

## NEURAL NETWORKS: ADVANTAGES AND LIMITATIONS FOR BIOSTATISTICAL MODELING

Philip H. Goodman, University of Nevada, Reno; Frank E. Harrell, Jr., University of Virginia, Charlottesville  
 Philip H. Goodman, Washoe Med Center, 77 Pringle Way, Reno, NV 89520 www.scs.unr.edu/nevprop

**Key Words:** Neural network, neurostatistical model, nonlinear regression, prediction, inference

- I. Rationale
- II. Basic Approach
- III. Examples
- III. New Directions
- V. Summary

### I. Rationale for “Neurostatistical” Models

Artificial neural networks (ANNs), also known as neurocomputational models, are computer algorithms that attempt to simulate the parallel, highly interactive distributed processing in brain tissue. But how does brain function relate to the analysis of predictive data sets with binary outcomes? Below, we describe the biostatistical application of ANNs to the analysis of observational health care data, in the form of what we call “neurostatistical” modeling. The methods are generally applicable, however, to any complex data.

A major methodological obstacle in drawing inference and making prediction from observational studies is that outcomes will be influenced *as much by differences in the underlying biological substrates* in the populations, as by the quality and content of medical interventions, because the treatments have not been randomly allocated to the patients. Health care researchers, providers, and insurers share a common need for accurate outcomes adjustment methodologies to assess the impact of treatment and determine the quality of medical care. An ideal adjustment system would: (1) rapidly adapt to changing disease and demographic patterns, (2) be robust to noise and errors in data entry, (3) optimally adjust outcomes for confounding influences, (4) squeeze maximal predictive information out of the data that would generalize to future patients, and yet (5) not be labor-intensive nor overly sensitive to the preferences of individual data analysts.

#### I.1 Generalized Linear Models

The most popular biostatistical methods are based on the *generalized linear model* (GLM):

$$y_i = \mathbf{X}\mathbf{b} + \mathbf{e}_i = \sum_j \mathbf{b}_j x_{ij} + \mathbf{e}_i$$

That is, the explanatory variable measurements,  $\mathbf{X}$ , obtained on patient  $i$  are multiplied by weights,  $\mathbf{b}$ , obtained by statistical estimation, and summed together for a score,  $Y$ . Because the random error term,  $\mathbf{e}$ , is assumed to have a mean of zero and constant variance, we can focus on the expected, systematic component of the regression:

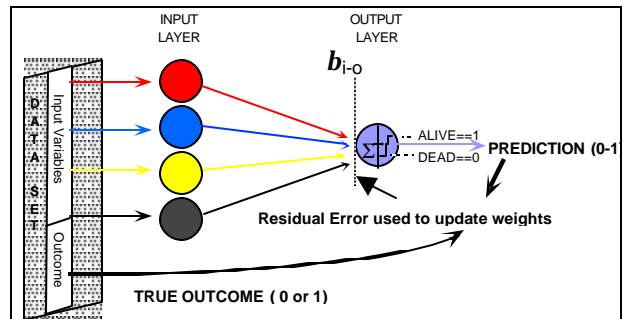
$$\mathbf{h}_i = \mathbf{X}\mathbf{b}$$

For simple linear prediction,  $\mathbf{h}$  may have intrinsic meaning (e.g., length of stay, cost of care). For binary outcome events (e.g., survival, cancer recurrence, quality of life), we require a link function that monotonically constrains the output prediction to lie between 0 and 1 (or some other limiting values). Because we are most often interested the probability,  $p_i$ , of an event, we usually use a logistic (log-odds) link:

$$\text{logit}(p_i) = \log(p_i/1-p_i) = \mathbf{h}_i = \mathbf{X}\mathbf{b} \text{ , or,}$$

$$p_i = \frac{\exp(\mathbf{h}_i)}{1 + \exp(\mathbf{h}_i)} = \frac{\exp(\mathbf{X}\mathbf{b})}{1 + \exp(\mathbf{X}\mathbf{b})}$$

Figure 1 shows a directed graph corresponding to the GLM. Recommendations for the appropriate biostatistical application of linear logistic models can be found in Harrell et al (1996).



**Figure 1. Directed Graph Representation of A Generalized Linear Model with Binary Outcome** showing flow in information (rightward), and conceptual flow of error gradient used for optimization (leftward). Value of input explanatory (predictor) variables are transmitted to the output unit, where they are multiplied by corresponding the weights,  $\mathbf{b}$ , summed, and passed through a monotonic inverse link function that constrains the output between the binary levels. Optimization proceeds through matrix inversion in the case of a linear link, or iterative gradient-based methods for nonlinear links functions.

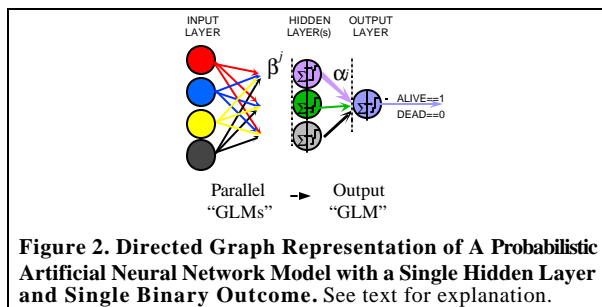
## I.2 Artificial Neural Network Models

There is no reason, *a priori*, to believe that health risk factors work together in a linear manner. But in the absence of detailed knowledge of the exact interactions of organ systems, genetics, and disease, how can we determine if the linear model is the best approach? And if not, how do we construct a better, nonlinear, predictive instrument?

We do know that human brains process data in an automated, highly nonlinear fashion. Despite the fact that these brain cells conduct signals much more slowly than silicon-based electrical circuits, we know that human brains do it faster and better using parallel intercellular connections. Our brains very rapidly recognize complex patterns and anticipate outcomes in complicated sensory environments, exceeding the ability and accuracy of even the fastest supercomputers. Yet when it comes to exact numeric processing, our brains are *not* good at giving precise estimates of risk or uncertainty—things that can be done perfectly by a hand-held calculator.

The goal, therefore, of applying ANNs to binary outcome data is to augment linear modeling with brain-like algorithms. It is hoped that this added flexibility will model but not overfit underlying nonlinear and interactive relationships among variables, yielding superior bias-adjusted accuracy in the analysis of complex health care databases.

The simplest formulation of a probabilistic ANN can be visualized as the serial linkage of logistic GLMs (Figure 2). Explanatory variable measurements at the input layer are transmitted *simultaneously* to *several* parallel probabilistic “GLMs” (only the systematic structure of the GLM is actually used; optimization of the weights is performed globally on the full ANN model). Weighting, summation, and logistic transformation occur at “hidden” units. The computed values at these hidden units are then transmitted to (one or more) output unit(s), which again perform weighting, summation, and logistic transformation to yield predicted probability. It is the parallel layering and diverging-converging flow of signals that is analogous to layers of interconnected neurons (neural networks) in the brain. This general design is also referred to as a multilayer perceptron.



**Figure 2. Directed Graph Representation of A Probabilistic Artificial Neural Network Model with a Single Hidden Layer and Single Binary Outcome.** See text for explanation.

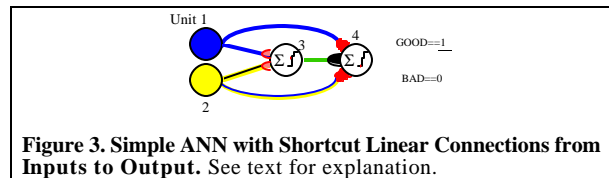
The input-to-hidden “parallel GLMs” essentially perform a piecewise-linear transformation of the data into hidden components, which are then combined at the output unit GLM. This is also analogous to simultaneously fitting nonlinearities and interactions using nonparametric splines. However, as Barron (1993) has shown, for high dimensional data, ANNs offer a theoretical and computational advantage: the residual error falls as  $O(1/M)$ , where  $M$  is the number of hidden units in the ANN, irrespective of the number of inputs; whereas the error falls as  $O(1/M^{2/d})$  for  $d$  inputs using fixed-function polynomial (e.g., spline) or any other series expansion in a GLM. The tradeoff for this efficient dimensional scaling is the computational burden necessary to optimize and interpret effects in the nonlinear ANN regression model compared to the GLM.

Using a sufficient number of hidden units, this piecewise-linear transformation can model the systematic component of any smooth function, including arbitrary interactions among the input explanatory variables. Conceptually, then, the ANN formulation captures in a natural way, the arbitrarily complex nonlinearities in predictors as well as interactions among them. That is, there is no need transform predictors nor create artificial variables (such as multiplicative products of predictors or explicit spline terms).

The equation corresponding to the systematic prediction on case  $i$  using a single hidden-layer ANN is:

$$\begin{aligned} \log_i(p_i) &= \sum_j^{\text{hidden units}} \mathbf{a}_j \langle \text{hidden unit output} \rangle_j \\ &= \sum_j^{\text{hidden units}} \mathbf{a}_j \left\langle \frac{\exp(\mathbf{h}_j^i)}{1 + \exp(\mathbf{h}_j^i)} \right\rangle = \sum_j^{\text{hidden units}} \mathbf{a}_j \left\langle \frac{\exp(\mathbf{X}\mathbf{b}^j)}{1 + \exp(\mathbf{X}\mathbf{b}^j)} \right\rangle \end{aligned}$$

where we recognize that there are  $J$  sets of GLM-like input-to-hidden weights  $\mathbf{b}^j$ .



**Figure 3. Simple ANN with Shortcut Linear Connections from Inputs to Output.** See text for explanation.

In practice, we often add direct shortcutting connections from input to output in an ANN model (Figure 3). Conceptually, these linear terms represent direct linear effects, which are usually present in real-world data sets. During optimization, then, the hidden unit weights are available to fit predictor nonlinearities and interactions (although they may pick up piece-wise portions of the linear components as well, as there is no constraint on this behavior). In general, adding these linear terms lends stability and accelerates convergence of optimization. More detailed discussion of ANNs from an

applied statistical perspective may be found in Bishop (1995) and Goodman (1998).

### I.3 Frequentist Optimization vs Bayesian Analysis

The actual process of fitting parameters to a any statistical model using existing data may be approached using either “frequentist” analysis or “Bayesian” analysis. ANNs provide a hybrid modeling environment.

Frequentist analysis seeks the single, most likely set of parameters that satisfy the constraints of a model—e.g., logistic regression for the probability models under consideration here. While regression may be computationally frugal, the need to determine the statistical significance of these results involves a retrospective evaluation of the estimation procedure over the distribution of possible observations that could have been observed assuming a “true” but unobserved set of parameters. The interpretation of resulting  $p$  values or confidence intervals is, therefore, not straightforward.

In particular, for nonlinear regression models like ANNs, it is the model’s *predictions* and the *net effects of predictors* that are of intrinsic interest, not the architecturally- and initialization-sensitive values of the many individual, interacting parameters. Therefore, frequentist assumptions are suspect, and the generation of  $p$  values or confidence intervals for individual parameters is problematic.

Alternatively, the “Bayesian” approach bases inference on parameters and predictions directly on the observed data—no single correct solution to the model is assumed. Bayesian models are formulated in probabilistic terms, which facilitates the interpretation of confidence boundaries. The objective is to properly weight competing models, and, within each model, to integrate over a *distribution* of parameters (rather than a single most-likely solution). Weaknesses of Bayesian analysis include (often extreme) computational intensity, and a need to make assumptions (subjectively but explicitly) about the forms of prior parameter distributions.

In certain knowledge domains, scientists can formulate very specific mathematical models for the processes they wish to better understand. For instance, in drug pharmacokinetics (distribution in multiple body compartments based on well-characterized tissue affinities), and in nerve cell signal processing (accepted approximations of cable theory and synaptic physiology). Although both of these examples comprise nonlinear models, the relationships among their predictors are explicit, and inference follows readily after maximum likelihood optimization. Such inference can be used to test the scientific hypotheses that motivated the models and data collection. Frequentist methods are very appropriate in such models.

In other knowledge domains (especially with complex, but poorly understood relationships among variables, as we assume to be true in the present case) explicit or analytic formulations of parametrized models are simply not available. Modern computation power has promoted the emergence of several approaches to the description of these relationships, including ANNs and artificial evolutionary programs (genetic algorithms). Both methods emulate “tricks” nature uses to search for good solutions.

While ANNs are usually implemented as regression models (i.e., maximum likelihood optimization is used to fit parameters), the sensitivity of parametric fit to initialization and data-sensitive connectivity is reminiscent of Bayesian concerns that distributions of parameters be considered. While direct application of Bayesian methods to ANNs is theoretically and computationally complex, such methods *can* be hybridized with maximum likelihood methods improve ANN inference. For example, as discussed in section III, implementation of dynamic hyperpriors into the optimization algorithm can be used to obtain data-focused rather than individual parameter-driven estimates of effects, and to determine the relevance of predictors without the use of stepwise procedures.

## II. Basic Approach to Neurostatistical Modeling

Because of computational and interpretive simplicity, we would certainly prefer a GLM if it *is* the “true” model, but how do we know when that’s the case? A traditional approach would be to statistically compare a series of linear models, with and without various (judiciously chosen) nonlinear and interaction terms involving the predictors, and with various subsets and transformations of the predictors. The (statistically) dangerous parts of this activity are making type I errors by performing multiple tests on the same data, and overfitting. Furthermore, choices of nonlinear transformation and forms of interaction will vary among analysts.

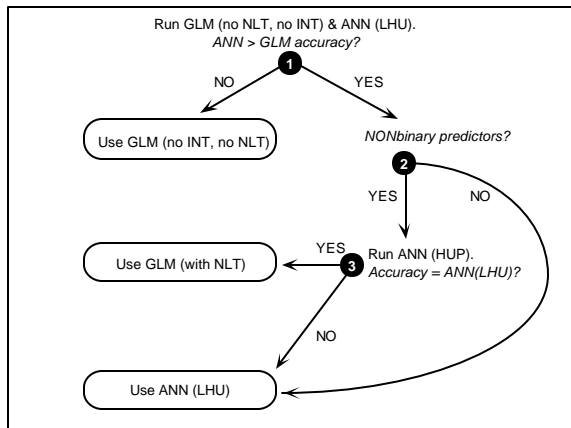
A basic approach to augmenting linear analysis with neurostatistical modeling is summarized in Table 1. Basic data cleaning should be carried out, as for any biostatistical modeling. Missing data may be imputed or deleted (as long as the subsequent inference would be acceptable). At this early stage, variable transformation and selection (by consensus, exploratory analysis, or stepwise regression) should be avoided.

Based on our work to date (Goodman and Harrell, 1998), we suggest the following integrated ANN-GLM strategy (Figure 4). First, a single hidden-layer ANN is used to “screen” for nonlinear relationships. As discussed in section I, a sufficiently flexible ANN (i.e., with enough hidden units) can emulate arbitrarily

complex nonlinear transformations and interactions among the predictors.

Table 1. Basic Approach To Effective and Efficient Neural Network Modeling	
Step 1.	Confirm data integrity; appropriately impute
Step 2.	Define linear model
Step 3.	Define nonlinear model
Step 4.	Define regularization
Step 5.	Fit linear model
Step 6.	Fit nonlinear ANN model
Step 7.	Compare numeric discrimination statistics (e.g., C index / ROC area)
Step 8.	Compare numeric global fit statistics (e.g., -LL, Nagelkerke R2, Brier Score)
Step 9.	Compare graphical calibration statistics (e.g., apparent:predicted probabilities)
<i>If the ANN outperforms the GLM:</i>	
Step 10.	Compare mean effects
Step 11.	Compare variable selection

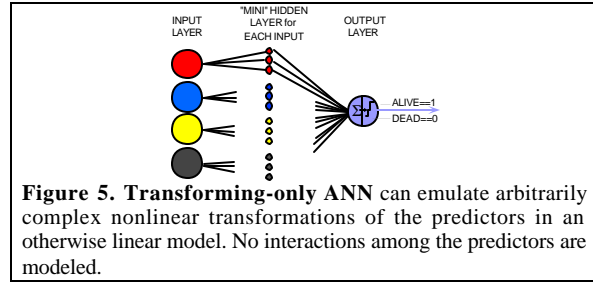
If the predictive ability of an ANN is not statistically better than a simple GLM (i.e., with no interactions or nonlinear terms), there is no point in entertaining more complex GLM models. This approach results in more efficient modeling and fewer type I errors. An example is provided in the next section.



Abbreviations: GLM, generalized linear model (e.g., multiple linear regression, logistic regression)  
 NLT, nonlinear transformations of predictors (e.g., spline, polynomial)  
 INT, interactions terms involving predictors  
 ANN, artificial neural network with adaptive regularization  
 LHU, layer of hidden units common to all predictors  
 HUP, several hidden units from each predictor, none in common

**Figure 4. Basic Approach to Screening for Complexity.** By comparing the accuracies of a GLM and ANN, an initial decision may be made about the need for incorporating any predictor nonlinearities or interactions.

The second stage of Figure 4 need be taken only if the ANN outperformed the GLM. If all predictors are binary, the improved performance of the ANN must be due to interactions; therefore, proceed to use the full ANN model (or use ANN software to discover important interactions, and incorporate these into a GLM). If there are nonbinary predictors, fit a



**Figure 5. Transforming-only ANN** can emulate arbitrarily complex nonlinear transformations of the predictors in an otherwise linear model. No interactions among the predictors are modeled.

“transforming-only” ANN model (Figure 5), which uses several hidden units for each nonbinary predictor, but no hidden layer common to the predictors. This is analogous to flexible spline transformation. If the accuracy of the “transforming-only” ANN is essentially the same as the full ANN, the improvement is attributable to predictor nonlinearities rather than interactions; it is therefore best to return to the GLM and find suitable transformations (e.g., restricted splines). However, if the full ANN model remains significantly better, use the full ANN implementation.

### III. Example Analyses of Binary-Outcome Data

#### III.1 Simulated Surgery Data Set

It makes sense to first assess the neurostatistical approach to “designer” data that incorporates the types of predictor nonlinearity, interaction, and noise we want to identify in real-world databases. We took several approaches to generate sets of explanatory variables, including the modification of real data sets and creation entirely *de novo*. In all cases, the neurostatistical approach correctly identified those data sets with only linear relationships, and performed close to (known) Bayes-optimal accuracy with properly regularized nonlinear ANNs when nonlinearities or interactions were present. The methods and findings are described in detail in Lowe et al (1995). For example, we generated 3 continuous (normally distributed with differing variances) and 2 binary predictors (binomial with differing variances), and an outcome variable that is the sum of nonlinear transformations and multiplicative interactions among the predictors. We used a Cauchy cumulative distribution function to constrain the predictions to [0,1] and allow us to flexibly introduce noise.

**Table 2. Simulated Data Set: Linear vs ANN Discrimination**

	Logistic GLM ("vanilla LRM")	Preprocessed GLM ("preprocess LRM")	ANN w/stopped training	BAYES OPTIMAL
c Index, bias corr.				
Before:	.505	.679	.901	
After:	.491	.675	.889	.902
c Index on separate validation data:	.493	.668	.879	.902

As shown in Table 2, using the c index (approximate area under the ROC curve) as a measure of discriminative accuracy, the out-of-sample validation c index for the simple logistic GLM was 0.493, which improved to 0.668 when restricted cubic spline terms were introduced to fit predictor nonlinearity. The corresponding nonlinear ANN c index was 0.879, close to the known Bayes optimum of 0.902. A similar pattern was observed for the AIC. These and other simulations make us quite confident that the neurostatistical approach will readily detect clinically meaningful nonlinearity and interactions present in real data distributions.

### III.2 Duke Coronary Artery Surgery Data Set

The Duke coronary artery bypass grafting data set is described in Table 3.

**Table 3. Description of Duke CABG Data Set**

5516 patients underwent CABG, 1/1/86 and 12/31/92

**Outcome variable:** 30-day mortality (binary)

**Pre-operative risk-factor variables:** (34: some transformation, imputation, and aggregation into indexes performed in 1994 by Dr. Harrell at Duke)

*Binary:* (15 recoded as 22) sex, race, redo CABG, previous M, Q-waves on ECG, index for recent MI/pattern of angina (coded as 8 binaries), HTN, DM, smoking, dyslipidemia, FH of premature CAD, mitral insufficiency, cardiogenic shock, PTCA performed within 48 hours, preoperative insertion of intraortic balloon

*Polytomous:* (2) NYHA CHF functional classification (1-4), index of conduction defects (0-6)

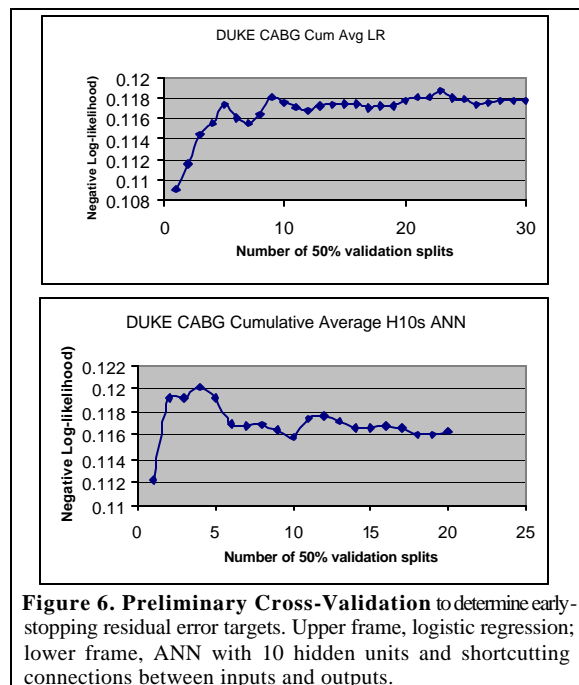
*Continuous:* (10) date, age, weight, BSA, height, duration of CAD, LVEF, surgeon's volume of similar surgeries, index of vascular disease, index of CAD

Step 1. Data integrity confirmed: NevProp4 ANN software (Goodman, 1998) produced descriptive statistics that were identical to Dr. Harrell's earlier documentation generated using S-Plus® software.

Step 2. Linear model defined: Commercial regression package or ANN without hidden units (direct linear shortcuts between inputs and the output variable).

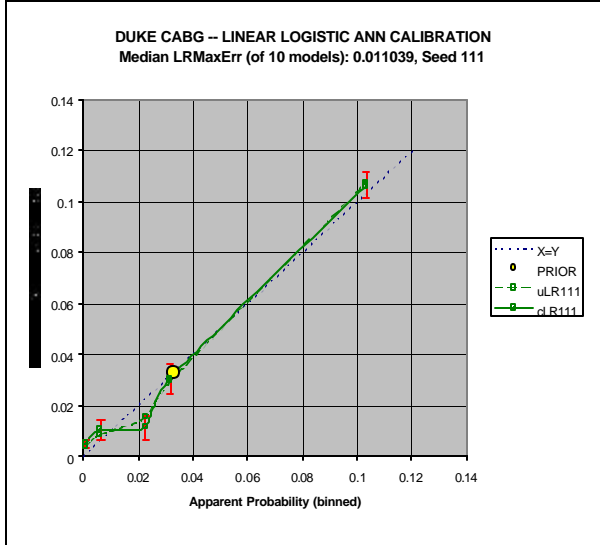
Step 3. Nonlinear ANN model defined: In addition to linear shortcuts, 10 hidden units (completed connected from inputs) were included run to determine whether that parameterization provided sufficient flexibility to easily overfit the data. Within 200 iterations of optimization, the model log-likelihood approached zero, indicating more than sufficient parametric flexibility.

Step 4. Regularization defined: Rather than splitting the data into training and testing sets, we generally prefer to use all the data for model development (even with thousands of cases, data space is sparsely represented when there are many predictors) For discussions on this issue, see Harrell et al (1996), Hutton et al (1995). To adjust the final estimates of accuracy for optimistic bias, the data is bootstrapped and the entire model development process is repeated in an automated fashion. In addition, to prevent overfitting of the data (especially by the nonlinear ANN), we used early stopping of optimization based on reaching a mean target error determined by an automated, preliminary phase of multiple 50% data-splitting cross-validation (for actual model development, the intact data set is used). For both the linear (logistic regression) and nonlinear ANN model, NevProp4 produced a graphic summary of the convergence of these cumulative mean targets as a function of the number of cross-validation splits for each model. In this case, convergence patterns indicated that up to 30 splits were needed for stability on the logistic regression, whereas only 15 splits was sufficient for convergence of the ANN (Figure 6).

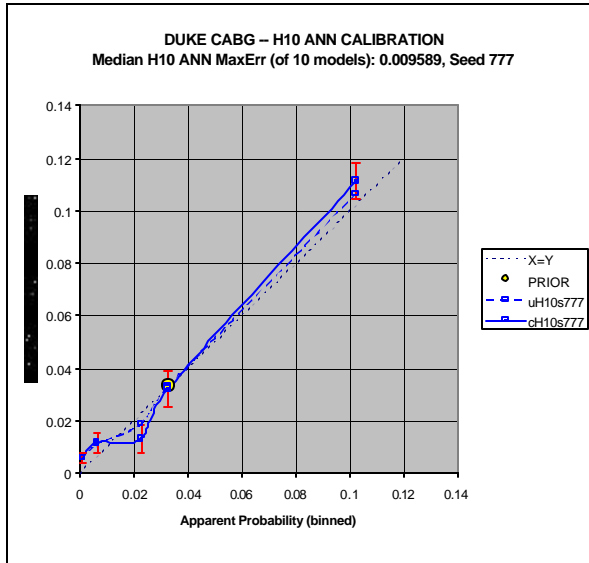


Step 5 & 6. Linear (logistic regression) and nonlinear ANN models fitted: Because random seed selection for the pseudorandom number generator affects the stopping target determined by cross-validation, 10 repetitions of the optimization were performed to determine consistency of the method. A visual check confirmed that 100 bootstrapped models was more than





**Figure 9. Graphical Calibration Estimates** for the linear logistic regression model. Results shown here were typical. Dashed line, smoothed raw calibration relationship; solid line, calibration adjusted for optimistic bias using bootstrapped models.



**Figure 10. Graphical Calibration Estimates** for the nonlinear ANN with 10 hidden units. Results shown here were typical. Dashed line, smoothed raw calibration relationship; solid line, calibration adjusted for optimistic bias using bootstrapped models.

We conclude, therefore, that a basic linear logistic model fully utilizes the information present in the Duke database to predict 30-day surgical mortality. That is, based on the data provided, no curvilinear transformation of the predictors, nor addition of spline or interaction terms, will meaningfully improve upon the generalized accuracy of mortality prediction in this population. Armed with this information, the

biostatistician would not feel compelled to explore additional GLMs.

Although recent cancer survival data sets showed improved performance using ANNs (Burke et al, 1997), we recently applied the neurostatistical approach to 4 large surgical mortality data sets and found no evidence of nonlinearity or interaction. We were somewhat surprised by these findings, and hypothesize that this phenomenon represents three influences, which warrant further methodological investigation:

- the assortment of predictors used by the original investigators reflects “what worked” in prior studies that were analyzed with simple linear models, thereby precluding other measurement that might have added information if modeled with nonlinear effects;
- incorporating a large number of predictors in the model may, in aggregate, provide piecewise linear (hence nonmonotonic) predictions that are essentially nonlinear effects (but this would not account for interactions among predictors); and/or,
- measurements available were weak, first-order surrogates for the true underlying (highly interactive) biological processes that ultimately determine survival and response to medical intervention.

### III.3 Comparing Effects and Selecting Variables

Had the bias-adjusted nonlinear ANN outperformed the GLM by some meaningful measure of accuracy, we would prefer the ANN for prediction on future data drawn from the same population. Drawing inference on the predictors is more complex, however, because there no longer exists a *unique* identifiable relationship between a specific weight and the effect of a predictor. An approach using a *mean effect* (with confidence intervals obtained by bootstrapping) is presented in Goodman (1998). The mean effect for a predictor is defined as:

$$MeanEffect^{ANN} \text{ of } x_i = \mathbf{b}_i^{IP(Path_{i,j})} + \frac{1}{N} \sum_{T=1}^{N_{cases}} \left( \sum_{j \in paths} \left[ \mathbf{b}_{path}^{H-O} \prod_{j \in path} (\mathbf{b}_j^{H-H} Act_j^{H-I}) \right] \right)$$

where Act’ is the derivative of the activation (inverse link) function at a hidden unit along one of the paths from the particular input to the output unit. Note that there is a specific effect for each case, because the predictor vector determines the activation, hence its derivative, at each hidden unit. Although the weights are obtained by optimization methods, this data-dependent, rather than  $\mathbf{b}$ -focused approach to inference reflects a Bayesian attribute of nonlinear ANNs.

Another way to introduce Bayesian estimation is in the selection of relevant predictor variables. Although stepwise variable selection processed are equally applicable to nonlinear ANN models, the computational overhead (typically 1-2 orders of magnitude for a nonlinear ANN compared with an iteratively-optimized GLM) led to our interest in Bayesian-motivated model reduction. The approach we used is called automatic relevance determination (ARD), which incorporates an adaptive hyperpenalty function to adjust the penalty on each parameter based on the iteration-by-iteration variance and eigenstructure represented in the Hessian matrix (Bishop 1995, chapter 10; Mackay 1995, 1998). In its application to nonlinear ANN models, it includes additional penalty hyperparameters that generate a competition among connections from the multiple inputs to each hidden unit, and between hidden units. ARD provides an aggressive regularization (which should allow use of the full data set to estimate generalizable accuracy) which dynamically shrinks parameters with high variance, resulting in a relevance estimate upon conclusion of model fitting.

No. of well-determined parameters in the model: 21.2, or 60%, of 35 total.

GROUP	FROM	TO	ARD HYPERPARAMETERS	# Well-determined
0	bias	output1	0.07	1.0, or 99%, of 1
1	input1	output1	137.25	0.5, or 59%, of 1
2	input2	output1	2071.05	0.1, or 5%, of 1
3	input3	output1	244.93	0.4, or 36%, of 1
4	input4	output1	48.10	0.8, or 75%, of 1
5	input5	output1	814.43	0.1, or 10%, of 1
6	input6	output1	381.06	0.2, or 22%, of 1
7	input7	output1	40.47	0.7, or 70%, of 1
8	input8	output1	2594.67	0.6, or 63%, of 1
9	input9	output1	183.32	0.5, or 45%, of 1
10	input10	output1	8.51	0.9, or 93%, of 1
11	input11	output1	54.80	0.8, or 82%, of 1
12	input12	output1	123.64	0.5, or 52%, of 1
13	input13	output1	31.80	0.8, or 79%, of 1
14	input14	output1	192.18	0.6, or 58%, of 1
15	input15	output1	70.79	0.6, or 59%, of 1
16	input16	output1	50.07	0.6, or 75%, of 1
17	input17	output1	324.20	0.3, or 34%, of 1
18	input18	output1	2444.65	0.0, or 0%, of 1
19	input19	output1	250.49	0.4, or 38%, of 1
20	input20	output1	221.63	0.4, or 42%, of 1
21	input21	output1	63.53	0.6, or 61%, of 1
22	input22	output1	565.61	0.2, or 19%, of 1
23	input23	output1	24.24	0.9, or 85%, of 1
24	input24	output1	29.83	0.9, or 90%, of 1
25	input25	output1	360.03	1.0, or 98%, of 1
26	input26	output1	41.79	0.9, or 87%, of 1
27	input27	output1	186.78	0.5, or 46%, of 1
28	input28	output1	1893.31	0.0, or 2%, of 1
29	input29	output1	23.54	0.9, or 91%, of 1
30	input30	output1	117.90	0.7, or 67%, of 1
31	input31	output1	61.67	0.7, or 71%, of 1
32	input32	output1	27.75	1.0, or 95%, of 1
33	input33	output1	8.12	1.0, or 97%, of 1
34	input34	output1	315.61	1.0, or 99%, of 1

A \*\*\* indicates this variable was selected to stay in the SAS Proc Logistic stepwise elimination model (F to leave 0.05).

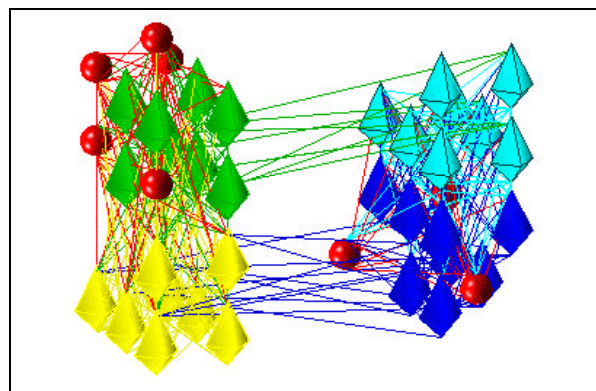
**Figure 11. Automatic Relevance Determination (ARD).** Hyperparameters (hyperpriors) and number of well-determined parameters are computed from a recursively updated Hessian matrix.

Figure 11 shows the results of applying ARD during optimization of the Duke CABG data set. Because the neurostatistical approach showed no meaningful nonlinearity, we expect that ARD should find the same predictors to be important as we obtain by performing stepwise regression on the GLM. Note the strong correspondence of the 11 ARD-selected variables (those > 80% determined) to the results of the SAS® Proc Logistic backwards elimination. The only discordant variables were input 1, NewMI category 1, (retained by

stepwise, but only 59% determined) and input 22, Q waves on EKG (retained by stepwise, but only 19% determined).

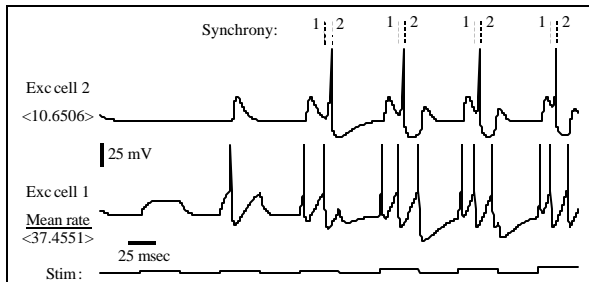
#### IV. New Directions in Neuro-Statistical Modeling

As discussed above, artificial neural networks currently in use for statistical modeling are actually nonlinearly linked GLMs. As such, they pick up systematic (mean) structural relationships in the conditional distribution of data. This ANN design is thus motivated by the concept of *mean* neuronal activity in brain networks (i.e., the greater the net signal arriving at a neuron, the greater its output). However, there is no good evidence that these basic ANNs really represent the powerful pattern-recognizing properties demonstrated even by animals with the most primitive central nervous systems. Real neurons transmit information in the form of brief voltage spikes, called action potentials, that are generated when the cell's membrane voltage is raised above a threshold. Whether threshold is reached reflects the number and timing of spikes arriving from other cells, as well as a host of other influences, including the history of recent firing (mainly due to intracellular calcium accumulation and ion channel dynamics), surrounding ongoing activity, and hormonal stimulation. One of the authors (PHG) has programmed compartmental models of biological spiking networks (Figure 12), which readily demonstrates that signalling in even such simple biological networks are not consistent with a systematic input-output relationship of the multilayer perceptron (Figure 13).



**Figure 12. Biologically Realistic Neural Network.** Neurons on the left represent a brain region interconnected with probability 0.1 to cells in the region on the right. Spheres are inhibitory, pyramidal are excitatory cells. Within a region pyramids, cells are divided into input (lighter) and output (darker) layers.

There is now a general consensus among neuroscientists that the central nervous system represents information by spike-dependent synchrony principles (Maass, 1997a), rather than mean rates of firing. It has recently been shown (Maass, 1997b) that biologically-realistic spike-synchronous networks can, as a minimum, readily encode all the information needed to function as multilayer perceptron ANNs. Simultaneously, they encode temporal binding information, which may be the attribute necessary for pattern recognition. Essentially, spike-synchronous networks add the temporal dimension of information above the space containing the stationary distribution of data. Spike-synchronous artificial neural networks thus offer the promise of improved prediction and inference, especially when there are temporal relationships among predictor variables, sequential outcomes, or time-series of events. Improved pattern-recognition may also enhance the ability of ANNs to automatically identify relevant predictors.



**Figure 13. Intracellular Voltage Recordings from Two Neurons in the Compartment Model.** As steps of increasing stimulation current (lower square-wave tracing) are injected into neuron (lower cell tracing), that cell responds initially with increased number of spikes (average activity), but then only with closer spacing of the spikes. The upper cell, which receives input from the lower cell, reaches threshold upon the receipt of 2 (or more) spikes. Although it does not produce a second spike with increased input, its firing remains synchronous with the second input spike (indicated by synchrony bars along top of figure).

## V. Summary

The current generation of artificial neural networks may be conceptualized as nonlinearly-linked GLM systematic structures. If properly regularized, ANNs will therefore reduce to GLMs in the absence of significant predictor nonlinearity and interaction. In such cases, ANNs essentially subserve a screening role, allowing the analyst to avoid complex transformations of explanatory variables and trial-and-error searching for interactions.

When a properly-regularized and bias-adjusted ANN outperforms the corresponding GLM, its superior predictive performance makes it a valuable addition to the armamentarium of the biostatistician. Interpreting effects and selecting variables remains computationally burdensome, however. The advantages and disadvantages are summarized in Table 6.

A promising new direction of research is to develop biologically-realistic spike-synchronous models that retain the advantages of current ANNs while adding pattern-recognition abilities.

## ACKNOWLEDGMENTS

This work was supported in part by grants from the U.S. Public Health Service Agency for Health Care Policy and Research (HS06830: "Outcomes by Neurocomputing") and the U.S. Department of Defense (DAMD17-94-4-4383: "Developing the AJCC Prognostic System for Breast Cancer").

**Table 6. Neurostatistical versus Generalized Linear Modeling**

<u>ADVANTAGES</u>	<u>LIMITATIONS</u>
+ Same link function as linear logistic regression	– Predictive CIs more computationally demanding
+ Capture predictor nonlinearity	– Nonlinear effects more difficult to interpret
+ Capture interactions	– Interactions may be difficult to identify
+ Minimize preprocessing biases	– Minimize preprocessing biases
+ Inherent Bayesian attributes	– Full Bayesian requires multiple ANNs
+ Minimize FP associations due to data snooping	– Overfitting occurs if not properly regularized
+ Screen large data sets for meaningful nonlinearities & interactions	– If ANN > LR, inference on predictive effects more difficult to draw (if needed)
+ Capability may be expected by collaborators	– Less established theory and accepted software

## REFERENCES

- Barron, AR (1993). Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transaction on Information Theory* 39(3):930-945.
- Bishop CM (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.
- Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857-62 1996 Feb 15.
- Goodman PG (1998). NevProp4 Artificial Neural Network Software and Manual. Available from <http://www.scs.unr.edu/nevprop>.
- Goodman PG, Harrell FE Jr (1998). Effectiveness & Outcomes by Neurocomputing. Narrative summary of USPHS/AHCPR project HS06830. Available from <http://www.scs.unr.edu/nevprop>.
- Harrell FE Jr, Lee KL, Mark DB (1996). Tutorial in Biostatistics. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* 15:361-387.
- Hutton LV, Goodman PH, Lowe W. Estimating Out-of-Sample-Performance with Bootstrapped Neural Networks. *Proceedings of the Fourth Golden West International Conference on Intelligent Systems*, June 12-14, 1995, San Francisco.
- Lowe W, Goodman PH, Hutton LV (1995). Simulating Nonlinear Prediction Data Sets. *Proceedings of the Fourth Golden West International Conference on Intelligent Systems*, June 12-14, 1995, San Francisco.
- Maass, W (1997a). Networks of Spiking Neurons: The Third Generation of Neural Network Models. *Neural Networks* 10(9):1659.
- Maass, W (1997b). Fast Sigmoidal Networks via Spiking Neurons. *Neural computation* 9(2):279.
- Mackay DJ (1995). Probable Networks and Plausible Predictions - A Review of Practical Bayesian Methods for Supervised Neural Network. Available from <http://wol.ra.phy.cam.ac.uk/mackay/abstracts/network.html>.
- Mackay DJ (1998). *Information Theory, Inference and Learning Algorithms*. Prepublication available at <http://wol.ra.phy.cam.ac.uk/pub/mackay/itprnn/#book>.