Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 11426–11430, December 1991
Medical Sciences

Neural networks as a tool for utilizing laboratory information: Comparison with linear discriminant analysis and with classification and regression trees

(parallel processing/pattern matching/clinical chemistry/medical diagnosis/liver diseases)

Gilbert Reibnegger*, Günter Weiss*, Gabriele Werner-Felmayer*, Gerd Judmaier†, and Helmut Wachter*‡

*Institute for Medical Chemistry and Biochemistry, and †Department of Internal Medicine, University of Innsbruck, Fritz Pregl Strasse 3, A-6020 Innsbruck, Austria

Communicated by Leon N Cooper, September 16, 1991 (received for review June 10, 1991)

ABSTRACT **Successful applications of neural network architecture have been described in various fields of science and technology. We have applied one such technique, error back-propagation, to a medical classification problem stemming from clinical chemistry, and we have compared the performance of two different neural networks with results obtained by conventional linear discriminant analysis or by the technique of classification and regression trees. The results obtained by the various models were tested for robustness by jackknife validation ("leave *n* out" method). Compared with the two other techniques, neural networks show a unique ability to detect features hidden in the input data which are not explicitly formulated as input. Thus, neural network techniques appear promising in the field of clinical chemistry, and their application, particularly in situations with complex data structures, should be investigated with more emphasis.**

Although studied long since by several pioneers (1–3), neural networks have attracted a high level of attention in the scientific world only during the past decade (4–6). Nowadays "neural network architecture" represents a serious alternative to conventional "von Neumann architecture" of computing devices (hardware and software), and computational neuroscience is not only of eminent interest for neuroscientists but has penetrated into fields as diverse as speech generation (7), stock price prediction (8), the traveling-salesman problem (9), prediction of secondary protein structures (10), and recognition of signals in a noisy environment (11), to mention just a few.

To investigate the suitability of neural network technique to medical pattern-matching problems which are typical in the process of generating medical diagnoses, we have simulated neural networks designed to differentiate between three distinct disease entities [fatty liver (FL), chronic persistent non-A, non-B hepatitis (CPH), and chronic aggressive non-A, non-B hepatitis (CAH)] on the basis of several characteristic laboratory data. For training of the networks, we have used a real data set from an earlier publication (12). We present a comparison of the resulting neural networks with classical linear discriminant analysis and with the more recent technique of classification and regression trees (13).

METHODS

Data. We have chosen to use a previously published data set on laboratory measurements in patients with mild chronic diseases of the liver (12): briefly, it was shown that the combined measurement of urinary neopterin, an indicator of activation of the cell-mediated immunity, and of the ratio of the concentrations in serum of aspartate aminotransferase (AST) and alanine aminotransferase (ALT) is of use for the differential diagnosis of FL, CPH, and CAH. In particular, neopterin as an immune activation marker differentiates well between FL (low and normal levels) and the two viral disease entities (high levels) but cannot discriminate between CPH and CAH. On the contrary, the AST/ALT ratio is low in CPH and high in both other conditions but cannot be used to differentiate between FL and CAH. Table 1 shows descriptive statistics of the data used in the present study.

Table 1. Laboratory findings on 42 patients with different types of liver disease (adapted from ref. 12)

		Laboratory findings*			
Diagnosis	No.	Neopterin	AST	ALT	AST/ALT
CPH	16	290 ± 82	47 ± 29	94 ± 49	0.49 ± 0.12
CAH	10	346 ± 172	62 ± 51	76 ± 57	0.87 ± 0.34
FL	16	135 ± 54	37 ± 26	45 ± 24	0.98 ± 0.63

*Given are mean values ± SD. Units: neopterin, μmol/mol of creatinine; AST and ALT, units per liter.

Network Formulation and Calculation. For training of the networks, we used error back-propagation (5). As basic network, a three-layered feed-forward network was used consisting of four input units (the units are henceforth called "neurons" for simplicity) receiving the input signals (i.e., appropriately transformed laboratory measurements), four hidden neurons (responsible for internal representations), and three output neurons delivering the output signals (corresponding to the three diagnostic classes). As is evident from Fig. 1, the four input neurons are responsible to receive as "input signals" the actually measured laboratory variables—neopterin, AST, ALT, and AST/ALT ratio. In order to explore the ability of the neural network to extract hidden information from the data, a second network was studied which differed from the one shown in Fig. 1 by omission of the input unit corresponding to the AST/ALT ratio. Thus, this truncated network did not receive the AST/ALT ratio as an explicit input signal.

To adjust the connection strengths between the neurons (i.e., the "learning phase"), the generalized delta rule for gradient descent was used. The details of computations have been described by others in recent papers and monographs; we have specifically used the formulation given in ref. 11. The process can be externally controlled by two parameters, the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; CAH, chronic aggressive non-A, non-B hepatitis; CPH, chronic persistent non-A, non-B hepatitis; FL, fatty liver; CART, classification and regression trees.

‡To whom reprint requests should be addressed.

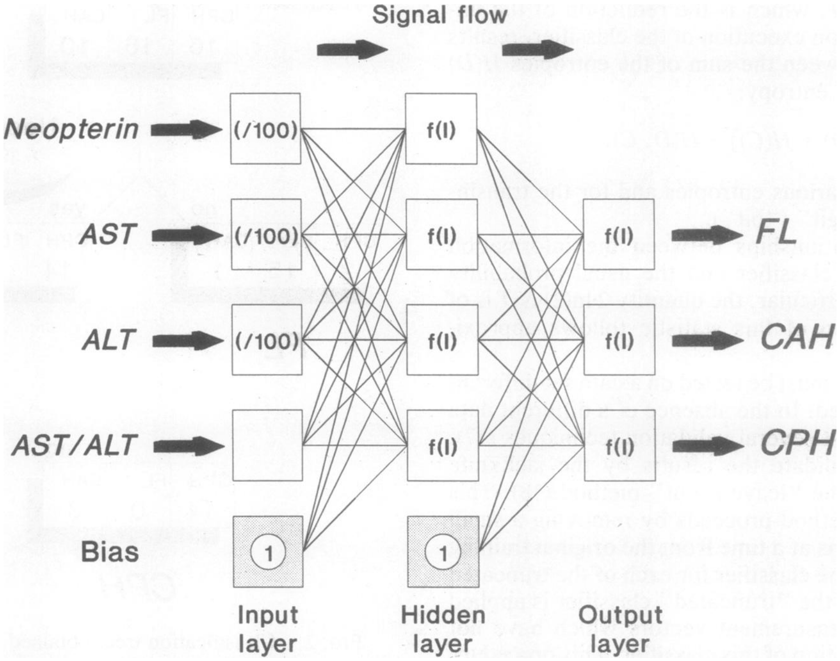

FIG. 1. The basic feed-forward network. The signal flow proceeds from the input layer via the hidden layer to the output layer. The input neurons receive the input signals—i.e., real-valued measurements of laboratory variables neopterin, ALT, AST, and AST/ALT ratio. Neopterin values (in μmol per mol of creatinine) and ALT and AST activity values (in units per liter) are divided by 100 for numerical convenience. The "bias" neurons are always "firing" (i.e., their output is 1.0). Each neuron of the hidden and output layers computes a weighted sum of its incoming signals (I) which is then transformed according to $f(I) = 1/[1 + \exp(-I)]$ to yield the output of this neuron. The weights used for these computations are adjusted during the supervised learning process. The output neurons correspond to three diagnostic categories: FL, CPH, and CAH.

learning rate ϕ and the momentum factor μ. For all network computations reported here, these parameters were set to $\phi = 0.20$ and $\mu = 0.50$. The training set used consisted of complete data records of 42 patients, 16 with FL, 16 with CPH, and 10 with CAH (Table 1). Initially the connection strengths were randomly set.

Evaluation of Results. In order to judge the potential of neural computing in the chosen setting, the same data were analyzed by two additional statistical techniques: linear discriminant analysis (using program BMDP7M from the commercially available BMDP software, University of California Press) and the classification and regression trees (CART) technique (13) were employed for comparison. While linear discriminant analysis is a well-known statistical tool with a long-standing tradition, the CART technique is of more recent origin. Briefly, the method allows one to construct a binary tree structured classifier. The method starts with the whole measurement space itself (which by definition is the matrix containing all measurement vectors) and proceeds by repeated splits of subsets of the measurement space into two descendant subsets. The fundamental idea is to select such splits that the data in the descendant subsets are "purer" with respect to the classification problem at hand; i.e., each subset should contain as many as possible members (measurement vectors) belonging to one certain class and as few as possible members belonging to all remaining classes.

Evidently, each classification procedure (neural network techniques, linear discriminant analysis, and the CART method) eventually produces a classification matrix N_{ij} showing the number of correctly and falsely classified members. Such a matrix gives detailed information about how the resulting classifier performs: the elements n_{ij} are simply the frequencies of data vectors belonging to class i which were classified into class j.

To compare the resulting classifiers statistically, an information theoretical technique was employed (14). This approach is based on the formal analogy between the transmission of information from a transmitter to a receiver, on the one hand, and the perception of a pathological process (e.g., a certain disease class) by an observer (e.g., a physician) on the other. The main idea here is that a certain classifier (which could be a single clinical chemical test or, as in the present work, a more complex classifier such as a trained neural network, a specified binary decision tree, or a set of linear discriminant functions) reduces the initial uncertainty concerning class membership of a measurement vector. This reduction in uncertainty can be expressed, following Shannon's concept (15), by a single value, the so-called transinformation (or "information content"). The following paragraph shows briefly the necessary computations.

The transinformation is calculated via so-called "entropies," which are measures of uncertainty. The input entropy, *H(D)*, which is due to the medical classification problem, is defined as

$$H(D) = \log_2(N) - N^{-1} \cdot \sum_{i=1}^{m} n_i \cdot \log_2 n_i,$$

where the n_i ($i = 1, 2, \ldots, m$) are the numbers of measurement vectors belonging to class i, and N is the total number of measurement vectors (i.e., $N = 42$ in the present example). Thus, this entropy reflects the *a priori* uncertainty (i.e., before the classifier under consideration has been used). Applying the classifier provides the output entropy, *H(C)*, resulting from the j ($j = 1, 2, \ldots, k$) classifier's outcomes:

$$H(C) = \log_2(N) - N^{-1} \cdot \sum_{j=1}^{k} n_j \cdot \log_2 n_j.$$

As a combined measure of the uncertainty of the classification system, the joint entropy *H(D, C)*, which will assume different values in accordance with the mutual dependence of the classification problem and the classifier, is defined:

$$H(D, C) = \log_2(N) - N^{-1} \cdot \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij} \cdot \log_2 n_{ij}.$$

The transinformation, T, which is the reduction of the uncertainty of the system on execution of the classifier, results from the difference between the sum of the entropies $H(D)$ and $H(C)$, and the joint entropy:

$$T = [H(D) + H(C)] - H(D, C).$$

The unit used for the various entropies and for the transinformation is "binary digit" ("bit").

There are close relationships between the information theoretical model of a classifier and the usual probability theoretical model. In particular, the quantity $2 \cdot \ln(2) \cdot N \cdot T$ is of interest; the distribution of this statistic follows approximately a χ^2 (16).

Generally, a classifier must be tested on a data set different from the training set used. In the absence of a different data set, however, there exist several validation techniques (17). We have chosen to validate the results by the jackknife technique, also called the "leave n out" method (18). This simple but laborious method proceeds by removing a small number n of observations at a time from the original training data and recalculating the classifier for each of the truncated data sets. At each step, the "truncated" classifier is applied to just the removed measurement vectors which have not been used for the estimation of this classifier. This procedure is repeated until all measurement vectors have been classified by an appropriate truncated classifier; thus, the overall classification power obtained is a reliable estimate for the true value that can be expected for the classifier under consideration when it is applied to a new data set. Specifically, in our example with a training data set consisting of 42 measurement vectors, 14 separate networks (all starting with randomly set initial connection weights) were trained on truncated data sets consisting of 39 measurement vectors each. These truncated subsets were obtained in the following way. First, the complete data set consisting of 42 measurement vectors was randomly divided into 14 groups of three members each. One after another, each truncated subset consisting of 39 measurement vectors was obtained by omitting one of these triplets of measurement vectors from the full data set. After training the network on the truncated set, these three measurement vectors were categorized using this specifically trained network. A completely analogous procedure was used to validate by the jackknife technique the CART results as well as the linear discriminant model.

RESULTS

Table 2 shows the results obtained with the different methods for classification (the rate of misclassifications is only a crude indicator for the quality of a classification scheme; therefore, the transinformation values associated with the classification matrices are shown also). The total error function E after 10,000 iteration steps is 1.0748, and the change in the error function per iteration is smaller than 0.00001. The ability of the resulting network to reclassify the measurement vectors of the training data set is excellent: 41 of 42 cases are correctly classified with a corresponding output activity exceeding 0.90. Only one case, corresponding to a patient with CAH, is not classifiable by this network; for this measurement vector, the activity of all three output neurons is below 0.10. The associated transinformation of 1.4231 bits compares favorably with the maximum possible value of 1.5538 bits. However, jackknife validation of this network yields a significantly poorer classification result, which in fact is similar to the results obtained by the CART technique and by linear discriminant analysis.

Table 2. Results of the models studied

Model	Numbers of classifications		Transinformation	
	Correct	False	Value, bit(s)	χ^2
NN-4	41	1	1.4231	82.9
Jackknife	31	11	0.6110	35.6
NN-3	40	2	1.3381	77.9
Jackknife	31	11	0.5454	31.8
CART	35	7	0.8389	48.8
Jackknife	33	9	0.6180	36.0
DA	32	10	0.5432	31.6
Jackknife	31	11	0.4912	28.6

NN-4, neural network with four input neurons; NN-3, truncated neural network with three input neurons (ratio of transaminase activities omitted); CART, binary decision tree as shown in Fig. 2 (neopterin and ratio of transaminase activities included); DA, stepwise linear discriminant analysis (neopterin and ratio of transaminases were *jointly* significant).

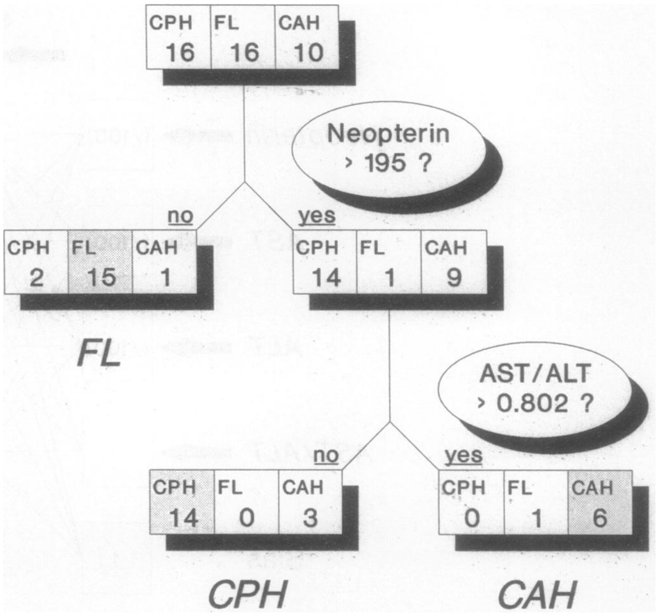

FIG. 2. Classification tree, obtained by the CART method.

The CART method yields the binary decision tree shown in Fig. 2: two of the four variables tested are selected by the algorithm—namely, neopterin and the AST/ALT ratio. The first decision ("neopterin above 195 μmol per mol of creatinine?") separates patients with FL from hepatitis patients, and the second decision ("AST/ALT above 0.802?") differentiates hepatitis patients into those with CPH and CAH. The CART model seems robust with respect to jackknife validation (Table 2).

Stepwise linear discriminant analysis identifies the same two variables (neopterin and AST/ALT ratio) as jointly significant classifiers. The rate of correct classifications is moderate and remains nearly unchanged after jackknife validation.

Importantly, when neural network architecture is used, the omission of the AST/ALT ratio from the measurement vectors (or, equivalently, removing the corresponding neuron from the neural network) has no marked effect on the results (compare NN-4 and NN-3 in Table 2). In strong contrast, the omission of this variable has a dramatic influence on the other two methods: of the three remaining variables (neopterin, AST, and ALT) only one, neopterin, is included into both the binary decision tree and the linear discriminant model, due to a lack of statistical significance for the enzyme concentrations taken singly. As a consequence, neither the CART technique nor linear discriminant analysis, under this setting without explicit "knowledge" of the AST/

ALT ratio, is able to differentiate between the two types of non-A, non-B hepatitis. The only differentiation possible is that between FL and hepatitis. This is easily seen in Fig. 2: when only neopterin is used for classification, the patients with hepatitis fall into one common subset which the algorithm does not further split. In an analogous manner, the linear discriminant model using only neopterin is not able to differentiate between the two types of hepatitis patients (not shown in detail).

Fig. 3 shows the response of both neural networks (NN-4 and NN-3) to varying input values of neopterin and AST/ALT ratio and the corresponding response surfaces obtained with the linear discriminant model (combining neopterin and AST/ALT ratio). Both neural networks yield similar response surfaces, and it is obvious from Fig. 3 that neural networks reproduce details of the actual distributions of training data which, for conceptual reasons, cannot be modeled as accurately by the classical statistical approach. Furthermore, the regions associated with different classes are very sharply separated in the case of neural networks as a consequence of the steep sigmoidal response function employed; linear discriminant analysis yields much less steep response surfaces. As a consequence, the activations of output neurons in most situations are very close to 0.0 or 1.0, whereas in linear discrimination intermediate values are very often obtained for the probabilities of categories.

DISCUSSION

Neural network architecture presents a valuable candidate model for medical diagnosis based on laboratory information. The ability of error back-propagation networks to correctly predict a diagnosis is at least as high as that of other methods, such as CART or linear discriminant analysis. In the most important case, however, that features being hidden in the data are important for efficient discrimination (such as, in our example, the AST/ALT ratio), neural networks are obviously far superior to the conventional approaches: by the training process, the networks learn to extract such hidden properties. In contrast, methods based on classical statistical

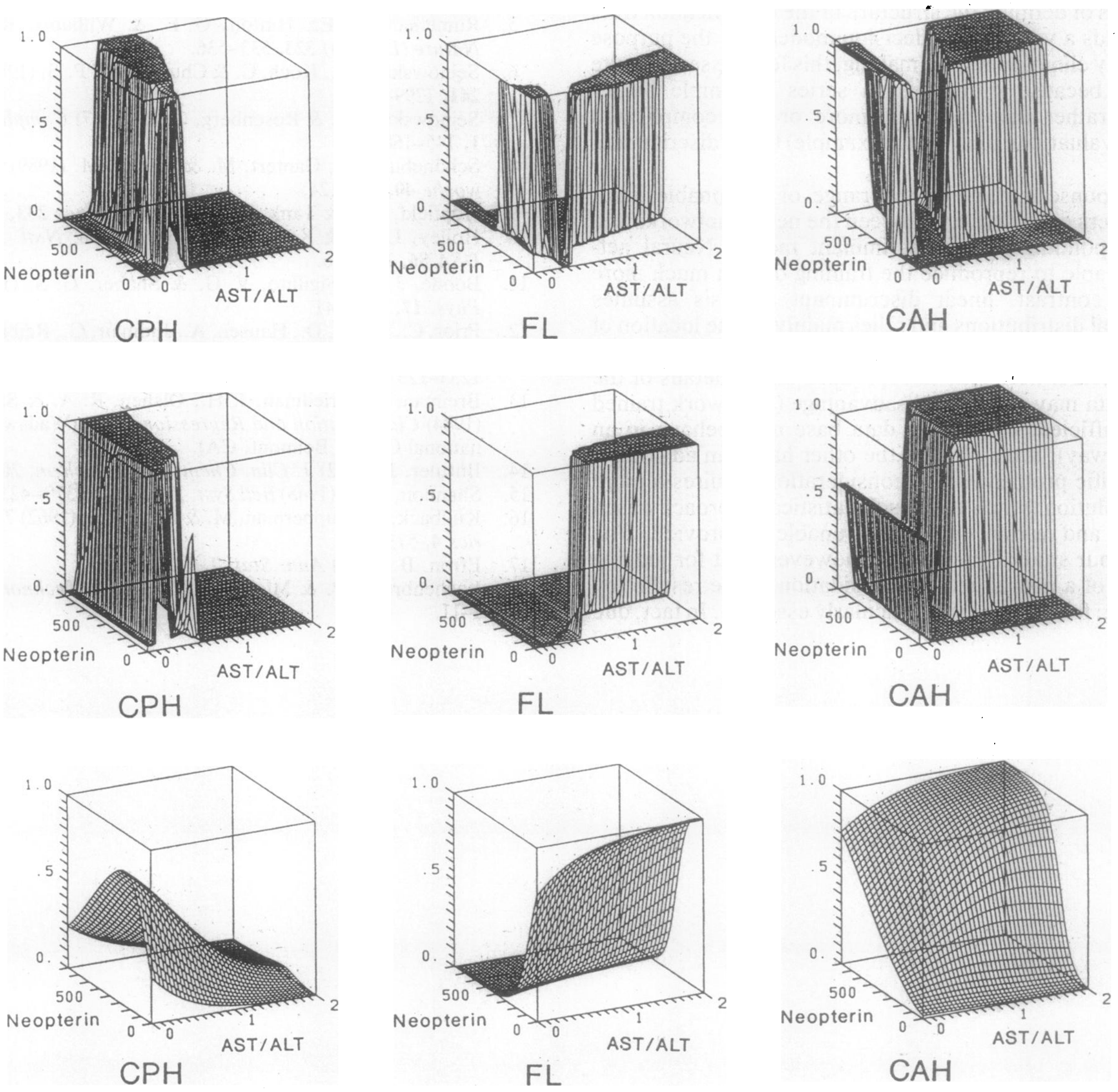

FIG. 3. Response surfaces of neural networks to different input values compared with results obtained by linear discriminant analysis (response values and probability values are shown along the *z* axis). (*Top*) Neural network with four input neurons (see Fig. 1). Profiles of output activity of output neurons after processing varying input signals, which were laboratory measurements of neopterin (μmol/mol of creatinine) and the ratio between serum activities of AST and ALT (AST activity was held constant at 50 units per liter). (*Middle*) Analogous response surfaces as above, but obtained for the truncated neural network with only three input neurons (AST/ALT ratio was omitted). (*Bottom*) Probabilities of diagnostic categories, obtained by linear discriminant analysis, in dependence on the same variables (again, AST activity was held constant at 50 units per liter).

algorithms fail to show this adaptive behavior; they rely on explicit statements of such relationships.

The data set and the discrimination problem studied here were relatively simple. The advantages of neural network methodology, particularly the ability to extract hidden features from the input signals, can be expected to be particularly useful when data sets with very complicated correlation structures are to be analyzed or when data sets contain, for example, irregularly spaced time series of many highly intercorrelated variables, perhaps including missing data. In such situations, which are typically met in everyday practice of laboratory-based medical diagnosis, classical statistical models may be difficult or even impossible to apply. Finally, the usual probability theoretical method for evaluation of diagnostic information (e.g., by application of Bayes' theorem) requires that the diagnostic cases are mutually exclusive. This requirement is not always fulfilled in clinical practice. Neural networks, however, could easily be trained to recognize overlapping diagnostic categories.

An advantage of the CART method should be mentioned: this technique, although it requires laborious computations in the process of defining the structure of the classification tree, finally yields a very simple decision model. For the purpose of everyday clinical decision making, this feature seems quite attractive because it involves a series of simple binary decisions rather than requiring more or less complicated function evaluations such as (for example) linear discriminant models.

The response surface to the range of reasonable input values differs considerably between the neural networks and the corresponding linear discriminant model. Neural networks are able to reproduce the training data in much more detail; in contrast, linear discriminant analysis assumes multinormal distributions and relies mainly on the location of mean values and on the associated variances. Clearly, the ability of neural networks to reproduce minor details of the training data may also be a disadvantage (a network trained on an insufficient or illogical data base may behave in an undesired way); it may be, on the other hand, an advantage if the specific problem under consideration requires a more detailed solution which the usual statistical approach based on means and variances may be unable to provide. The results of our study demonstrate, however, that for judging the merits of a neural network, validation of the results (by jackknifing, for example) is particularly essential. In fact, due to the excellent reproduction of minor details of the training data set by the neural networks, validation is even more important than with the classical statistical approaches.

The main advantage of neural network-based decision models lies in their adaptive behavior. For the practice of clinical decision making, it would be interesting to inquire with more emphasis into the similarities and dissimilarities between models involving parallel distributed processing and the strictly rule-based so-called "expert systems." Possibly, these alternative ways of extracting a decision from incomplete and uncertain information may show complementary behavior.

1. McCulloch, W. S. & Pitts, W. (1943) *Bull. Math. Biophys.* **5,** 115–133.
2. Rosenblatt, F. (1962) *Principles of Neurodynamics* (Spartan, New York).
3. Cooper, L. N. (1973) in *Proceedings of the Nobel Symposium on Collective Properties of Physical Systems*, eds. Lundqüist, B. & Lundquist, S. (Academic, New York), pp. 252–264.
4. Hopfield, J. J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2554–2558.
5. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) *Nature (London)* **323,** 533–536.
6. Sejnowski, T. J., Koch, C. & Churchland, P. S. (1988) *Science* **241,** 1299–1306.
7. Sejnowski, T. J. & Rosenberg, C. R. (1987) *Complex Systems* **1,** 145–168.
8. Schöneburg, E., Gantert, M. & Reiner, M. (1989) *Computerwoche* **40,** 121–124.
9. Hopfield, J. J. & Tank, D. W. (1986) *Science* **233,** 625–633.
10. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
11. Boone, J. M., Sigillito, V. G. & Shaber, G. S. (1990) *Med. Phys.* **17,** 234–241.
12. Prior, C., Fuchs, D., Hausen, A., Judmair, G., Reibnegger, G., Werner, E. R., Vogel, W. & Wachter, H. (1987) *Lancet* **ii,** 1235–1237.
13. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984) *Classification and Regression Trees* (Wadsworth International Group, Belmont, CA).
14. Büttner, J. (1982) *J. Clin. Chem. Clin. Biochem.* **20,** 477–490.
15. Shannon, C. E. (1948) *Bell Syst. Tech. J.* **27,** 379–423, 623–656.
16. Kullback, S., Kupperman, M. & Ku, H. H. (1962) *Technometrics* **4,** 573–608.
17. Efron, B. (1979) *Ann. Stat.* **7,** 1–26.
18. Lachenbruch, P. & Mickey, R. M. (1968) *Technometrics* **10,** 1–11.