AD_____

GRANT NUMBER DAMD17-94-J-4137

TITLE: Predicting Time-to-Relapse in Breast Cancer Using Neural Networks

PRINCIPAL INVESTIGATOR: Jonathan D. Buckley, Ph.D.

CONTRACTING ORGANIZATION: University of Southern California
Los Angeles, California 90033

REPORT DATE: December 1997

TYPE OF REPORT: Final

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

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

19980617 060

DTIC QUALITY INSPECTED 1

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 1997	3. REPORT TYPE AND DATES COVERED Final (15 Sep 94 - 14 Nov 97)

4. TITLE AND SUBTITLE Predicting Time-to-Relapse in Breast Cancer Using Neural Networks	5. FUNDING NUMBERS DAMD17-94-J-4137
6. AUTHOR(S) Buckley, Jonathan D., Ph.D.	

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Southern California Los Angeles, California 90033	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)*

We implemented neural network (NN) algorithms for analysis of censored-data in predicting time to relapse for breast cancer patients, including a generalization of the Buckley-James approach to censored linear regression, and the methods proposed by Faraggi and Simon and Liestol et al. The data set available included 236 women with node negative breast cancer treated with surgery only.

In Cox models HER-2/neu amplification, tumor size, treatment center, and age were univariately associated with outcome, but only HER-2/neu and treatment center significant in multivariate analyses. The recursive partitioning method selected HER-2/neu as the strongest predictor, and divided the non-amplified group by treatment center, and the amplified group by nuclear grade.

The NN initially appeared to be more predictive than any single covariate. To avoid overfitting problems (only 31 events in 133 cases with complete data) a 'bootstrap' analysis was conducted and proved no better than the Cox model in predicting outcome.

It is difficult to draw firm conclusions about the value of alternate methods because of the small dataset available. Further work is needed with other, large breast cancer databases to fully evaluate the potential of these alternative methods of analysis of prognostic factors.

14. SUBJECT TERMS Breast Cancer			15. NUMBER OF PAGES 25
			16. PRICE CODE

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

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).

_____ 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.

PI - Signature   2/27/9?   Date

TABLE OF CONTENTS

INTRODUCTION

Neural networks - a general description

NNs have been so-named because they mimic, in some respects, the structure and function of neurons in the brain. A NN consists of layers of nodes (analogous to neurons) linked by inter-connections (axons/dendrites), together with rules that specify how the output of each node is determined by input values from all nodes at the level below. A layered architecture of neurons in the brain can be used to provide progressively more abstract representation of input stimuli as the information is filtered through successive layers; NNs attempt to reproduce this effect, although most networks are limited in practice to three or four layers in total.

Output node(s)

Weights

Input nodes
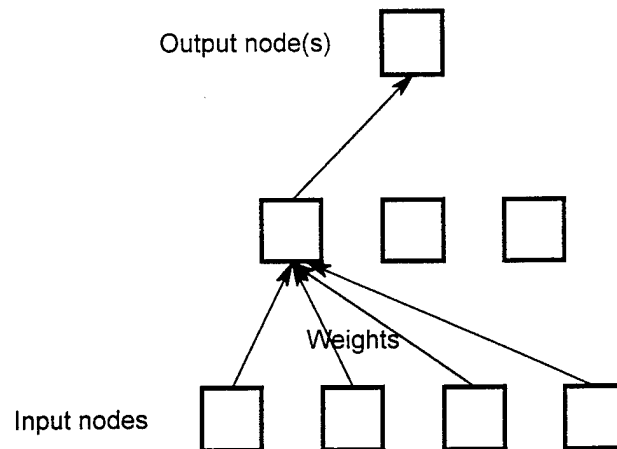
The lowest level of nodes in a NN is used to represent the **input** values, and the node or nodes in the highest level provide **output** from the NN. Since each node receives input from all nodes at the level below, generally combined as a **weighted** sum, the number of interconnections (and thus the number of weights) can be very large. Determining values for these weights *a priori* in order to obtained desired outputs for given inputs is clearly impractical for all but the most trivial networks. Useful NNs are made possible by the application of a learning algorithm that iteratively modifies the weights to minimize an "error function". The error function summarizes the differences between the actual output of the NN and the desired (or "true") output (Rumelhart, 1986).

The role of prognostic grouping and outcome prediction in clinical medicine

Establishing the prognosis for a patient may assist that patient in making **choices about treatment** and/or lifestyle changes. For breast cancer, determining the prognosis of a patient has become an essential first step to determining treatment: patients with a poor prognosis (at high risk of relapse or recurrence and/or with substantial residual disease) will generally be placed on the most intensive treatments while those with a good prognosis may be spared the acute toxicity and risks

of long term effects associated with aggressive treatment.

The traditional methods used to identify prognostic variables are logistic regression for categorical outcomes, such as death/non-response/remission, or Cox regression (Cox, 1972) for survival-type outcomes. These multivariate methods generally combine the explanatory variables in a single linear expression; more complex relationships between the explanatory and outcome variables can be modelled using stratification and interaction terms but incorporation of such terms tend to be limited. However, a complex "prognostic syndrome" that involves several variables in a non-linear fashion would almost certainly escape attention in a traditional Cox regression analysis.

In recent years, the largely clinical data that have been used to separate prognostic groups have come to be supplemented to an increasing degree by laboratory data. New analytic methods can provide information on such things as specific mutations, gene amplification, gross chromosomal abnormalities such as translocations, deletions, ploidy changes etc., presence or absence or cell surface markers including antigens and receptor proteins, and immunological parameters. Not unexpectedly, many of these biological characteristics correlate with outcome, but all too commonly new factors are reported without analysis of the extent to which they provide **independent** prognostic information, nor any guidance as to their use in conjunction with other factors in clinical decision making.

Use of neural networks to predict time to relapse in breast cancer

NNs have been used successfully to predict categorical clinical outcomes but there is no established method for dealing with potentially censored output values.

Ravdin et al (1992) published the first report of the use of NNs for clinical prediction with a survival-type outcome. This analysis attempted to relate six prognostic factors (tumor hormone status, DNA index, S-phase determination, tumor size, number of axillary nodes involved, and patient age) to time to relapse for women with node-positive breast cancer. A rather complex ad hoc method was used to adapt conventional NN programs to handle the censored data: (1) Selected input variables were log transformed, and normalized to lie within -1 to 1, (2) The database was split into a training set, evaluation set and validation set, (3) Time intervals (from the Kaplan-Meier curve) that corresponded to estimated rates of 0.90, 0.80, ... 0.10 were determined, (3) Each patient-record was split into m_i patient-time records (for patient i), where the time from study entry to time of analysis (the maximum follow-up) was T_i years and m_i of the time intervals come before T_i, (4) The NN was constructed with one output node (dead/alive) and a time variable (1, 2, up to T_i) as an input value. Patients that died before T_i were represented as dead in all patient-time records for intervals after their time of death, (5) To correct for bias due to non-uniform follow-up, the number of patient-time records corresponding to each interval was adjusted (by random elimination of records) to ensure that the ratio of records with "alive" status to those with "dead" status matched the observed Kaplan-Meier rates for the study group, (6) The output prediction was interpreted as a measure of relapse risk, and used to create risk subsets, (7) NN and Cox regression were compared in their ability to define groups with different Kaplan-Meier disease-free curves in an independent validation dataset, (8).

This approach was effective in defining prognostic groups: generally, the NN defined high and low risk subsets as efficiently as Cox regression and in some respects performed better. For instance, although having ten or more positive nodes was identified as a poor prognostic factor (32% relapse rate at 3 years), the NN placed only 54% of such patients in the high risk tertile; 40% were in the mid tertile and 6% were assigned to the lowest tertile. When the actual outcomes of the women with 10+ nodes were compared to all other women, within each predicted-risk tertile, relapse rates were very similar, indicating that the NN had correctly identified subgroups of apparent high risk (according to conventional methods) that belonged in lower risk groups.

Aims of this project

The aims of this project were:

1. To develop a program designed to apply neural network methods to the analysis of clinical data. Development will involve two stages:

   a. Software development of a neural network (NN) program, based on established methods.
   b. Extension of the NN program to handle censored data.

2. To integrate this program with existing software, in order to:

   a. Provide the neural network program access to a wide range of database management/data transformation functions.
   b. Provide a single package that will perform traditional analyses of clinical data (e.g. Cox regression) and neural network modelling.

3. To evaluate alternative methods for identification of prognostic factors. The methods included Cox regression (including recursive partitioning), censored linear regression, and four different neural network methods.

BODY

1. Development of a general NN program

The neural network program has been developed as a procedure (PROC NEURAL) within the statistical package Epilog Plus, in order to benefit from the broad range of data management features of this program and to facilitate comaprisons with more conventional methods. PROC NEURAL has been developed to have the following basic features (not specifically related to analysis of survival-type data):

Basic structure: Feed-forward neural network, with up to four layers, up to 50 input nodes and 50 output nodes. Logistic transfer function. Dynamic changes to network structure through switching on or off of nodes.

Training: Back-propagation of errors (calculated as sum of square of prediction error). Logicon

Projection offered as an option for weight initialization (see below). Weight updating following each record, or batched (e.g. after each 'run' through the training dataset). User-specifiable learning coefficient and momentum term, with the option to change these learning parameters after a preset number of runs - repeated such adjustments are allowed.

Epilog commands: Initial set-up determined by Epilog-style commands. Training pauses after a preset number of runs, or when the user 'breaks'. At this point, the commands can be modified, using the Epilog Plus command editor - if the changes do not alter the network structure, training continues from the previous point, otherwise the weights are re-initialized.

Weight analysis. In addition to the weight histogram, which gives an overall indication of weight sizes, the user can examine individual weights. By clicking on a node, the values of all interconnections into and out of that node can be shown. Combined with a scatterplot that shows the relationship between a nodes output and the networks output(s), this information can be used to investigate the contribution of a given node to overall performance.

Weight matrix - Save and recall. The network periodically writes the weights to a disk file. This can be at set intervals (number of runs), or when the RMS error on the test dataset begins to increase (suggesting that the network is beginning to overfit the training dataset). The weights can also be written to a disk file by selecting a menu choice. The weights can subsequently be read back, to pick up training or other activities involving the network from any point at which a weight 'dump' occurred.

Node enable/disable. The structure of a network can be readily altered (without complete respecification) by toggling a node on or off. When off, the network functions as if the node was not present. Nodes may be switched on or off at the command line, or interactively, through a menu choice.

Model predictions. A single key-stroke requests that the network's **predicted outputs** be written into the database. These predictions are then available to all other Epilog procedures, for example to determine means, medians, distributions, and to graphically display using PROC GRAPH.

Learning parameters. Several key learning parameters can be set at the start of a training session. Often the best values for these parameters early in training can be different from those needed later, when the network is getting close to a (local) minimum. To accommodate this, the user may pre-specify time points (in terms of run number) at which the parameter values should change. Alternatively, the user may make changes interactively, during a training session.

At any time the test and training dataset can be interchanged. This allows the user to examine network function with respect to test data as well as training data.

2. Windows 95 (32-bit) implementation

For applications such as neural networks which are very computationally intensive, and which can consume large amounts of memory to store the extensive weight matrix, the advantages of 32-bit

are obvious. Furthermore, the newest FORTRAN standard (FORTRAN 90) includes extensions to permit dynamic allocation of memory, which is also highly desirable for this application. For these reasons, the neural networks program was converted to Windows 95, using the zApp application framework as an intermediary API.

Fig 2 shows the Windows 95 display for a neural network training run.
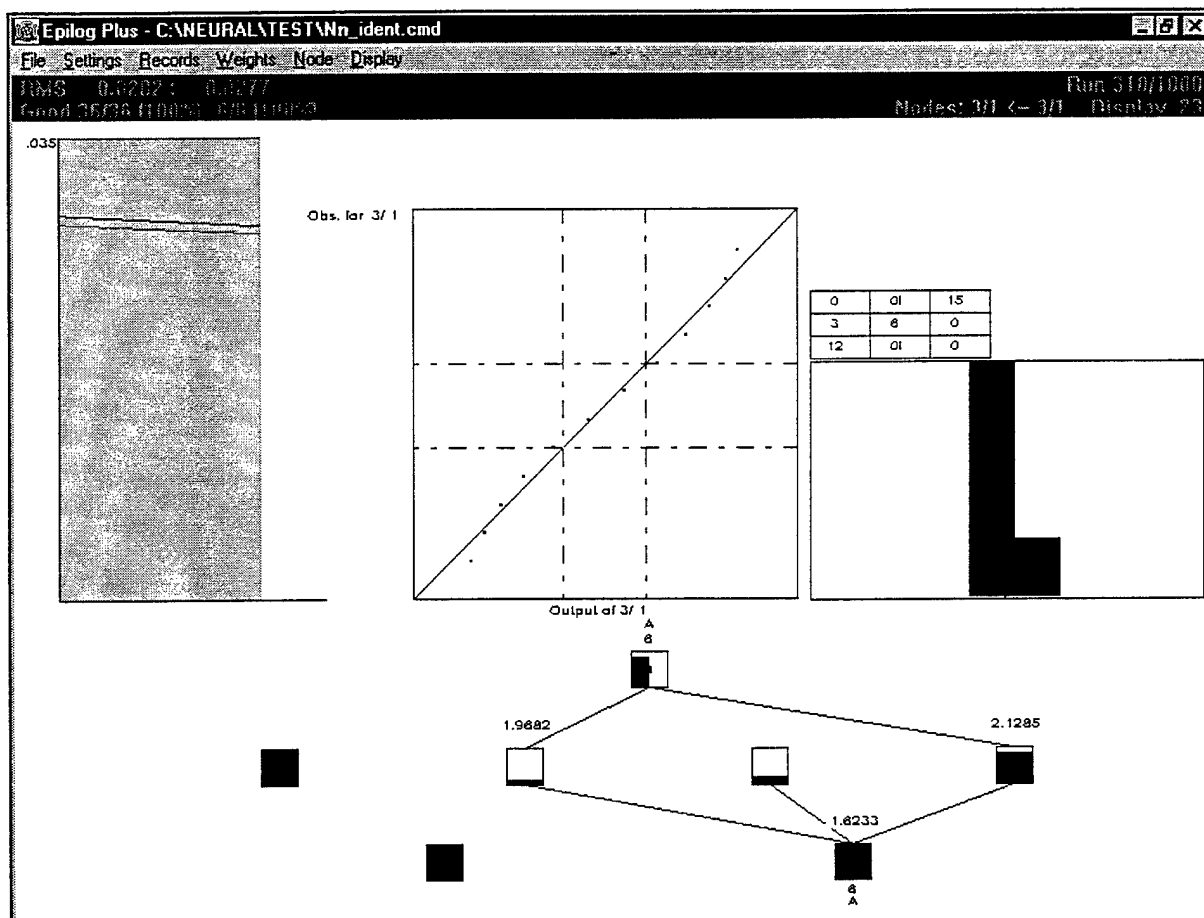
Figure 2. The training display of PROC NEURAL

The **upper panels** show (1) the name of the current Epilog command file, (2) the neural network menu choices, and (3) an information panel that shows the run number, the root mean square (RMS) prediction error (for both the training dataset and a test dataset if it is available) and the proportion of records for which the output prediction is classified as 'good' (within a specfied tolerance of the true value)..

Below this are **Plots** that show (1) RMS error by run number (for training and test data separately) and percent of predictions that are within a user-specified tolerance value of the true value (also for training and test datasets), by run number;(2) a scatter plot of predicted vs. actual value for a specified output node; (3) a table of predicted vs. actual,

5

based on the scatterplot and user-specified cutpoints;and (4) a **histogram of the NN weights**, which is useful for monitoring training progress.

The **scatterplot** will normally show the relationship between the predicted output for a node and the true value. When there is more than one node, any one of them can be selected for the scatterplot display by clicking on the node. As an alternative, the scatterplot can show the input or output of *any* node vs. the input or output of any other node or, for the top level (output) nodes, the true value. This provides avaluable tool for exploring network function.

The scatterplot can be replaced by a number of alternatives, depending on the type of data being modeled. One option is a **residual plot**, that shows a histogram of residual values (true-predicted) for a given output node. This distribution should become progrssively narrower as the network performance improves.

Another plot is the **Kaplan-Meier curve**, which is analogous to the residual plot - it displays in a Kaplan-Meier type survival curve the plot of residual survival times (true-predicted). This plot is intended for survival-type outputs.

The last type of plot is an **ROC curve**, intended for binary outcomes. The predicted value is continuous, and depending on where this value is cut, the prediction will have a range of values of sensitivity and specificity. The ROC curve is a widely used means of graphically displaying the relationship of sensitivity vs. specificity, and the area under the ROC curve is a measure of the performance of the predictor. When this plot is selected, the area is also calculated and shown.

A schematic of the network. Each node is shown, and for a selected training record, the values (and variable names) for each input and output node (for this record) are shown. The magnitude of the output from each node is indicated by a 'thermometer' that fills from 0% to 100% of the node interior. The record selected to be shown in this way can be fixed, or may change every time the display is refreshed. The display can be refreshed after every n runs (user specifiable). Selected interconnections between nodes are shown: only those with a weight (or optionally, a weight times node value) that exceed a threshold are shown, with negative and positive connections distinguished by color.

3. Extensions to the NN program to handle censored data

Undefined node method.

The simplest approach is to represent the outcome as a series of indicator variables corresponding to periods of follow-up. The interval in which a death occurred is represented by a 1, and all other intervals (output nodes) take value 0.

Death in Interval 3
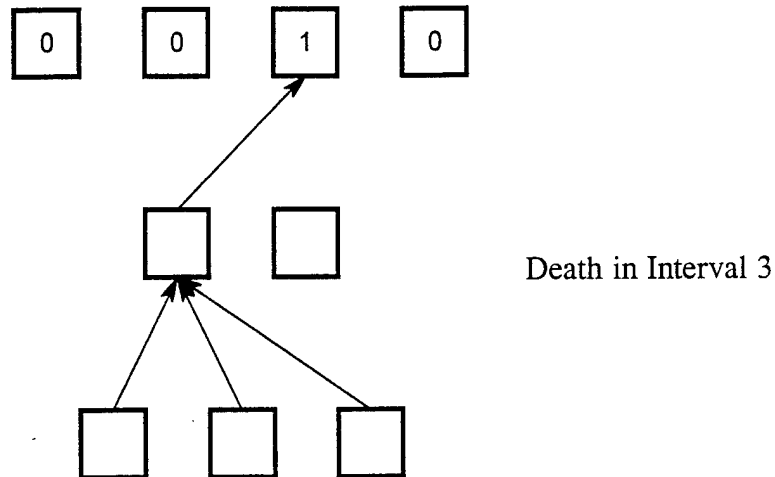
Intervals after a censoring event are considered to have an undefined value. In practice this means that this node does not contribute to the prediction error - essentially, it has no influence in the error that is used (through back-propagation) to adjust the weights. This method represents a relatively straight-forward modification to a 'standard' NN.

Censored in Interval 3

This node does not contibute to the error
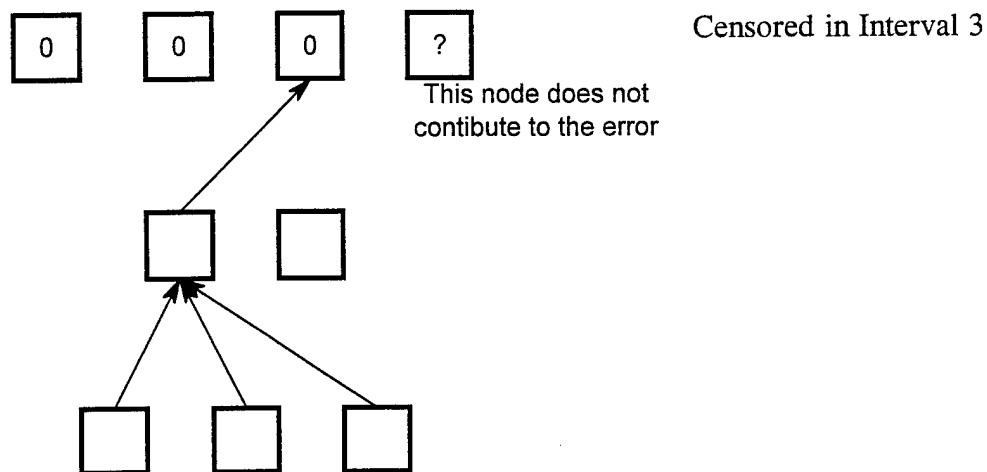
Buckley-James method.

In Buckley-James (expected value) method, NN predictions are compared to the actual values on each run and the differences (residuals) used to calculate a Kaplan-Meier-type curve. Based on the residual distribution (as reflected in the Kaplan-Meier curve) it is possible to estimate the expected survival for any person who was censored. The Buckley-James approach, as it was described for linear regression (Buckley and James, 1979) and as it is generalized to the NN setting, is to determine the expected survival for all censored individuals (based on the current weight matrix), and to substitute the expected value for the censored value when determining and

back-propagating the error.

Survival time

Censored observation is replaced
by its expected value, given the
current regression line

Regression line

Covariate
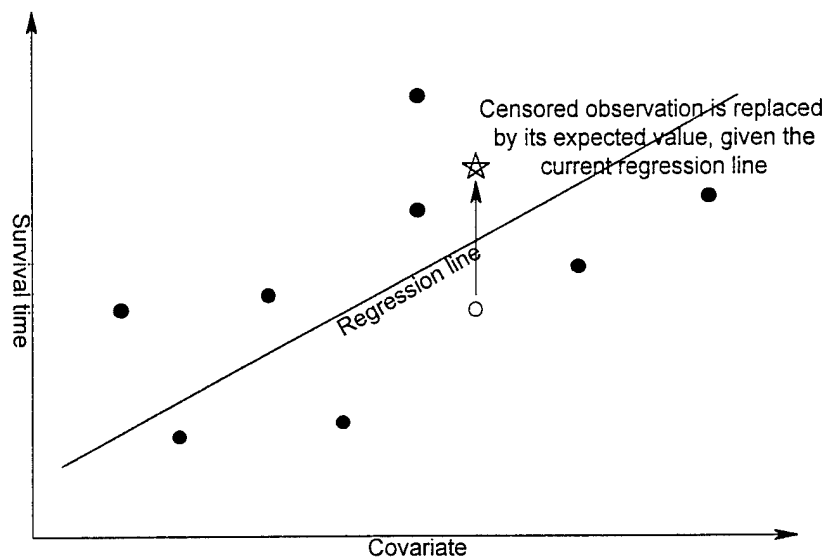
Modified error function.

An alternative is to modify the error function. For uncensored observations, the error function
remains quadratic:

Error
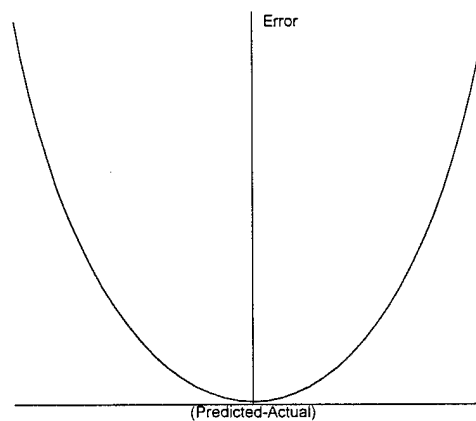
(Predicted-Actual)

However, for censored observations the error to be the square of the difference been actual and
predicted values, except when the prediction exceeds a censored actual value, when the error is
set to zero.

8

Error

Quadratic error when
prediction is less then
actual value

Zero error when prediction
exceeds actual value

(Predicted-Actual)

Method of Ravdin

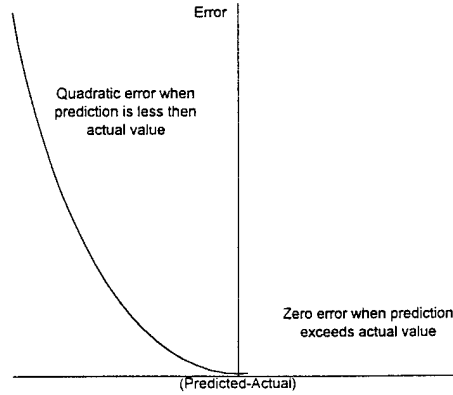This method does not require any specific programming, since it was developed to use existing NN software. However, the incorporation of our NN software within a more general purpose software package will make the Ravdin method much simpler to apply.

Liestol et al

This is a very promising method that creates a NN that has the property of being identical to Cox regression, when certain constraints are imposed. By choosing the appropriate error function and transfer function, and stipulating that certain weights be identical in value, it can be shown that the NN is the same as Cox regression. Clearly this is not very useful in itself, since there are better ways to implement Cox regression that through a NN, but the advantage is that the constraints can be selectively relaxed to produce a model that shares some properties with Cox regression, but also takes advantage of the power and flexibilty that is inherent in most NN models.

Farragi and Simon

Another generalization of Cox method has been proposed by Faraggi and Simon. They replace the usual linear function of covariates, in the Cox model:

$$\lambda(t,x_i) = \lambda(t) . \frac{e^{(\beta.x_i)}}{\prod_{R_i} e^{(\beta.x_i)}}$$

with a NN model (that is, a non-linear function, f(x)):

9

$$\lambda(t,x_i) = \lambda(t) \cdot \frac{e^{(f(x_i))}}{\prod_{R_i} e^{(f(x_i))}}$$

The numerical method for parameter (weight) estimation suggested by them is Newton-Raphson, although any of the estimation methods can be used.

4. Other additions relevant to this project

PROC COX and PROC CENREG

As part of the strategy of providing a powerful suite of routines, in one package, that could be used for clinical prediction using censored data, we have modified PROC COX (Cox regression) and PROC CENREG (censored linear regression) to generate predictions on a case-by-case basis and to write these back into the database.

PROC TREE

Recursive partitioning is an iterative process that repeatedly subdivides the dataset in order to achieve maximum separation of the subgroups. The steps are as follows:

(1) At each partitioning stage, the variable that provides the 'best' division of an existing partition is used to create a new partition. The criterion for deciding on the best division may be based on ratios of observed to expected events, on degree of separation of the survival curves, or on the logrank statistic (p-value). For continuous variables, finding the best separation involves examination of all possible cutpoints. For nominal variables, finding the best separation involves comparisons of all possible combinations of categories.

(2) Partitioning ceases when the criterion (for the best division) does not exceed a prespecified threshold value.

(3) Once partitioning is complete, the program can examine all pairwise combinations of partitions to determine if any are sufficiently similar (based on O/E, separation or logrank statistic) to combine. After any two partitions are combined, all pairs are again examined, until no further combinations are possible.

5. Evaluation of performance

As an inevitable consequence of developing alternative NN methods for handling censored data, and comparison of them with conventional approaches, we have given a good deal of thought to methods for evaluation and comparison. Simple scatterplots of observed and predicted (and correlation coefficients) are not useful, since the 'true' event time is not known for censored

individuals. Alternatives that have been used by others, and/or have been considered by us are:

Group comparison.

The most straight-forward approach, and one that is widely used, is to define prognostic groups based on the predicted value, and to examine the Kaplan-Meier curves for each group. The better method would generally be expected to give greater separation of the curves (based on observed/expected ratios) and higher chi-square test statistics for trend across categories. However, this method does not provide a direct test for the superiority of one method over another.

Cox goodness of fit.

A very useful approach is to fit a Cox model to the covariates of interest, and treat this as a baseline for comparison to other models which include predictors derived from alternative methods. The advantage of this is that a NN prediction, say, when treated as if it were just another covariate may be shown to significantly improve the model fit when compared to a simple (no interaction terms) Cox model; an example of this can be seen below. The disadvantage is that it evaluates the NN prediction within the context of a proportional hazards model, which may or may not be appropriate, and thus may unfairly weigh against the NN prediction.

ROC curves and the c statistic.

Receiver operating characteristic (ROC) curves are important tools for analyzing model performance when the outcome is dichotomous. A high neural network score may have excellent specificity but poor sensitivity for predicting clinical outcomes, whereas a low neural network score will be more sensitive but less specific. The ROC curve plots combinations of sensitivity and specificity for the entire range of model predictions, providing an overall view of performance (Hanley and McNeil, 1982) and the area under the ROC curve is a useful measure of prognostic accuracy. ROC curves have been used in several neural network studies (Ebell, 1993; Patil et al., 1993; Buchman et al., 1994).

In survival analysis the area under the ROC curve can be been used as a measure of prognostic accuracy at a specific time point - that is, after dichotomizing the output (Knaus et al., 1995). However a more powerful approach is to calculate the c statistic which can be considered a generalization of the area under the ROC curve. It is calculated by considering all possible pairings of patients and evaluating the concordance between model predictions and outcomes (Harrell et al., 1984). Model predictions are concordant if the patient predicted to live longer had a later event. However, if two patients both were censored then the concordance of model predictions cannot be determined. If only one patient is censored then concordance can still be determined, as long as the censored patient stayed in the study long enough. The c statistic derived from these comparisons has been used in evaluating neural networks for survival analysis (Faraggi and Simon, 1995).

11

6. Learning methods (parameter estimation)

These methods all require estimation of a large number of parameters in order to minimize the value of an error function. The most common approach is 'back-propagation' which is essentially a gradient descent approach. In practice, because of the large number of weights used in many NNs, convergence to an error minimum can be slow. Furthermore, this minimum may be a local rather than a global minimum. While we still feel that the back-propagation approach is extremely useful, the problems of long training time and local minima have led us to develop alternative training algorithms.

Logicon projection.

This is a patented algorithm developed by scientists at Logicon Inc, Los Angeles (Wilensky G and Manukian N, 1992). It requires that the user specify a 'prototype' individual to correspond with each hidden node. A method of N-dimensional projection is used to calculate initial weights into these hidden nodes so that that node fires maximally for the prototype individual. Starting the NN with such weights, instead of randomly assigned one, can reduce training time by one to two orders of magnitude. Even if no care is taken to select appropriate prototypes, and they are drawn at random from the training database, training times can be substantially reduced.

Numerical methods

Three numerical approaches for parameter estimation have been implemented:

> a. Gradient Method (back propagation)
> b. Newton Method
> c. Conjugate Gradients Method

Genetic algorithms.

A radically different approach to weight optimization may be used to try to avoid getting caught in a local minimum. With the so-called 'genetic algorithm' the weights are represented conceptually as genes on a chromosome. Instead of a single NN, a whole 'population' of networks with the same structure are created. The performance of each is evaluated, and the best (smallest error) are selected for 'mating', while the worst are removed (die).

The weight-chromosome for each offspring is derived from the parent chromosomes through a process analogous to meiotic recombination with or without point mutations. Through many generations, with only the fittest being allowed to pass on their weight-chromosomes to new individuals, network performance improves. As a result of the discontinuous nature of the recombination process the weight matrix makes jumps in the parameter space that potentially avoid the trap of a local minimum and hopefully allows for exploration of the entire space for a global minimum (Narayanan and Lucas, 1993).

7. Evaluation using simulated data

The most useful database for initial evaluation of network performance is one in which the relationship between the input and output variables in known.

The simulation database included 1000 training records and 1000 testing records with four binary input covariates (A, B, C and D):

Covariate	Prob(value=1)	Prob(value=0)
A	0.05	0.95
B	0.10	0.90
C	0.25	0.75
D	0.50	0.50

Cases were assigned to a Low, Intermediate or High risk group, based on their covariate values:

Covariate 'pattern'	Group	Distribution
C=1 or (B=1 and D=1)	Low risk	exp(0.005)
{all other combinations}	Intermediate risk	exp(0.01)
A=1 or, C=1 and (D=1 or B=1)	High risk	exp(0.02)

The censor time was drawn from a uniform(0,365) distribution.

Cox regression was used to fit the four covariates (A - D), the NN prediction (P), a model with A,B,C,D and P, to determine whether P contained useful prognostic information not provided by the covariates (as main effects). Note, these evaluations were based on the test group only.

Table 2. Cox regression goodness-of-fit chi-squares

	Model	χ^2	D.O.F	p-value
1	A,B,C,D	44.00	4	<0.0001
2	P	90.86	1	<0.00014
3	A,B,C,D, plus P	106.54	5	<0.0001
	vs. Model 1	62.54	1	<0.0001

4	A,B,C,D and all two-way interactions	105.96	10	<0.0001
5	A,B,C,D and all two-way interactions, + P ...... vs. Model 4	118.68 12.72	11 1	<0.0001 <0.0001
6	Risk groups (Low, Intermediate, High)	111.54	2	<0.0001
7	Risk groups, + P ...... vs. Model 6	111.84 0.30	4 1	<0.0001 N.S.

From these results we conclude that the NN prediction was able to substantially improve the Cox regression model fit when added to the four covariates, and even improved on a Cox model that included all two-way covariate interactions. As expected, it did not improve on a model in which the 'true' risk group assignments were represented.

Since the 4 covariate variables can take only 16 possible combination of values, we examined the NN prediction for all input combinations. In addition we calculated the median of the fitted Cox distribution for models with A-D, and A-D, plus P, as shown in the table below.

Group		Cases		Actual values		Difference (Predicted - Actual)			
ABCD	Risk	p	N	Mean	Median	NN vs. mean	NN vs. median	Cox (A-D)	Cox (P,A-D)
0000	Int	0.32	290	99	68	-12	+19	+6	+1
0001	Int	0.32	350	114	76	-24	+14	-7	-3
0010	Low	0.11	102	196	116	-73	+7	-41	+4
0011	High	0.11	100	64	51	+1	+14	+19	-2
0100	Int	0.04	35	108	97	-19	-8	-8	-23
0101	Low	0.04	49	183	132	-59	-8	-53	+5
0110	High	0.01	11	36	14	+16	+38	+76	+28
0111	Int	0.01	10	114	67	-66	-19	+13	-28
1000	High	0.016	14	46	16	+2	+32	+16	+16
1001	High	0.016	23	42	17	+3	+28	+10	+15
1010	High	0.006	4	51	41	+10	+20	-9	+1
1011	High	0.006	5	40	44	-9	-13	-17	-22
1100	High	0.002	2	26	9	-20	-3	+29	+6

1101	High	0.002	4	62	31	0	+31	+3	+16
1110	High	0.001	1	29	29	-29	-29	+10	-15
1111	High	0.001	0	-	-	-	-	-	-
Weighted absolute difference (days)						23.8	14.1	14.6	4.3

8. Breast cancer

The data

The original proposal included plans to analyze breast cancer from NSABP. This clinical trials group has several very large databases that would be ideal for NN modelling. At that time we had written assurances of collaboration from the NSABP. Subsequently, it became apparent that internal political difficulties, in the NSABP, was creating hold ups in the release of data to us. We visited their data center in Pittsburgh, to present an outline of our plans and methods, and had a return visit from a senior member of their statistical office. We had repeated assurances that they would be able to provide the data as promised. Unfortunately, the problems at NSABP got worse, with law suits being filed, and we were told that no data would be forthcoming until these matters were resolved.

At this point we sought data from other sources. Not any data set would do - we needed one that included fairly detailed information on each case, and with a large number of patients. For some time it appeared that we would be able to have access to data from the Mayo Clinic. We had approval from the statistician and senior investigator responsible for the data, but despite repeated reminders the data were never forthcoming.

We were more successful on our third try. We obtained data from Dr. Leslie Bernstein who has information on prognostic data from breast cancer patients treated at the Norris Cancer Hospital. The data set included 236 women with node negative breast cancer treated with surgery only. Covariates available included age, tumor size, nuclear grade, HER-2/neu amplification, histology, estrogen receptor status, and progesterone receptor status, although not all information was available for all patients.

Cox regression

In univariate Cox models, with disease free survival (DFS) as the endpoint, HER-2/neu amplification, tumor size, treatment center, and age all had significant associations with outcome. However in a stepwise multivariate Cox model, only HER-2/neu amplification and treatment center were accepted into the model: all other covariates were non-significant when controlling for the two selected covariates. Due to missing values, only 133 patients were available for the multivariate analysis.

Recursive Partitioning

The recursive partitioning method selected HER-2/neu amplification as the strongest predictor, and further divided the non-amplified group according to treatment center. The amplified group was also split, based on nuclear grade (3 vs. 1 and 2).

```
                                   No.    Obs.    Exp.
         Partitioned Tree          Cases  Evnt    Evnt    X²/Pct
           :-5...SITE = IOWA/WISC    76      8    14.25   98.7%
           |
         :-2...Her/2 = amplified    107     20    25.51    9.60
         | |
         | :-4...SITE = USC          31     12     5.75   93.5%
         |
       --1                          133     31             6.77
         |
         | :-7...GRADE > 2.00        16      5     7.99   93.8%
         | |
         :-3...Her/2 = not amplified 26     11     5.49    4.35
           |
           :-6...GRADE <= 2.00       10      6     3.01   80.0%
```

The initial partition was based on NER-2/neu amplification, which had a logrank chi-square value of 6.77. The amplified group was further split, based on treating hospital: the USC patients did significantly worse (chi-square = 9.60), although this probably reflected differences in referral patterns rather than in efficacy of treatment. The non-amplified group was split into two groups based on tumor grade (greater than 2 vs. 2 or less). This partitioning thus resulted in four prognostic groups, with one year DFS for these groups being 80%, 94%, 94% and 99%.

Neural Networks

A two level feed-forward neural network (NN) was fitted, using all covariates and the Buckley-James method for handling censored observations. In initial evaluations the NN predictor was far more significant than any single covariate. However, with only 31 events the 133 cases with complete data, there was a risk that the NN had over-fitted the data. This is a common problem with NNs, since they typically involve a large number of estimated parameters.

To avoid this bias a second 'bootstrap' analysis was conducted. The NN model was fitted 133 times, each time omitting one patient's data, and the predicted survival time was then calculated for that patient only. Using this approach, the NN predictions were much less valuable, and in fact in a stepwise regression, the amplification and site covariates were chosen for inclusion and the NN predictor did not provide significant additional information.

CONCLUSIONS

It is difficult to draw firm conclusions about the relative value of the alternate methods because of the limited number of patients, and events, in the dataset. The original proposal for this project called for analysis of data on women treated by the NSABP, but due to internal problems within

that organization this collaboration was not possible. Application of these methods to other, large breast cancer databases will be needed to fully evaluate the potential of these alternative methods of analysis of prognostic factors.

REFERENCES

Ashutosh K, Lee H, Mohan CK, Ranka S, Merotra K, Alexander C. Prediction criterai for successful weaning from respiratory support: statistical and connectionist analyses. Crit. Care Med. 20:1295-301, 1992.

Astion ML, Wilding P. The application of backpropagation neural networks to problems in pathology and laboratory medicine. Arch Pathol Lab Med 116:995-1001, 1992.

Buchman TG, Kubos KL, Seidler AJ, Siegforth MJ. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive care unit. Critical Care Medicine 22:750-762, 1994.

Buckley JD, James IR. Linear regression with censored data. Biometrika 66:429-36, 1979.

Buckley JD. Epilog Plus manual. Epicenter Software, 1993.

Buckley JD, Lampkin BC, Nesbit ME, Bernstein ID, Feig SA, Kersey JH, Piomelli S, Kim ZT, Hammond GD. Remission induction in children with acute non-lymphocytic leukemia using cytosine arabinoside and doxorubicin or daunorubicin. Med Ped Oncol. 17:382-90, 1989.

Ciampi A, Lawless JF, McKinney SM, Singhal K. Regression and recursive strategies in the analysis of medical survival data. J Clin Epidemiol. 41:737-48, 1988.

Cox DR. Regression models and life tables. JR Stat Soc B 34:187-220, 1972.

Davis GE, Lowell WE, and Davis GL. A neural network that predicts psychiatric length of stay. MD Computing 10:87-92, 1993.

Ebell MH. Artificial neural netowrks for predicting failure to survive following in-hospital cardiopulmonary resuscitation. J Fam. Pract. 36:297-303, 1993.

Edenbrandt L, Devine B, and Macfarlane PW. Classification of electrocardiographic ST-T segments-- human expert vs. artificial neural network. European Heart Journal 14:464-468, 1993.

Iezzoni LI and Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. Journal on Quality Improvement 20:305-316, 1994.

Knaus WA, Harrell FE , Lynn J et al. The SUPPORT prognostic model: objective estimates of survival for seriously ill hospitalized patients. Ann Intern Med 122:191-203, 1995.

Faraggi D and Simon R. A neural network model for survival data. Statistics in Medicine 14:73-82, 1995.

Fisher E, Redmond C, Fisher B et al. Prognostic factors in NSABP studies of women with node-negative breast cancer. JNCI Monogr. 11:77-83, 1992.

Gasparini G, Pozza F, Harris A. Evaluating the potential usefulness of new prognostic and predictive indicators in node-negative breast cancer patients. JNCI 85:1206-19, 1993.

Green J, Wintfeld N, Sharkey P, and Passman LJ. The importance of severity of illness in assessing hospital mortality. JAMA 263:241-246, 1990.

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

Harrell FE, Lee KL, Califf RM, Pryor DB, and Rosati RA. Regression modelling strategies for improvied prognostic prediciton. Statistics in Medicine 3:143-152, 1984.

Katz S. Personal communication, 1993.

Lapuerta P, Azen SP and LaBree L. The use of neural networks in predicting risk of coronary artery disease. Comput Biomed Res 28:38-52, 1995.

Liestol K, Andersen PK and Andersen U. Survival analysis and neural nets. Statistics in Medicine 13:1189-200, 1994.

Localio AR, Hamory BH, Sharp TJ, Weaver SL, TenHave TR and Landis JR. Comparing hospital mortality in adult patients with pneumonia: a case study of statistical methods in a managed care program. Ann Intern Med 122:125-132, 1995.

McCullagh P, Nelder JA. Generalized linear models. Chapman and Hall, New York, 1983.

Narayanan MN and Lucas SB. A genetic algorithm to improve a neural network to predict a patient's response to warfarin. Meth Inform Med 32:55-8, 1993.

Patil S, Henry JW, Rubenfire M, and Stein PD. Neural network in the clinical diagnosis of acute pulmonary embolism. Chest 104:1685-89, 1993.

Ravdin PM, Clark GM, Hilsenbeck SG, Owens MA, Vendely P, Pandian MR, McGuire WL. Demonstration that breast cancer recurrence can be predicted by neural network analysis. Breast Canc. Res. and Treat. 21:47-53, 1992.

Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back propagating. Nature 323:533-6, 1986.

Somoza E and Somoza JR. A neural network approach to predicting admission decisions in a psychiatric emergency room. Med Decis Making 13:273-280, 1993.

Tu JV and Guerriere MRJ. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. Computers and Biomedical Research 26:220-229, 1993.

Wilensky G and Manukian N. The projection neural network. Intl Joint Conf on Neural networks Vol II:358-67, 1992.

MEETING ABSTRACTS

Grant No. *DAMD17-94-J-4137*

P.I. Jonthan D. Buckley

Buckley, J., Van Tornout, J., Bernstein, L., Press, M., Flom, K. Neural networks for breast cancer prognosis. Proceedings of the Department of Defense Breast Cancer Research Program Meeting, Vol. III:1031-1032, 1997.

LIST OF PERSONNEL

PAID FROM GRANT NO. *DAMD17-94-J-4137*

P.I. Dr. Jonathan D. Buckley

Buckley, Jonathan D.
Howells, William B.
LaPuerta, Pablo
Meyers, Jeffrey L
Sosa-Cardenas, Maria
Van Tornout, Jan M.