Applying Artificial Neural Network Models to Clinical Decision Making

Rumi Kato Price
Washington University School of Medicine

Thomas J. Downey and Donald J. Meyer
Partek, Inc.

Edward L. Spitznagel
Washington University

Nathan K. Risk
Washington University School of Medicine

Omar G. El-Ghazzawy
Washington University

Because psychological assessment typically lacks biological gold standards, it traditionally has relied on clinician's expert knowledge. A more empirically based approach frequently has applied linear models to data to derive meaningful constructs and appropriate measures. Statistical inference are then used to assess the generality of the findings. This article introduces artificial neural networks (ANNs), flexible nonlinear modeling techniques that test a model's generality by applying its estimates against "future" data. ANNs have potential for overcoming some shortcomings of linear models. The basics of ANNs and their applications to psychological assessment are reviewed. Two examples of clinical decision making are described in which an ANN is compared with linear models, and the complexity of the network performance is examined. Issues salient to psychological assessment are addressed.

In the area of psychological assessment, a computer-based algorithm approach has relied heavily on the expert knowledge of clinicians. In a more empirically based approach—used when insufficient knowledge is available, existing knowledge is debated, or a special population is investigated—linear models such as factor analysis or regression models are frequently applied to empirical data to derive meaningful constructs and appropriate measures. In this approach, statistical inferences are needed to assess the generality of constructs and rules because the gold standard of expert knowledge may not be applied.

Making a statistical inference generally requires estimating the standard error of the parameter; hence linear models have most frequently been used. Linear models, however, are not appropriate for all instances. Loss of information is inevitable when extreme values of a measure must be truncated or rescaled to fit data to a linear scale. In multivariate linear models, emphasis on statistical significance often leads to situations in which a substantively meaningful predictive measure with a large effect and a large standard error may be dropped in favor of another measure with a small effect and an even smaller standard error (Archen 1982). Such practice limits the ability of these models to construct a potentially more predictive model.

A more drastic approach than that exemplified by the current linear models in psychological research can be taken by relying primarily on empirical knowledge and minimizing the importance of generalization by statistical inference, if the assumption is accepted that gold standards in psychological assessment exist only at a data-specific level. Such an assumption may be reasonable for psychological assessment given the lack of biological gold standards. Once such an approach is taken, validity assessment depends more on the ability of the constructed gold standards to predict "future" data. The importance of making statistical inferences is reduced if future data are indeed available to validate the empirically obtained constructs and rules. The importance of capturing nonlinear association becomes an even more pertinent issue because increased precision for prediction is needed with validity testing against real data.

In this article, we introduce artificial neural networks (ANNs) suitable for a variety of clinical decision problems such as assessment of psychological state, diagnosis of psychiatric disorders, and prediction of behavior outcomes such as suicide attempts, hospitalization, and death. ANNs have become important in a variety of areas in psychology and medicine, but their application to diagnostic decision making in psychology and psychiatry has been minimal.

The basics of multilayer perceptron (MLP) modeling, the most frequently used type of ANN, are described after a brief introduction to ANNs. We provide two examples of how an ANN can be used in judgment tasks. The first example uses item questions to predict a psychiatric diagnosis; the second example uses behavioral measures collected in early adulthood to predict long-term

Rumi Kato Price and Nathan K. Risk, Department of Psychiatry, Washington University School of Medicine; Edward L. Spitznagel, Department of Mathematics, Washington University; Thomas J. Downey and Donald J. Meyer, Partek, Inc., St. Peters, Missouri; Omar G. El-Ghazzawy, Department of Chemistry, Washington University.

Preparation of this article was supported in part by the Independent Scientist Award (K02DA00221) and Research Grants R01DA07939, R01DA09281, and R01DA10021 from the National Institute on Drug Abuse of National Institutes of Health. A portion of this article was first presented at the Annual Scientific Meeting of the College on Problems of Drug Dependence, San Juan, Puerto Rico, June 1996. We thank Paul Monroe for his critical and constructive comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Rumi K. Price, Department of Psychiatry, Washington University School of Medicine, 40 N. Kingshighway, Suite 2, St. Louis, Missouri 63108. Electronic mail may be sent to price@rkp.wustl.edu
mortality. The details of ANN methodology in these contexts include ANN parameter specification, software, cross-validation, and the receiver operating characteristic (ROC) evaluation. The performance of the ANN is compared with the performance of the linear statistical analyses frequently used in the area of psychological assessment: linear and quadratic discriminant analyses and logistic regression analyses. We examine the performance of ANN models with varying numbers of hidden neurons to show when complex network architecture is needed to achieve a better level of prediction than that achieved in linear models. We then discuss issues for research practitioners to consider in order to make optimal use of ANNs. Future directions of ANN applications and current salient issues in psychological assessment are discussed.

Artificial Neural Networks for Psychological Assessment

Background

Since the first article on neural network modeling (McCulloch & Pitts, 1943) appeared over half a century ago, continuing fascination with this technique appears to be related to its historical linkage to artificial intelligence, in that ANNs mimic neural activities of the human brain (Cheng & Titterington, 1994). Neural network modeling has been used for various tasks, including speech recognition, stock market prediction, mine detection, cancerous cell identification, and handwriting recognition. Current applications of ANNs usually fall into three areas: (a) in modeling biological nervous systems, (b) as real-time adaptive signal processors or controllers implemented in hardware, and (c) as data-analytic methods (Sarle, 1994). The applications described in this article fall into the third area.

With increased communication between the fields of neural network modeling and statistics, areas of overlap have become better defined. Today, the ANN can be characterized as a highly flexible nonlinear regression technique. Limitations of linear models have been shown in the “exclusive-or (XOR)” problem (Rumelhart & McClelland, 1986). The simplest case consists of two vectors: An input vector consisting of (0, 0) and (1, 1) belongs to Class 1, and another input vector of (0, 0) and (0, 1) belongs to Class 2. There is no single linear decision boundary that can correctly classify all four points. This problem can be generalized to higher dimensions. Thus, an ANN’s ability to transform nonlinear input covariances to linearly separable ones via activation functions supposedly gives the ANN an edge over the linear modeling technique in some situations.

In practice, however, reports of the ability of ANNs to surpass linear models for classification and prediction problems have been inconsistent (Anderer, Saletu, Klöpfl, Semlitsch, & Werner, 1994; Baxt, 1995; Doig, Inman, Sibbald, Martin, & Robertson, 1993; Duh, Walker, Pagano, & Kronlund, 1998; Dybowski, Weller, Chang, & Gant, 1996). Attempts to examine weight matrices, analogous to regression coefficients in linear models such as logistic regression, are a logical step in identifying reasons for the inconsistent performance of ANNs relative to linear models (Duh, Walker, & Ayman, 1998). Nonetheless, examination of individual weights becomes infeasible for a complex network model because the number and dimension of weight matrices increase with the complexity of models.

ANN performance associated with variations in complexity is rarely examined in the context of comparison with linear models. Our results from two experiments relate to the differences in ANN performance with varying numbers of neurons, types of activations, and input variables. We suggest that better prediction of ANN performance depends in part on the complexity of models relative to the complexity of predictive input variables.

Multilayer Perceptron Model of Artificial Neural Networks

Of several types of ANN models amenable to data-analytic applications, MLP modeling has been used most often in clinical medicine and the psychological literature. An MLP model consists of multiple layers of interconnected neurons. Figure 1 illustrates the iterative process of MLP training for two-layer (or single hidden-layer) models, used in two application examples described later. In this instance, the two layers include hidden and output layers, because input variables are not usually counted as a layer. A hidden layer is used for internal transformation of the ANN: Hidden neurons transform the “signals” from the input layer and generate signals to be sent to the next layer, so that the MLP model is sometimes called the “feedforward” neural network.

The input layer consists of input variables plus an input bias, which adjusts the weighted mean of the input variables. All input variables are regressed on the next layer, in the same fashion as estimation of multiple regression. Thus, if the model includes only a single output without a hidden layer, the ANN becomes a logistic regression with a logistic activation function; it becomes a linear discriminant function with a threshold activation function at output. However, the activation functions of the hidden neurons between the input variables and the output layer make the ANN a highly flexible model. Any function can be approximated with increasing numbers of hidden neurons using the sigmoid (logistic) activation function (Bishop, 1995).

Once the predicted values are forwarded from the hidden layer to the output layer, the error amount is computed on the basis of the target (actual) value of outputs. Weights are recalculated after each observation until all observations are processed.

For a more general multiple-hidden-neuron, multiple-hidden-layer model, the output for pattern (observation) \( p \) at the \( j \)th neuron on layer \( l \) can be expressed formally as

\[
\mathbf{x}_{ij} = f(\mathbf{w}_{ij}) + \mathbf{B}_{i} \cdot \mathbf{x}_{i-1}
\]

where

\[
x_i \in \{0, 1\}
\]

\[
\mathbf{w}_{ij} \in \mathbb{R}
\]

\[
\mathbf{B}_{i} \in \mathbb{R}^{n \times n}
\]

3 MLP models are considered to be parametric, quasi-parametric, or nonparametric (infinitely parametric, to be more precise) depending on the complexity of the model. A reference point, the generalized additive model (Ilastic & Tibshirani, 1990) can be placed somewhere between parametric linear regression and fully nonparametric nonlinear regression such as MLP, because the generalized additive model uses a nonlinear transformation estimated by a nonparametric smoother (Cheng & Titterington, 1994; Sarle, 1994, White, 1992).

Because both the ANN and logistic regression use an iterative maximum-likelihood method to estimate parameters, the result of an ANN with a single output and the result of multiple logistic regression should be identical. On the other hand, although an ANN with a single threshold activation function without a hidden layer uses the same model as a linear discriminant analysis, the two methods would not yield exactly the same results because the parameters in a linear discriminant analysis are computed by ordinary least squares and an ANN uses a maximum-likelihood method.
where the previous layer, \( l - 1 \), is input to a node in layer \( l \) through \( v_{ij}^l \):

\[
v_{ij}^l = \sum_{i=1}^{n} w_{ij}^l x_{ij}^{l-1} + B_{ij}^l,
\]

(2)

where \( w_{ij}^l \) is the weight (coefficient) from node \( i \) on layer \( l - 1 \) to node \( j \) on layer \( l \) and \( x_{ij}^{l-1} \) is node \( i \) on the previous layer \( l - 1 \), which can be an input variable or an output from the previous layer. \( B_{ij}^l \), the output bias, can be added to adjust the decision threshold of the predicted outputs to evaluate MLP performance free of thresholds. \( v_{ij}^l \), the input bias, can be used to adjust the weighted mean of the input variables. These biases are not usually applied to hidden layers.

In our applications, two activation functions are used for \( f(v_{ij}^l) \). The sigmoid (logistic) activation function is most commonly used in MLP modeling. This activation function takes the form

\[
f_{sig}(v_{ij}^l) = \frac{1}{1 + e^{-\gamma v_{ij}^l}}.
\]

(3)

The generalized Gaussian activation function is expressed as

\[
f_{gen}(v_{ij}^l) = e^{-\gamma^2 v_{ij}^l}.\]

(4)

where the right-hand side is a generalized activation function of the Gaussian distribution. Additional parameters can be added to change the shape of these functions.

The two functions are plotted along the axis of \( v_{ij} \) in Figure 2. Although the sigmoid activation function is more commonly used, the Gaussian activation function is more suitable for examining a finer range of the distribution of an input vector, as the shape of this function suggests. In contrast, a sigmoid neuron “turns on” and stays “on” once passing a threshold value. The maximum value of the activation function can be adjusted to scale the distribution range of the output vector. For classification and prediction problems involving dichotomous outcomes, the maximum value is set to 1.

Our MLP training used the “back-propagation” algorithm in the “on line” mode, in which the weights are updated after processing each pattern (Rumelhart, Hinton, & Williams, 1986). In this algorithm, small initial weights are randomly assigned to the input neurons and are fed forward through the network. The error of the estimate with respect to the target at the \( p \)th output for the \( p \)th pattern is

\[
F_{ij} = \frac{1}{2} (t_{ij} - y_{ij})^2.
\]

(5)

The total error at the output layer is

\[
E = \sum_{p} W_{p} \sum_{ij} E_{ij}.
\]

(6)

It sums errors over all outputs and over all patterns. \( W_{p} \) is the optional pattern weight to adjust the error function to give weighting to individual patterns. Pattern weighting was introduced because the minimization function can be problematic for maximizing prediction when using population survey data in which a prevalence estimate is very low. The pattern weight is typically set as the inverse of the prevalence, and \( W_{p} = 1 \) when pattern weighting is not added. In practice, we found pattern weights to be more helpful in increasing the efficiency of estimation than in improving prediction.

The information on the total error is used to adjust the weights in the output layer, which are propagated back to the hidden layer to adjust weights on the hidden layer. These changes are accomplished with the use of the generalized delta rule, which employs a gradient descent on parameter space. Each epoch (iteration) of back-propagation continues until the error is minimized globally.

Specification of a few other parameters is usually necessary for the back-propagation algorithm. The learning rate is a step size
taken in the downward direction of the gradient of the error function, \( E \), with respect to a parameter. The momentum smooths out the changes in parameters such as weights (coefficients) to avoid short-term error oscillations during learning and to speed up convergence.\(^2\) These parameters are concerned with convergence speed or assurance of the global minima. The other basics of MLP modeling are found in neural network textbooks (Anderson, 1991; Bishop, 1995; Ripley, 1996; Rumelhart & McClelland, 1986).

**Applications to Psychiatric and Psychological Research**

In early 1995, we conducted a MEDLINE literature search on neural networks. We identified, cross-sectionally, over 1,200 entries for current articles in engineering, computer science, medicine, psychology, and neuroscience and methodological articles relating to statistics and new techniques. This search produced about 30 entries on topics in medicine and psychology.

Ten of the articles were on psychiatric or psychological topics. They can be divided into two groups. (a) conceptual or simulation models that use neural network models as a frame of reference to describe the complexity of psychological phenomena and (b) applications of ANN models to improve the outcome of interest. The topics in the first group included a neural network model of cortical processing to provide a framework for understanding pathological processes in schizophrenia (Chen, 1994); neural network simulation of psychiatric symptom recognition to illustrate the deficiency of traditional algorithmic models of psychiatric diagnoses (Bertin & Chen, 1993); neural network simulation illustrating the structure of the relationships among situation, behavior, and childhood experience in order to predict personality traits (Gonzalez-Irbydrich, 1993); a mathematical approach to hypnosis that connects general ideas about the organization of the nervous system and neural network models (Kuzin, 1995); and a review of the relation between dream construction and memory using the neural network as a model architecture (Palombo, 1992).

In this literature, a neural network model is seen as a machine analogue of the biological neural network system.

In the literature in the second group, neural network modeling was used as an analytical tool. This group included use of ANN modeling as a diagnostic test to predict the admission decision among psychiatric emergency patients (Somoza & Somoza, 1993); a comparison of treatment decisions made by ANNs and clinicians for depressed and psychotic patients (Modai, Stoler, Inbar-Saban, & Saban, 1993); use of an ANN to predict length of hospital stay given psychiatric diagnoses, demographics, illness severity, and other predictors (Davis, Walter, & Davis, 1993; Lowell & Davis, 1994); and an application of Bayesian decision models to an ANN in order to predict teenagers’ substance use (Wu & Gustafson, 1994).

About half of the articles in medicine and psychology were published by investigators in institutions outside the United States. Most noticeably, fewer than 40% of the investigators attempted to compare or incorporate neural network models with standard statistical methods, which supports the view that there is a need for increased communication between statisticians and neural network researchers (Ripley, 1994). Comparison with, or incorporation of, multiple logistic regression occurred in three studies (Doig et al., 1993; Lett et al., 1994; Spackman, 1992). Comparison with other methods, such as the Z statistic and discriminant analysis, was proposed (Anderer et al., 1994). Simple ROC methods were applied to compare the performance of an ANN and other methods in two studies (Lienard, Andersen, & Andersen, 1994; Spackman, 1992).

A more recent review of literature suggests that the use of ANNs in psychiatric and psychological research has been increasing in the last few years. An accelerating diffusion into diverse disciplines is indicated by the fact that articles on the use of ANNs are appearing in highly regarded nonmedical journals. For example, an introduction to ANNs by Baxt (1995) and an ANN application that used a large database collected on intensive-care-unit patients (Dybowski et al., 1996) appeared in The Lancet. Studies that used ANNs have also appeared in widely circulated journals in epidemiology (Ioannidis, McQuen, Gueder, & Kaslow, 1998), genetics (Lucek & Ott, 1997), psychiatry (Zou et al., 1996), and statistics (Warner & Mixa, 1996). In those articles, the results were often compared with results obtained through commonly used statistical methods (Dybowski et al., 1996; Ioannidis et al., 1998).
ANNs have been applied to studies involving standardized psychological assessment instruments. For example, an ANN was used to examine how varying the values of each of the four Stage of Change (SOC) scales (Prochaska & DiClemente, 1982) affected anxiety outcomes, thus allowing prediction of the amount of changes in anxiety outcomes given a certain level of change in an SOC scale (Reid, Nair, Misry, & Beinmann, 1996). Measures from the Diagnostic Interview for Children and Adolescents (Herjanic & Reich, 1982) were input to an ANN to predict adolescent hopelessness (Khashani, Nair, Rao, Nair, & Reid, 1996). The Composite International Diagnostic Interview (Robins et al., 1988) was analyzed with an ANN to compare ANN-derived diagnoses with those derived from computer algorithms based on the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III; American Psychiatric Association, 1980); clinician-based diagnoses were used as the gold standard (Zhu et al., 1996). Overall, however, we found that ANNs are still rarely used in psychological assessment research.

Empirical Examples

In this section, using two large-scale psychiatric surveys, we describe applications of ANNs in two problems in psychological assessment: (a) whether an ANN is able to “mimic” a computer diagnostic algorithm that translates the logic of the clinician diagnostic process and (b) whether an ANN is able to predict a behavioral outcome with various self-report measures expected to contain a level of measurement error. To assess the utility of ANNs, we also examined (c) whether ANNs outperform common statistical methods and (d) whether the patterns of performance improvement differ depending on architecture, initialization, and complexity of input variables.

Method

Data Sets

Experiment 1: St. Louis Epidemiologic Catchment Area (ECA) Project. The ECA project is one of the largest epidemiological psychiatric studies ever conducted in the United States (Robins & Regier, 1991). Survey respondents were persons 18 years of age or older residing in five catchment areas of U.S. communities. For each catchment area, both community and institutional samples were obtained with complex multistage sampling methods, so that the sample represented the demographic distribution of the U.S. population according to the 1980 census with proper sampling weights (Eaton & Kessler, 1985). For this article, the household sample of the data set from St. Louis was used (N = 2,809 after elimination of missing values).

The ECA data set was chosen for several reasons: the advantage of using general-population surveys for drawing inferences that can be generalized, the importance of having a large sample when prevalence (for a disorder is low, and, finally, familiarity with the data set (Price, Robins, Helzer, & Cottler, 1998). The diagnostic variables were computed on the basis of the DSM-III (American Psychiatric Association, 1980), because the surveys were conducted between 1981 and 1983. Psychiatric nosology has subsequently changed with the publication of the revised DSM-III (DSM-III-R; American Psychiatric Association, 1987) and the fourth edition (DSM-IV; American Psychiatric Association, 1994). It should be noted that the appropriateness of a specific nosological system is a peripheral issue in this article because the ANN was tested to mimic the clinical decision process given a specific nosological system; in this case, the DSM-III.

Experiment 2: Vietnam Era Study—Phase III. For the second set of experiments, we used data from the Washington University Vietnam Era Study—Phase III (VES). The previous surveys were conducted in 1972 and 1974. About half of the veterans in the target sample were drawn randomly from the list of Army enlisted (EI-E9) returnees whose urine tested positive for opiates, amphetamines, or barbiturates at the time of departure from Vietnam in September 1971. The drug-positive veteran population comprised an estimated 10.5% of the total 13,760 Army enlisted returnees who departed at that time. The other half, the general sample, comprised veterans who were randomly drawn from the same total population and included 4.1% of the sample already drawn from the drug-positive list (Robins, 1974). A civilian control sample was ascertained from Selective Service registrations and individually matched to the general sample members interviewed in 1974 for draft eligibility, draft board location, age, and education (Robins & Helzer, 1975). The cohort, although not a general-population sample, was ascertainment from a large nonclinical population, and veterans arguably can be considered a general-population sample weighted to slightly lower socioeconomic status (O’Donnell, Voss, Clayton, Slatin, & Room, 1976).

As part of the feasibility phase of the 25-year follow-up study in 1996, mortality information was collected. Further details of the follow-up study are described elsewhere (Price, Risk, & Spirt, 1999; Price, Risk, Stawley, et al., 1999). The sample for Experiment 2 was composed of veterans and civilians who were interviewed in 1974 (N = 855).

Measures

Experiment 1. Using data from the St. Louis ECA study, we assessed the judgment ability of ANNs in the absence of information on DSM-III rules for the diagnosis of antisocial personality disorder (ASPD). The input variables were measures of 12 childhood and 9 adult DSM-III symptom criteria for ASPD obtained from the Diagnostic Interview Schedule, a well-validated structured instrument designed to be administered by non-clinicians (Helzer et al., 1985; Robins, Helzer, Coughran, & Ratcliff, 1981). The childhood symptoms included immaturity, school expulsion, arrest, running away from home, lying, sex before age 15, getting drunk or engaging in illicit drug use before age 15, stealing, vandalism, poor grades, serious trouble at school, and starting fights. The adult symptoms included job troubles, neglectful toward children: a nontraffic arrest for prostitution, punching, or drug sale: marital or intimate relationship problems; violence: trouble with drugs; vagrancy: lying; and traffic offenses. These measures have been shown to correlate highly with each other (e.g., Iether, 1988).

The output variable was the ASPD diagnosis. Validity studies reported kappa's of .63 for test-retest agreement (Robins, Helzer, Ratcliff, & Seyfried, 1982) and .62 for agreement between a psychiatrist and a trained nonclinician interviewer (Robins et al., 1981). The exclusion criteria of ASPD—that is, mental retardation, schizophrenia, or mania—were not applied to the operational diagnostic measure of ASPD because the information for the exclusion criteria is external to the information contained in the childhood and adult symptoms.

Experiment 2. For the VES, the output variable was mortality. That is, the judgment task was to predict which veterans and civilians would be dead at the time of the follow-up study more than two decades after their return from Vietnam. It was worthwhile to test the ability of an ANN with a clearly defined behavioral measure that avoids fuzzy classifications typical in psychological assessment. We do not claim that our measure of mortality is perfect; nonetheless, we felt it was relatively free from classification errors.

By the end of December 1993, 97 of the 1,227 (7.9%) participants in the original study had been identified as deceased (Price, Et. al., 1999). Three sources of death information were used, which were later augmented by information from family members in order to maximize accuracy for case identification (Price et al., 1999). We input two sets of predictors to the ANN to see if differences in measures related to the ANN’s performance relative to that of logistic models.

The first set of input variables consisted of 23 significant predictors chosen separately by logistic regression for each of three periods—pre-
Vietnam, in-Vietnam, and post-Vietnam up to 1974 — plus lifetime questions and demographics. These variables were selected from over 120 variables covering 11 domains: demographics, socioeconomic status, military experience, family history, psychiatric problems, antisocial behavior, crime, social networks, drug use, alcohol use, and a residual category that were found to be associated with mortality by bivariate analyses. The input variables were as follows: (a) pre-Vietnam — heroin use, regular narcotics use, narcotics injection, types of drugs used, marijuana use level, too much drinking, and childhood antisocial behavior; (b) in-Vietnam — narcotics withdrawal, stimulant use by heavy drug use, and arrest; and (c) post-Vietnam up to 1974 — wanting to take narcotics, feeling addicted to narcotics, knowing where to buy heroin, taking drugs to relieve effects of other drugs, taking drugs to help depression, depression while drinking, arrest, and number of arrests. The lifetime behavioral predictors included heroin use level and use of opiates five or more times. Two demographic variables, African American race and veteran status, were also added. These variables were extensively analyzed previously (Robins, Helzer, & Davis, 1975), and findings from the earlier surveys continue to be cited (U.S. General Accounting Office, 1998).

The second set of input variables included 8 variables that were significant at the .05 probability level when the previous 22 variables were input to logistic regression with backward elimination. They included pre-Vietnam heroin use, narcotics injection, and too much drinking; in-Vietnam narcotics withdrawal; and post-Vietnam wanting to take narcotics, knowing where to buy heroin, and using drugs to help depression. African American race was also included. We hypothesized that an ANN would improve prediction when a number of highly collinear variables were input whereas the level of prediction with a logistic model would stay about the same because a linear regression technique does not gain much additional information from highly collinear variables.

 Scaling. The input variables in Experiment 1 with the ECA data set were all scaled to be dichotomous. Five of the input variables in Experiment 2 with the VES data set were not dichotomous: types of drugs used (0–4), marijuana use level (0–3), childhood antisocial behavior (0–10), number of arrests (0–4), and heroin use (0–5). These variables were scaled to vary between 0 and 1, a standard practice used to increase efficiency of estimation in MLP modeling. The same scaling was applied to the input variables of logistic regression for consistency.

 Analyses

Experiment 1. Two-output models were used because they provide a simple solution to the threshold decision in a classification problem since the difference in predictive values is used for assigning the class (Bishop, 1995). Both sigmoid and Gaussian activation functions were tried. The range of a hidden neuron varied between 0 and 1, which is a default for a classification problem. The number of hidden neurons was increased by an increment of one from 1 to 5, then to 7, to 10, and thereafter by an increment of five to a maximum of 30. After an initial series of experiments, the learning rate was set to .001 and the momentum to .01. Both are considered to be conservative, which is to say that convergence is achieved slowly. The maximum iteration was set to 600 after initial analyses, which used as many as 1,200 iterations. Estimation was accomplished both with and without pattern weighting.

The results of MLP neural network models were compared to the results of linear and quadratic discriminant analyses. We carried out analysis of suboptimal network performance both for sigmoid and Gaussian neurons with two different initializations in order to examine whether complex MLP architecture was required to improve prediction.

Experiment 2. Two-output models were also used for this experiment. The same set of activation functions and learning parameters were tried. The maximum number of iterations was set to 600 after initial trials. Pattern weights were used throughout for this experiment because the point prevalence was low at .079. Logistic regression analyses were used to compare the results because beta estimates illustrate a limitation of linear models when predictive collinear variables are used. ¹ We also examined suboptimal network performance to assess whether the improvement in prediction using the full set of predictors was related to the complexity of the MLP architecture.

Software. Numerous software packages for ANNs are available. We identified 52 free and 25 commercial packages at the time these experiments were planned (Surle, 1996). After a trial period we chose UNIX-based Partek, written by Thomas J. Diuere and Donald J. Meyer (Partek Inc., 1998). We also applied an ANN to the MCA data set using source code modified from the "Neural Network Classes for the Next Computer," which is available from the University of Arizona. The results were judged to be similar, although this source code used a one-output model and Partek used two-output models.

Cross-validation. Because real future data are lacking in practice, cross-validation is an essential part of the ANN methodology, because it allows assessment of the generalizability of ANN results obtained from the training phase. Tenfold validation was used: 90% of the data were used to train the ANN, and the remaining 10% of the data were used to compute the numbers of positive and negative errors. Results of each validation were obtained from the predicted outcome for each observation based on estimates obtained from the training phase. The process was repeated 10 times, and the results were summed over 10 validation results to compute the sensitivity and specificity scores. We cross-validated corresponding linear models using the same procedure.

We selected tenfold validation, a typical choice in ANN modeling (Bishop, 1995). Tenfold validation was suitable for our problems because the outcome prevalence rates of both data sets were low (11.6% for ASPD without exclusion criteria and 19% for mortality). Therefore, we wanted to use as much information as possible for the positive cases for training without increasing the processing time too much.

Evaluation by ROC analysis. The ANN results were evaluated with a standard ROC method. Although it was originally invented as an engineering technique for radar detection, ROC analysis has been subsequently developed in the medical decision theory and signal detection literatures (Swets & Pickett, 1982; Weinstein & Fineberg, 1980). ROC analysis is a simple but useful tool for evaluating the predictive utility of a test. It combines information on both true positives and false negatives by plotting sensitivity and (1 specificity), therefore, ROC analysis is free of the problems of cutoffs (Dwyer, 1996). For assessing the overall predictive utility of measures, the area under the ROC curve (AUC) is commonly used, which ranges between .5 and 1.0. When prediction is made from self-reported behaviors, the AUC rarely extends beyond .5 even with established risk factors. For example, the AUC value was in the .5 to .8 range in our past analysis predicting adult ASPD from childhood conduct problems (Robins & Price, 1991). This range is similar to that of the AUC for a blood glucose test; earlier in the prediction of diabetes (Kendrick & Leve et al., 1981). We used the value of the complement of the area under the curve (CAUC), that is, 1 AUC, as an overall measure of predictive utility. Differences in prediction are easier to see with the use of CAUC when the prediction range is very good to excellent.

The ROCs for MLP models were obtained with the use of an output bias. The decision threshold was varied to obtain 101 coordinates of the summary sensitivity and (1 specificity); the ROCs for discriminant analyses

¹ Although it might have been preferable to use Cox regression, which estimates the impact of predictors on the proportional hazard rate, we chose logistic regression because initial analyses indicated that estimates obtained from Cox regression were nearly identical to those obtained from the logistic regression analysis, which is often the case with a low-prevalence dependent variable. Therefore, we used logistic regression estimated for ease of interpretation.

³ Because of time constraints, we did not use the leave-one-out cross-validation, which repeats validation for the same number of times as the sample size by using all observations but one to train the ANN and by testing on the one left.
were obtained by varying the threshold of the canonical discriminant function (Ripley, 1996); and the ROC for logistic regression was obtained by varying the output threshold probability level.

Results

Experiment 1

Figure 3 shows the results of MLP modeling in an ANN compared with the results of linear and quadratic discriminant analyses in the assessment of the diagnosis of ASPD. The best-fit MLP model was obtained with 30 Gaussian hidden neurons without pattern weights at the 600-iteration stop point. This model outperformed the best estimate of MLP with pattern weights as well as those of the linear and quadratic discriminant analyses. The CAUC was less than .001 for the MLP model, .023 for the linear discriminant analysis, and .083 for the quadratic discriminant analysis. The high level of diagnostic accuracy achieved by all three methods is not surprising because the information consisted of the DSM-III symptom criteria. However, it is noteworthy that the MLP model yielded a better result than the linear ratings of the symptoms.

Figure 4 plots the performance of suboptimal networks for both sigmoid and Gaussian hidden neurons. We chose two sets of randomly assigned initial weights for each type of activation function to assess whether the initial weights also have an impact on performance. Different initial weights resulted in some differences in the performance level within the range in which the CAUC was improving rapidly. When the number of hidden neurons was very small or very large, the choice of initial weights made little difference in the CAUC.

With the Gaussian activation, the MLP model’s performance improved rapidly with a small number of hidden neurons. With four neurons, the CAUC decreased to less than .01 and thereafter improved very slowly with larger numbers of hidden neurons. On the other hand, network performance with the sigmoid activation function was more affected by the number of hidden neurons used. The CAUC did not improve to the same level as for the Gaussian models until the number of neurons was increased to the 10-15 range.

Experiment 2

In Figure 5, MLP performance results were compared with results of logistic regressions in predicting the 25-year mortality outcome for two sets of input variables. The smallest CAUC value of .04 was achieved with 40 Gaussian hidden neurons and the full 23 variables. Given that the behavioral data were obtained two decades ago, we consider this level of the CAUC as excellent. The parallel logistic regression produced a much worse result (CAUC = .21) with the same 23 predictors. However, when only eight variables were used, the MLP performance was about the same as that of logistic regression, yielding a CAUC value of .19. The gain in the CAUC value by the MLP model with the 23 variables is primarily attributable to increased sensitivity; that is, the MLP model did much better in predicting deaths correctly.

As shown in Figure 6, the network performance results were different depending on whether 23 or 8 input variables were used. When all 23 variables were used, the increase in the number of hidden neurons yielded smaller CAUC values for both sigmoid and Gaussian activation functions. However, with the 8 input variables, neither the sigmoid nor the Gaussian models were able to reduce the CAUC much below a level of .19, and the minimum was achieved with only five sigmoid neurons. Beyond this level of complexity, the CAUC obtained with sigmoid hidden neurons essentially stayed flat, whereas the CAUC obtained with Gaussian neurons increased, which is likely to be a sign of overfitting. These results indicate that complex network architecture did not help when the number of predictive input variables was limited.

The predictors in the logistic regression without cross-validation are shown in Table 1. The 8 most significant variables chosen by the logistic regression were all dichotomous when the 23 mixed dichotomous and scaled variables were input for selection with backward elimination. The 23 variables included several clusters of collinear variables involving heroin use, narcotics injection, types of drugs used, marijuana use level, narcotics withdrawal, felt

---

* It would have been possible to achieve a slightly better fit with a higher number of neurons or of iterations. However, we judged further improvement was unnecessary for the purposes of this experiment given the small level of the CAUC achieved by the MLP model.

[1] No method exists to our knowledge for adjusting beta estimates obtained from multifold validation. Therefore, the logistic regression in Table 1 used 100% of the data, compared with 90% in a tenfold estimation.
addiction to narcotics, drug use to relieve the effects of other drugs, drug use to help depression, depression while drinking, arrest, number of arrests, heroin use level, and lifetime use of opiates five or more times. Odds ratios transformed from the beta estimates involving these 13 variables, therefore, are biased with respect to the magnitude of the predictive effects. Also, effects of 7 variables yielded odds ratios less than 1, erroneously indicating that they were protective factors. A usual treatment would be to remove highly collinear variables to make the results of odds ratios meaningful. Without such a treatment, odds ratios can result in misleading to outright wrong interpretations. Furthermore, prediction with logistic regression suffers from having these collinear variables, as shown by the higher CAUC with 23 variables than with 8 variables.

Discussion

Findings Summary

The results of the two experiments in using ANNs in this article were related in (a) improving judgment and decision making in the area of psychological assessment, (b) comparing ANN performance with the performance of linear statistical methods, and (c) delineating some of the reasons for the superiority of ANNs over linear methods. The ANN outperformed linear and quadratic discriminant analyses in a psychiatric diagnosis assessment when a set of clinical measures was provided a priori (the ECA example). It also made a substantial improvement over logistic regression analyses in predicting cumulative mortality from measures collected two decades earlier when the input variables consisted of a large number of measures of dichotomous and ordinal measures known to be associated with the outcome. When the input vari-

Figure 4. Multilayer perceptron network performance predicting antisocial personality disorder using sigmoid versus Gaussian neurons with different initializations. The sample corresponds to that used in Figure 3, and the same input variables were used. The number of neurons plotted in this figure varied between 1, 2, 3, 4, 5, 7, 10, 15, and 20 (the complement of the area under the curve [CAUC] almost reached a plateau at 20). Two sets of initial weights were randomly chosen for both sigmoid and Gaussian activation functions.

Figure 5. Comparison of 22-year mortality prediction between the MLP model and logistic regression (N = 855). The sample was a subset of the Vietnam Era Study data set, which consisted of veterans and nonveterans interviewed in the 1974 survey. All 23 variables listed in Table 1 were used, and this model was compared with a reduced model with 8 variables selected by backward elimination in a logistic regression. The receiver operating characteristic (ROC) curve for the MLP full model was obtained with 40 Gaussian hidden neurons; the ROC curve for the MLP reduced model was obtained with 5 sigmoid hidden neurons. Both used pattern weights. All ROC curves used tenfold cross-validation.
ables were limited to a smaller set of measures considered most significant by logistic regression, MLP modeling did only as well as logistic regression (the VES example).

Examination of network performance yielded several observations. Initial weights were found not to be critical if the number of hidden neurons was sufficiently large. The results indicate that the types of hidden neurons seem to affect the level of prediction. With the ECA data set, we found differences in a midsize range of network complexity: Gaussian models reached the low plateau of the CAUC with fewer neurons than did models with sigmoid neurons. A similar trend was observed with the VES data set; however, when the input variables were limited, Gaussian models risked overfitting even with a smaller number of hidden neurons.

There appear to be no solid rules on the number of hidden neurons other than a commonsense understanding that the number of response patterns is one factor (Zurada, 1992). Nevertheless, because of the Gaussian model’s purported sensitivity for finding a threshold in a finer region, better performance was expected with a smaller number of Gaussian hidden neurons. With increased numbers of hidden neurons, however, the distinction between sigmoid and Gaussian models was expected to blur.

Our second experiment suggests that the number and the nature of input variables are particularly important for MLP models to be able to improve prediction over linear models. The lack of improvement in the CAUC for the MLP model using the eight variables may be due in part to the fact that all of these eight variables were dichotomous (Warner & Misra, 1996). The information for potentially nonlinear association with the underlying distribution is diminished with the use of dichotomous variables. In the logistic analyses, highly correlated input variables hurt more than helped estimation and prediction. However, collinear variables have the potential to provide additional information to improve prediction in MLP models by transforming a linearly non-separable input covariate space to a potentially linearly separable “image” space in hidden layers (Zurada, 1992), which may be a reason for the much improved prediction by the MLP model with the 23 variables.

Future Directions for ANNs and Associated Methods

Our data analyses using ANNs could be improved in several ways. With respect to ANNs, a variety of algorithms are now available in addition to the back-propagation algorithm, which is known to be slow. These newer algorithms use the information from second derivatives and therefore reach convergence faster and require fewer learning parameters. For example, the learning rate and momentum parameters we used in our experiments are not needed with the use of newer algorithms (Bishop, 1995; Ripley, 1996). Employing these newer algorithms with the increasingly powerful hardware available is likely to remedy the current computational inefficiency associated with ANNs.

Our method in neural network estimation did not directly minimize errors based on the values of the CAUC, because its derivatives with respect to weights could not be mathematically obtained. It may be possible to explore alternative minimization functions in neural networks with the use of ROC analysis by linear transformation of the ROC or by using equivalent estimates such as Wilcoxon or C statistics (Hanley & McNeil, 1982). Alternatively, obtaining the derivatives numerically may become a feasible option, as is already done with nonlinear regression algorithms (Ralston & Jennrich, 1978).
Table 1

Estimates of Logistic Regression Coefficients With All 23 Variables Versus 8 Most Significant Variables (N = 855)

<table>
<thead>
<tr>
<th>Variable</th>
<th>All 23 variables</th>
<th>Only 8 most significant variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>p</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American (race)</td>
<td>3.27</td>
<td>.002</td>
</tr>
<tr>
<td>Veteran</td>
<td>2.00</td>
<td>.36</td>
</tr>
<tr>
<td>Pre-Vietnam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heroin use</td>
<td>0.05</td>
<td>.005</td>
</tr>
<tr>
<td>Regular narcotics use</td>
<td>1.29</td>
<td>.72</td>
</tr>
<tr>
<td>Narcotics injection</td>
<td>24.72</td>
<td>.0009</td>
</tr>
<tr>
<td>Types of drugs used (0–4)</td>
<td>0.87</td>
<td>.86</td>
</tr>
<tr>
<td>Marijuana use level (0–3)</td>
<td>2.19</td>
<td>.35</td>
</tr>
<tr>
<td>Too much drinking</td>
<td>2.46</td>
<td>.03</td>
</tr>
<tr>
<td>Childhood antisocial behavior (0–10)</td>
<td>2.83</td>
<td>.22</td>
</tr>
<tr>
<td>In-Vietnam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narcotics withdrawal</td>
<td>3.04</td>
<td>.13</td>
</tr>
<tr>
<td>Stimulant use</td>
<td>1.87</td>
<td>.18</td>
</tr>
<tr>
<td>Heavy drug use</td>
<td>1.95</td>
<td>.44</td>
</tr>
<tr>
<td>Arrest</td>
<td>1.59</td>
<td>.37</td>
</tr>
<tr>
<td>Post-Vietnam</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wanted to take narcotics</td>
<td>4.32</td>
<td>.003</td>
</tr>
<tr>
<td>Felt addicted to narcotics</td>
<td>2.03</td>
<td>.21</td>
</tr>
<tr>
<td>Know where to buy heroin</td>
<td>1.64</td>
<td>.19</td>
</tr>
<tr>
<td>Drug use to relieve effects of other drugs</td>
<td>0.53</td>
<td>.21</td>
</tr>
<tr>
<td>Drug use to help depression</td>
<td>2.76</td>
<td>.02</td>
</tr>
<tr>
<td>Depression while drinking</td>
<td>0.50</td>
<td>.20</td>
</tr>
<tr>
<td>Arrest</td>
<td>2.03</td>
<td>.21</td>
</tr>
<tr>
<td>Number of arrests (0–4)</td>
<td>0.26</td>
<td>.22</td>
</tr>
<tr>
<td>Lifetime</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heroin use level (0–3)</td>
<td>0.78</td>
<td>.78</td>
</tr>
<tr>
<td>Opium use five or more times</td>
<td>0.66</td>
<td>.58</td>
</tr>
</tbody>
</table>

Note. The two logistic regressions correspond to the two receiver operating characteristic curves shown in Figure 5. Estimates were one-tailed; that is, all available observations were used. OR (odds ratio) = e^β; p = probability level of the beta coefficient. The lifetime assessment refers to "ever" using up to the 1974 survey. All variables are dichotomous except those followed by numbers in parentheses. All ordinal variables were subsequently scaled to be between 0 and 1, to be identical to the scales used in the multilayer perceptron model.

Issues of Psychological Assessment

The results of our experiments demonstrated several advantages of ANNs over linear models for psychological assessment. ANNs are far from problem free, however. Perhaps most important is the "black-box" nature of current ANN applications. As mentioned, one obvious way to open the black box is with a detailed interpretation of weights, moving from a simple to a more complex model. For example, analyses of input and output spaces created from weight matrices can aid in deciding the optimal number of neurons that yield substantive interpretation of data (e.g., Duh et al., 1998). In a more complex model involving a large number of input variables and hidden neurons, it is still possible to assess the relative importance of input variables by taking a summary measure of the weights (Lucek & Ott, 1997). Alternatively, the structure of hidden layers can be explored by using the information stored in vectors of hidden neurons, once training is completed. Principal-components analysis, canonical discriminant analysis (e.g., Wiles & Sloesch, 1992), and cluster analysis (Elman, 1990) have been used to explore the network structure.

Linear methods and ANNs are both sensitive to low prevalence, which is often the case with rare diseases or psychiatric disorders. When the prevalence is very low, ANN does what any smart human would do: It throws all cases into the negative category at the cost of a small number of false negatives. This practice yields excellent specificity and unimpressive sensitivity and could yield unfortunate clinical consequences. For example, a clinician may miss diagnosing 1 depressive who subsequently commits suicide to be more "costly" than prescribing Prozac to 10 dysphoric individuals. We used pattern weights to address this issue, but this strategy was more helpful for improving the efficiency of estimation than prediction. The concern for low prevalence can be substantially diminished if the error function in MLP modeling is derived directly to minimize the CAUC in the ROC method.

Guidelines are needed for the sample size. The literature available on sample size requirements for neural network modeling is limited. In the case of a rare disease or condition, the sample size needed is probably related to the number of positives, so enriched samples or populations will be valuable. It would be possible to obtain the sampling distribution of weights by a bootstrap method to estimate necessary sample size (White, 1989).

Equally important for the purposes of psychological assessment, information is needed on the conditions under which ANNs outperform conventional statistical methods. The examination of network performance and comparisons with logistic analysis in our application provided empirical observations about some likely exploration of fuzzy neural networks, which applies fuzzy set theory to neural network architecture, might prove useful for diagnostic decision making (Fu & Shann, 1994). However, this type of ANN architecture involves multiple hidden layers and is much more complex than the type of ANN we have modeled so far.

One way to describe the validity of a clinical diagnosis is to evaluate how well it predicts longitudinal outcomes. Neural network application to longitudinal measures at this time remains theoretical at best (De Laurentibus & Ravdin, 1994; Liestol et al., 1994). In our example of behavioral outcome prediction, we used logistic regression to predict cumulative mortality. Nonrecurring events, such as death, however, are better handled by survival analysis and Cox regression given their ability to take the timing of events into account.

Neural network analogues to the statistical methods for longitudinal measures have been tested with small data sets. The approach involves (a) starting with a one-layer model with the logistic activation function, which, with back-propagation, should compute the predicted conditional probability of the event, and then (b) extending it to a nonlinear nonproportional hazard model by introducing a hidden layer (Liestol et al., 1994). The ROC methodology can be applied in a more sophisticated fashion. The probit-probit transformation of the ROC (Swets & Pickett, 1982) would yield linear estimation. Thus, regression coefficients can be estimated with Z statistics to test for differences in prediction (Engel & Lee, 1981). The recently developed "two-truth" method (Phipps & Hutton, 1995) appears to improve the estimate of the accuracy of a test when the estimate of the true rates is uncertain. It may provide a solution to the problem of noisy data, frequently found with psychosocial data.
conditions that make application of ANNs suitable. We attempted to provide some information as to how to better model ANNs, such as the number and nature of input variables and performance variation associated with types of activation functions and the number of hidden neurons. The next step will be to conduct simulation studies in which linearity, noise level, prevalence, and sample size are concurrently varied with other parameters of ANN modeling. Such studies are likely to provide information useful to research practitioners for deciding when and how to apply ANN models to their data.

It should be stressed that an ANN is only as good as the data it analyzes. We have had several failed experiments when input variables were limited or were not constructed to be useful for ANN modeling. Greater attention needs to be paid to improving techniques for selecting optimal measures. Newer techniques such as genetic algorithms can be applied to selecting optimal measures (Goldberg, 1989; Downey & Meyer, 1994). Genetic algorithms can be combined with ANN modeling (So & Karpus, 1996) to further improve ANN prediction. Given ongoing refinements to the ANN methodology, we expect that ANNs will gain increased acceptance in psychiatric and psychological research and will be applied more widely in clinical decision making.

References


Loober, R. (1988). Natural histories of conduct problems, delinquency, and


