral part of the disease process but is related to the disease process, such as being a byproduct of it (i.e., a secondary infection). The extrinsic predictive power of the factor depends on the question being asked, i.e., the specific factor-outcome relationship being examined. For a specific disease process and outcome, the predictive accuracy of a factor depends on:

1. How closely connected the factor is to the disease process (individual factor power) and its relationship to the other factors (degree of predictive overlap);
2. How easy it is to collect and measure the factor; and
3. The degree to which the selected statistical method is able to capture the individual factor's predictive information and to integrate it with the information of other factors.

It is rarely the case that one factor is sufficiently predictive, i.e., that it is able to predict the outcome of interest with 100% accuracy. The usual strategy, when dealing with predictive factors, is to combine several in a predictive model. The most useful grouping of factors is one in which all of the factors are powerful and predictively orthogonal to each other, i.e., they index independent aspects of the disease process. If they represent aspects of the disease that are not independent of each other, then to the degree that their information overlaps is the degree to which one will not add predictive power. The statistical method employed must be able to capture the complexity of the disease process indexed by the predictive factors.

A predictive model for a specific outcome is the result of entering one or more predictive factors into a statistical method. The statistical method attempts to capture the relationship between the factors and the outcome. For example, the mathematical formula generated by the logistic regression statistical method relates the predictive factors (input variables), in terms of their β-coefficients, to a binary disease outcome (relapse, death, and so forth). It should be noted that the predictive power of a factor depends on the specific statistical method selected and on the other factors selected to be included in the model. The statistical model that results from the application of a statistical method, learning the relationship between the factors and the outcome, may or may not be the most efficient at capturing the predictive power of the factors.

Before discussing specific statistical methods, it is important to distinguish among significance, accuracy, and importance ***(2)***. Model significance asks if the observed predictions are really different from those produced by another model or from those resulting from chance.

Significance is not accuracy. Accuracy is the association between the model's predictions and the known outcomes in a test population. The importance of a model or a factor is determined by whether the model or factor possesses sufficient accuracy to be useful in answering a particular clinical question. Finally, the assessment of model or factor significance, accuracy, and importance must be based on test data set results, not on training data set results.

3. Advantages and Disadvantages of Statistical Methods

Many methods can be used to combine predictive factors. In cancer, they include bins, stages, and indexes; decision trees; and regression methods, including logistic, proportional hazards, and artificial neural networks.

Bins are the result of the mutually exclusive and exhaustive partitioning of discrete variables. Each combination of variable values is a bin, and all patients are placed in the bin corresponding to their variable value combination *(2)*. An example is the TNM classification of breast cancer *(3)*. Tumor size (Tis, T1, T2, T3, T4), number of positive regional lymph nodes (N0, N1, N2, N3), and existence of metastases (M0, M1) produce 40 bins *(2)*.

Each patient in a bin receives the same prediction; namely, the most frequent outcome. If there are enough patients in each bin, it can be shown that the most frequent outcome is the best predictor of the true outcome. In other words, no prediction model can be more accurate than a bin model if the variables are discrete and the population is large. Problems with bin models *(2)* include:

1. Continuous variables must be cut up into discrete variables. This almost always results in a loss of predictive information and therefore a loss of accuracy.
2. As the number of discrete variables increases, the number of bins increases exponentially. In order to maintain accuracy, there must be a corresponding exponential increase in the size of the patient population.
3. The proliferation of bins reduces the ability to understand the phenomena. Bin proliferation negates the main advantage of a bin model; namely, its ease of understanding and ease of use.

Bin models are rarely used in situations in which there are more than two or three predictive factors or where each factor possesses more than a few strata.

A partial solution to the problems of a bin model is a stage model *(2)*. A stage model is the grouping of bins into super-bins. The justification for the grouping is the assumption that the factors selected represent "stages" of the disease process. For example, in breast cancer, the TNM staging system combines 40 TNM classification bins into six super-bins (TNM stages) based on decreasing survival ("stages of survival").

A small set of stages has the potential to maintain explanatory simplicity and ease of use. Problems with stage models include:

1. The combining of bins into super-bins/stages can substantially reduce predictive accuracy.
2. Stage systems do not overcome the exponential increase in bins and patients associated with adding a variable to the analysis: They just delay the problem at a cost in predictive accuracy. If the stages are held constant when variables (and their associated bins) are added to the staging system, the potential improvement

in accuracy associated with the additional bins will be small to nonexistent. But, if the stages are expanded to accommodate additional bins, the system loses its ease of understanding and usefulness. Thus, attempts to improve predictive accuracy by adding variables to a bin/stage model are rarely successful.

3. The problems of cutting up continuous variables, with the resulting loss in predictive accuracy, remains.
4. Finally, if a single staging system is used for more than one cancer site, the staging rules may be more applicable to some sites than to other sites. The sites to which they do not apply will experience major losses in predictive accuracy.

Indexes associate numerical scores (usually based on a bounded, linear scale) with bins or groups of bins. Each score is associated with one of a small number of disease stages (usually a severity of illness system). Each patient receives the prediction of the stage in which their score places them. Indexes offer some flexibility in the grouping of bins, but at the cost of further degradation in predictive accuracy because additional information is lost. The simplest example of an index is the Apgar. An example in breast cancer is the Nottingham Index ***(4)***.

The accuracy of different stratifications of a predictive factor(s) can be compared. For a specific site (i.e., breast) and predictor(s) (tumor size <2, 2–5, >5) any bin or group of bins, or stage (bin or index) or group of stages, can be compared, in terms of a specific outcome, with another stratification (tumor size <1, 1–<2, 2–<3, 3–<4, 4–<5, 5–>5). This contrast can be over a single time interval without respect to events within the interval (i.e., logistic regression) or with respect to the events within the interval ***(5,6)***. For a single interval without respect to events within the interval, accuracy has been assessed by several discriminative association approaches, including Goodman and Kruskall's Gamma ***(7)***, Kendall's Tau ***(8)***, or the area under the receiver operating characteristic ***(9)***.

The usual descriptive approach for contrasting predictive factors across a series of event time intervals is the Kaplan-Meier product-limit method ***(5)*** (inferential methods that can accommodate continuous variables, and that usually assume proportional hazards, will be discussed later when regression methods are presented). A Kaplan-Meier plot should always include confidence intervals for each stratum (i.e., each step function). A significant difference within a Kaplan-Meier stratification (tumor size <2, 2–5, >5) is usually assessed by a log-rank test ***(10)***. It is important to note that there is currently no method for comparing the accuracy of two different Kaplan-Meier plots (i.e., two different stratifications of the same predictive factors). It is incorrect to use the p-value of the log-rank test to select one stratification over another, because the log-rank test only determines whether a stratification is likely to have occurred by chance. An extreme stratification may result in smaller p-values, but it may also reduce predictive accuracy.

Decision trees split predictive factors to maximize predictive power using a loss function, such as the log-likelihood and a greedy search algorithm. A well-known decision tree approach is the Classification and Regression Trees (CART) recursive partitioning method ***(11)***. Empirically, we have not found CART, either pruned or shrunk, to be the most accurate statistical method when compared to regression methods. Its problems include the selection of the correct loss function, difficulty dealing with continuous variables, and overfitting when searching for the best predictors when there are more than two or three splits.

Univariate regression methods are not appropriate for determining whether a variable is a predictive factor. Univariate methods should not be used, because new variables must be assessed in the context of the known factors, and because some variables are only predictive when they interact with another variable.

Logistic regression assess the cumulative probability of a binary event occurring by a specific time. It uses a maximum likelihood loss function and a greedy search technique. It is a very efficient method for binary outcome problems (i.e., recurrence and survival). Its limitation is that it usually spans a large time interval and does not distinguish when events occur within the time interval. This limitation can be overcome if several sub-time intervals are created within the overall time interval. Logistic regression models can be created for each sub-time interval. Censoring can be accommodated by removing cases that are censored within the time interval that censoring occurs.

Proportional hazards methods include the Cox ***(6)*** and less commonly the Weibull or exponential ***(12)***. Proportional hazards methods assume that the hazard of each patient is proportional to the hazards of all the other patients, and that a patient's hazard is related to that patient's relative risk. The Cox model does not create survival curves. For Cox-related survival curves, a baseline hazard must be introduced (for example, Breslow-Cox estimates) ***(13)***. Some researchers incorrectly believe that the Cox is the only regression method that can deal with censoring (*see* paragraph on logistic regression above). Because, in cancer, the proportional hazards' assumption may be violated, researchers who use the Cox model must demonstrate that the proportional hazards assumption holds for their population.

Artificial neural networks are a general regression method ***(14,15)***. They can perform almost any regression task. In addition, three-layer artificial neural networks automatically capture nonlinearity and complex interactions. They can handle censoring in the same way that multi-interval logistic regression handles censoring. Artificial neural networks are as transparent as the phenomena contained in the data. For simple phenomena, artificial neural networks are easily understood; for complex phenomena they are complex and less easily understood. Artificial neural networks are especially recommended in the domain of complex systems (e.g., the molecular-genetic domain of cancer).

4. Reporting Predictive Factor Research Results

There is a great deal of variation in the reporting of predictive factor results. This variability makes it difficult to understand and compare results. The following is a recommended approach to reporting the discovery of a new predictive factor or the validation of an existing factor.

For a defined subset of patients with the ___a___ disease, ___b___ is a ___c___ predictive factor for ___d___ when assayed ___e___ by ___f___, for the ___g___ on a test data set with ___h___ characteristics, the ___i___ is significant at the ___j___ level using the ___k___ statistical method, which also incorporates ___l___ predictive factors, for ___m___ therapy. Using the ___n___ method to assess its accuracy, the ___k___ statistical model is ___o___ accurate on the test data set.

"Defined" means specification of collection method, inclusion and exclusion criteria, and so forth.

a: Name of disease.
b: Name of the predictive factor.
c: Type and subtype of predictive factor (i.e., risk, diagnosis, prognosis; natural history, therapy-dependent, post-therapy).
d: Outcome (i.e., 5-yr breast cancer-specific survival).
e: Time of assay (i.e., at discovery, prior to therapy, after therapy).
f: Specific laboratory method (i.e., immunohistochemistry).
g: If stratified, the specific range/cut-point/and so forth of the prognostic factor. If the variable value is based on rater judgment, then Cohen's κ should be reported.
h: Relevant characteristics of the data set, including:
   1. Data set size,
   2. Number of events, and
   3. Whether the therapy was randomized.
i: The value and confidence interval.
j: For example, $p < 0.05$ for one test of the data. If multiple tests of the data are performed, an adjustment may be required.
k: Type of multivariate statistical method (i.e., logistic regression, Cox).
l: Other relevant prognostic factors, if they are included in the multivariate model.
m: Specific type of surgery, chemotherapy, radiation therapy.
n: Area under the receiver operating characteristic (Az) R^2, χ-square, etc.
o: Numerical value and its range of possible values (i.e., Az = 0.75, 0.50, –1.0).

References

1. Burke, H. B. (1994) Increasing the power of surrogate endpoint biomarkers: aggregation of predictive factors. *J. Cell. Biochem.* **19,** 278–282.
2. Burke, H. B. and Henson, D. H. (1993) Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* **72,** 3131–3135.

3. Beahrs, O. H., Henson, D. E., Hutter, R. V. P., and Kennedy, B. J. (1992) *Manual for Staging of Cancer,* 4th ed., Lippincott, Philadelphia, PA.
4. Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholson, R. I., and Griffiths, K. (1982) A prognostic index in primary breast cancer. *Br. J. Cancer* **45,** 361–366.
5. Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53,** 457–481.
6. Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. Royal Stat. Soc. B.,* pp. 187–220.
7. Goodman, L. A. and Kruskal, W. H. (1954) Measures of association for cross classifications. *J. Am. Stat. Assoc.* **49,** 732–764.
8. Kendall, M. G. (1962) *Rank Correlation Methods.* Hafner, New York.
9. Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psy.* **12,** 387–415.
10. Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50,** 163–170.
11. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth and Brooks, Pacific Grove, CA.
12. Evans, M., Hastings, N., and Peacock, B. (1993) *Statistical Distributions,* 2nd ed., Wiley, New York.
13. Breslow, N. E. (1974) Covariance analysis of censored survival data. *Biometrics* **30,** 80–99.
14. Burke, H. B. (1994) Artificial neural networks for cancer research: outcome prediction. *Sem. Surg. Onc.* **10,** 73–79.
15. Burke, H. B., Rosen, D. B., and Goodman, P. H. (1995) Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival, in *Advances in Neural Information Processing Systems,* vol. 7 (Tesauro, G., Touretzky, D. S., Leen, T. K., eds.), MIT Press, Cambridge, MA, pp. 1063–1067.

BIBLIOGRAPHY

Book Chapters

Burke HB, Hutter RVP, Henson DE. Breast Carcinoma. In P Hermanek, MK Gospadoriwicz, DE Henson, RVP Hutter, LH Sobin (eds), UICC Prognostic Factors in Cancer. Berlin: Springer-Verlag, 1995, 165-176.

Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In G. Tesauro, D.S. Touretzky, T.K. Leen (eds), Advances in Neural Information Processing Systems 7. Cambridge, MA; MIT Press, 1995, 1063-67.

Burke HB. Statistical analysis of complex systems in biomedicine. In D. Fisher and H. Lenz (eds), Learning from Data: Artificial Intelligence and Statistics V. New York: Springer-Verlag, 1996, 251-258.

Burke HB. The importance of artificial neural networks and biomedicine. In P.E. Keller, S. Hashem, L.J. Kangas, R.T. Kouzes (eds), Applications of Neural Networks in Environment, Energy, and Health. Singapore; World Scientific Publishing Co., 1996, 145-153.

Burke HB. Integrating multiple clinical tests to increase predictive accuracy. In M. Hanausek, Z. Walaszek (eds), Methods in Molecular Biology:, Vol. XX: Tumor Marker Protocols. Totowa, N.J., Humana Press Inc., 1998, Chapter 1, 3 - 10.

Peer Reviewed Journals

Burke HB. Artificial neural networks for cancer research: outcome prediction. Sem Surg Onc 1994;10:73-79.

Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors. J Cell Biochem 1994;19S:278-82.

Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:857-62.

Burke HB. Henson DE. Histologic grade as a prognostic factor in breast carcinoma. Cancer 1997;80:1703-1705.

Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. Cancer 1998;82:874-7.

Burke HB. Applying artificial neural networks to clinical medicine. J Clin Ligand 1998;21:200-201.

Burke HB, Henson DE. Specimen banks for prognostic factor research. Arch Path Lab Med, 1998;122:871-874.

Burke HB. Neural networks and their application to oncology and prognostic factors. CME J Gyn Onc, in press.

Proceedings

Burke HB, Goodman PH, Rosen DB. Artificial neural networks for outcome prediction in cancer. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Assoc. Inc., 1994; 53-56.

Burke HB, Goodman PH, Rosen DB. Neural networks significantly improve cancer staging accuracy. Proceedings of the 1994 IEEE Seventh Symposium on Computer-Based Medical Systems 1994; 200.

Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of statistical models and artificial neural networks in breast cancer. Fifth International Workshop on Artificial Intelligence and Statistics 1995; 87.

Burke HB, Goodman PH, Rosen DB. Applying artificial neural networks to medical knowledge domain. Proceedings of the International Symposium on Integrating Knowledge and Neural Heuristics 1995.

Burke HB, Goodman PH, Rosen DB. A computerized prediction system for cancer patient survival that uses an artificial neural network. Proceedings of the First World Congress on Computational Medicine and Public Health 1995.

Burke HB. The importance of artificial neural networks in biomedicine. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Associates Inc., 1995, 725-30.

Burke HB, Hoang A, Rosen DB. Survival function estimates in cancer using artificial neural networks. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Assoc. Inc. 1995, 742-7.

Rosen DB, Burke HB, Goodman PH. Improving prediction accuracy using a calibration postprocessor. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Assoc. Inc. 1996, 1215-20.

Burke HB. Measuring classification/prediction accuracy. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Associates Inc. 1996, 1213-14.

Rosen DB, Burke HB. Applying a gaussian-bernoulli mixture model network to binary and continuous missing data in medicine. Sixth International Workshop on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, Ft. Lauderdale, FL, 1997, 429-437.

Burke HB. Evaluating artificial neural networks for medical applications. Proceedings of the 1997 International Conference on Neural Networks, Houston, TX, 1997, 2492-2496.

Invited paper

Burke HB. Defining the computer-aided diagnostic device domain. Proceedings of the World Congress on Neural Networks. Hillsdale, NJ: Lawrence Erlbaum Assoc. Inc. 1996, 1233-35.

Grand Rounds

Burke HB, Goodman PH, Rosen DB. Using artificial neural networks in molecular biology. Division of Molecular Cytometry, University of California, San Francisco, San Francisco CA, February 17, 1994.

Burke HB. Computer-based outcome prediction, Department of Medicine, New York Medical College, Valhalla NY, October 9, 1994.

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

Burke HB. Survival prediction in cancer using artificial neural networks. Center for Devices and Radiological Health, Division of Electronics and Computer Science, U.S. Food and Drug Administration, Rockville MD, April 3, 1995.

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

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

Burke HB. Creating a Clinical Decision Support System. Departments of Biostatistics and Pathology, Mayo Clinic, Rochester MN, October 14, 1996.

Burke HB. Medical informatics and clinical decision support systems. Division of Medical Informatics, Stanford University School of Medicine, Stanford CA, November 7 - 8, 1996.

Burke HB. Creating a cancer clinical decision support system. Specialized Program of Research Excellence in Breast Cancer, Duke University Medical Center, February 25-26, 1997.

Burke HB. Creating a clinical decision support system for primary care physicians. Division of Primary Care, New York Medical College, Valhalla NY, April 23, 1997.

Burke HB. A clinical decision support system for prostate cancer. University of Michigan Cancer Center, Ann Arbor MI, June 9, 1997.

Burke HB. Cancer outcome prediction using artificial neural networks. Karmanos Cancer Institute, Wayne State University, February 19, 1998.

Presented papers

Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors for chemoprevention trials. Quantitative pathology in chemoprevention trials: standardization and quality control of surrogate endpoint biomarker assays for colon, breast, and prostate. Sponsored by the Chemoprevention Investigational Studies Branch of the National Cancer Institute. San Diego CA, February 9 - 13, 1994.

Burke HB, Henson DE. The clinical use of the new cancer outcome prediction system. American Cancer Society Medical Affairs Committee, Atlanta GA, March 4, 1994.

Burke HB, Goodman PH, Rosen DB. The future of the TNM staging system and cancer outcome prediction. The Lancet Conference: The Challenge of Breast Cancer, Brugge Belgium, April 21 - 22, 1994.

Burke HB, Goodman PH, Rosen DB. A computerized prediction system for cancer survival that uses artificial neural networks. First World Congress on Computational Medicine, Public Health and Biotechnology, Austin TX, April 24 - 28, 1994.

Burke HB. Predictive factors and surrogate outcomes in Cancer Screening and Surveillance. National Cancer Institute: International Workshop on Colorectal Cancer Screening and Surveillance, Bethesda MD, June 6 - 8, 1994.

Burke HB, Goodman PH, Rosen DB. Artificial neural network generation of survival curves. 1994 World Congress on Neural Networks and 1994 International Neural Network Society Annual Meeting, San Diego CA, June 4 - 9, 1994.

Burke HB, Goodman PH, Rosen DB. Neural networks significantly improve cancer staging accuracy. Seventh Annual IEEE Symposium on Computer-Based Medical Systems, Winston-Salem NC, June 10 - 11, 1994.

Burke HB, Goodman PH, Rosen DB. Computerized Outcome Prediction in Cancer. 10th Annual Meeting of the International Society for Technology Assessment in Health Care, Baltimore MD, June 19 - 22, 1994.

Burke HB. Understanding and using prognostic factors in cancer. 1994 College of American Pathologists Conference XXVI: Clinical Relevance of Prognostic Markers in Solid Tumors, Snowbird Utah, June 23 - 26, 1994.

Burke HB. Clinical significance of statistical analysis. 1994 College of American Pathologists Conference XXVI: Clinical Relevance of Prognostic Markers in Solid Tumors, Snowbird Utah, June 23 - 26, 1994.

Burke HB, Goodman PH, Rosen DB. Methods to integrate multiple markers. 1994 College of American Pathologists Conference XXVI: Clinical Relevance of Prognostic Markers in Solid Tumors, Snowbird Utah, June 23 - 26, 1994.

Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. Neural Information Processing Systems (NIPS 94): Medical Applications of Neural Networks (postconference workshop), Denver CO, November 28 - December 3, 1994.

Rosen DB, Burke HB, Goodman PH. Local methods in high dimension: are they surprisingly good but miscalibrated? Neural Information Processing Systems (NIPS 94): Local Methods in High Dimensions (postconference workshop), Denver CO, November 28 - December 3, 1994.

Burke HB. The science of medical prediction. National Coordinating Workshop on Prognostic Factors in Cancer, National Cancer Institute, Bethesda MD, January 4 - 5, 1995.

Burke HB, Rosen DB, Hoang A. Prototype cancer prognostic system with survival curves. National Coordinating Workshop on Prognostic Factors in Cancer, National Cancer I Institute, Bethesda MD, January 4 - 5, 1995.

Burke HB, Rosen DB, Goodman PH. Comparing the predictive accuracy of statistical models and artificial neural networks in breast cancer. Fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL, January 4 - 7, 1995.

Burke HB, Rosen DB, Hoang A. Cancer survival curve prediction using artificial neural networks. American Joint Committee on Cancer National Meeting, Scottsdale AZ, January 14, 1995.

Rosen DB, Burke HB, Goodman PH. Local learning methods in high dimension: Beating the bias-variance dilemma via recalibration. Workshop: Machines That Learn - Neural Networks for Computing, Snowbird Utah, 1995.

Burke HB. The importance of artificial neural networks in biomedicine. Workshop on Environmental and Energy Applications of Neural Networks, Battelle - Pacific Northwest Laboratories, Richland WA, March 30 -31, 1995.

Burke HB. The past, present, and future of cancer prognostic factor research. Prognostic Factors in Cancer, Cambridge Healthtech Institute, Arlington VA, June 7 - 8, 1995.

Burke HB. Computer-based prognostic system in cancer. Prognostic Factors in Cancer, Cambridge Healthtech Institute, Arlington VA, June 7 - 8, 1995.

Burke HB, Rosen DB. Missing data solutions using artificial neural networks. National Cancer Institute, Bethesda MD, June 21, 1995.

Burke HB. Artificial neural networks and biomedical research. World Congress on Neural Networks, Washington D.C., July 17 - 21, 1995.

Burke HB, Rosen DB, Hoang A. Survival function estimation in cancer using artificial neural networks. World Congress on Neural Networks, Washington D.C., July 17 - 21, 1995.

Burke HB. Outcome prediction in cancer. 1995 International Conference on Health Policy Research: Methodologic Issues in Health Services and Outcome Research, Harvard Medical School, Boston MA, December 2 - 3, 1995.

Burke HB. Defining the Computer-aided Diagnostic Device Domain. U.S. Food and Drug Administration Computer-aided Diagnostic Device Workshop, Rockville MD, January 26, 1996.

Burke HB. Computer-based Clinical Decision Support System in Oncology. The Eli Lilly Lecture, Conference on Prognostic Factors and Rational Treatment of Cancer, Yorkshire Cancer Organization, Leeds UK, July 3, 1996.

Burke HB. Issues in the regulation of medical software devices. Plenary session, FDA/NLM Software Policy Workshop, National Institutes of Health, Bethesda MD, September 3 - 4, 1996.

Rosen DB, Burke HB, Goodman PH. Improving prediction accuracy using a calibration postprocessor. World Congress on Neural Networks, San Diego CA, September 15 - 20, 1996.

Burke HB. Measuring classification/prediction accuracy. World Congress on Neural Networks, San Diego CA, September 15 - 20, 1996.

Burke HB. Defining the computer-aided diagnostic device domain. World Congress on Neural Networks, San Diego CA, September 15 - 20, 1996.

Rosen DB, Burke HB. Applying a gaussian-bernoulli mixture model network to binary and continuous missing data in medicine. Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL, January 4 -7, 1997..

Burke HB. Evaluating artificial neural networks for medical applications. 1997 International Conference on Neural Networks, Houston TX, June 9 - 12, 1997.

Burke HB. Clinical decision support systems. United States Veterans Administration, Washington D.C., October 1, 1997.

Burke HB. Visiting Professor, Uniformed Services University of the Health Sciences and Walter Reed Army Medical Center, Bethesda, MD, October 8, 1997.

Burke HB. A clinical decision support system for breast cancer. Department of Defense Breast Cancer Research Program: An Era of Hope. Washington D.C., October 31 - November 4, 1997.

Burke HB. The clinical use of neural networks. Finnish Breast Cancer Group, Tampere, Finland, November 14, 1997.

Burke HB. Artificial neural networks in medicine (tutorial), Serum Tumor Markers conference. Clinical Ligand Assay Society, Rye, NY., May 5, 1998.

Burke HB. Neural networks and their application to oncology and prognostic factors. Consensus Meeting on Prognostic Factors in Epithelial Ovarian Carcinoma, 11th Congress of the European Society of Gynecologic Oncology, May 6 - 7, 1999, Budapest, Hungary.

Posters

Burke HB, Goodman PH, Rosen DB. Outcome prediction in medicine. 1994 Annual Meeting of t he American Association for the Advancement of Science (AAAS), San Francisco CA, February 21, 1994.

Burke HB, Rosen DB. Using artificial neural networks to discover causes in complex systems: investigating the molecular biology of cancer. Second International Conference on Intelligent Systems for Molecular Biology, Stanford University, Stanford CA, August 14 -17, 1994.

Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. Neural Information Processing Systems (NIPS 94), Denver CO, November 28 - December 3, 1994.

Burke HB, Walavalkar J, Ammon R, Gupta K. Mammography in elderly nursing home residents. 51st Annual Scientific Meeting, Gerontological Society of America, Philadelphia PA, November 23, 1998.

Conference exhibits

Burke HB, Goodman PH, Rosen DB. Improving breast cancer survival predictions. 1993 American College of Surgeons Clinical Congress, San Francisco CA, October 11 - 14, 1993.

Burke HB, Rosen DB, Hoang A. Computer-based prediction for cancer survival. American Society of Clinical Oncology Annual Meeting, Los Angeles CA, May 20 - 23, 1995.

Conference positions

1994 *Program Committee and Co-chair*. Biomedical Applications of Neural Networks section, World Congress on Neural Networks and 1994 International Neural Network Society Annual Meeting, San Diego CA, June 4-9, 1994.

1994 *Chair*, Medical Information Section. Seventh Annual IEEE Symposium on Computer-Based Medical Systems, Winston-Salem NC, June 10 - 11, 1994.

1995 *Co-organizer*. National Coordinating Workshop on Prognostic Factors in Cancer, National Cancer Institute, Bethesda MD, January 4 - 5, 1995.

1995 *Scientific Advisor*. Cancer Prognostic Factors Conference, Cambridge Healthtech Institute, Arlington, VA, June 7 - 8, 1995.

1995 *Program Committee and Co-chair*. Biomedical Applications of Neural Networks section, World Congress on Neural Networks and 1995 International Neural Network Society Annual Meeting, Washington DC, July 17 - 21, 1995.

1995 *Workshop Chair*, Missing Data: Methods and Models. Neural Information Processing Systems 1995, Vail CO, December 1, 1995.

1996 *Program Committee,* World Congress on Neural Networks, *Co-chair,* Biomedical Applications Section, World Congress on Neural Networks and 1996 International Neural Network Society Annual Meeting, San Diego CA, September 15 - 20, 1996.

1996 *Workshop Chair*, Model Accuracy: Issues and Methods. Neural Information Processing Systems 1996, Vail CO, December 2, 1996.

1997 *Annual Symposium Reviewer*, American Medical Informatics Association.

1997 *Co-chair*, Special Session on Biomedical Applications, International Congress on Neural Networks, June 11, 1997, Houston Tx.

1998 *Reviewer*, 1998 American Medical Informatics Association Annual Symposium.

1999 *Program Committee*, 1999 International Joint Conference on Neural Networks, July 10 - 15, 1999, Washington, D.C.

1999 *Chair*, Artificial Neural Networks as a Statistical Method, A Special Session of the 1999 International Joint Conference on Neural Networks, July 10 - 15, 1999, Washington, D.C.
1999 *Chair*, Applications Session, 1999 International Joint Conference on Neural Networks, July 10 - 15, 1999, Washington, D.C.
1999 *Reviewer*, 1999 American Medical Informatics Association Annual Symposium.

OTHER INFORMATION

Editorial positions related to this research (Harry B. Burke, M.D., Ph.D.)

1994 - Present	*Reviewer*, Journal of the National Cancer Institute.
1994 - Present	*Reviewer*, Cancer.
1996 - Present	*Reviewer*, Journal of Clinical Outcomes Management.
1998 - Present	*Editorial Board*, Journal of the National Cancer Institute..
1998 - Present	*Reviewer*, British Journal of Cancer.
1998 - Present	*Reviewer*, Oncology.
1998 - Present	*Reviewer*, IEEE Transactions on Neural Networks.

List of personnel receiving pay from this effort.

Phillip H. Goodman, M.D.
Harry Burke, M.D., Ph.D.
David Rosen, Ph.D.
Albert Hoang, Ph.D.