JACC Vol. 26, No. 7
December 1995:1586–93

One-Year Mortality Prognosis in Heart Failure: A Neural Network Approach Based on Echocardiographic Data

JUAREZ ORTIZ, MD, CLAUDIA G. M. GHEFTER, MD, CARLOS E. S. SILVA, MD, RENATO M. E. SABBATINI, PHD*

São Paulo and Campinas, Brazil

Objectives. **This study sought to assess the usefulness and accuracy of artificial neural networks in the prognosis of 1-year mortality in patients with heart failure.**

Background. **Artificial neural networks is a computational technique used to represent and process information by means of networks of interconnected processing elements, similar to neurons. They have found applications in medical decision support systems, particularly in prognosis.**

Methods. **Clinical and Doppler-derived echocardiographic data from 95 consecutive patients with diffuse impairment of myocardial contractility were studied. After 1 year, data regarding survival or death were obtained and produced the prognostic variable. The data base was divided randomly into a training data set (47 cases, 8 deaths) and a testing data set (48 cases, 7 deaths). Results of artificial neural network classification were compared with those from linear discriminant analysis, clinical judgment and conventional heuristically based programs.**

Results. **The study group included 57 male (47 survivors) and 38 female patients (33 survivors). Linear discriminant analysis was not efficient for separating survivors from nonsurvivors because the accuracy at the ideal cutoff value was only 67.4%, with a sensitivity of 67.5%, positive predictive value of 27.8% and negative predictive value of 91.5%. In contrast, all artificial neural networks were able to predict outcome with an accuracy of 90%, specificity of 93% and sensitivity of 71.4%, for the best artificial neural network. Both clinical judgment and automatic heuristic methods were also inferior in performance.**

Conclusions. **The artificial neural network method has proved to be reliable for implementing quantitative prognosis of mortality in patients with heart failure. Additional studies with larger numbers of patients are required to better assess the usefulness of artificial neural networks.**

(J Am Coll Cardiol 1995;26:1586–93)

Prognosis, or prediction of the future evolution of disease, is an important step in the evaluation of patients with chronic heart disease. Prognosis can be expressed in many forms, such as quality of life and symptom-free period, but the most common type of prognostic evaluation in patients with severe heart disease is survival within a given period, or outcome.

Because outcome is usually influenced by a conglomeration of abnormalities, found in one or more measurable variables, multivariable pattern classification and recognition techniques represent a useful approach to quantitative prognosis. Linear discriminant analysis, multilinear regression analysis and logistic regression analysis of known patient data have all been used extensively for such types of causal–structure evaluation in medical prognosis. These methods are based on linear multidimensional models, that is, patterns that are geometrically represented as points in a multidimensional cartesian space, where the axes of the space are measurements or predictor variables (also called features). Classes of patients (e.g., survivors and nonsurvivors) are said to be linearly separable whenever the statistical method is able to find a hyperplane (a plane in several dimensions) that can separate their representative points into two regions of the feature space. Computations for linear statistical analysis are usually straightforward and based on numeric matrix algebra procedures (1).

However, many problems in medical pattern classification cannot be approached by linear methods because separability of patient classes in the feature space can be achieved only by more complex, nonlinear decision surfaces. When this occurs, conventional numeric algorithms for nonlinear pattern classification can be developed, but they are usually computationally awkward and inefficient. In this respect, several investigations have demonstrated that a recently developed technology of artificial intelligence, artificial neural networks, offers noticeable advantages for implementing a system of diagnosis and risk stratification in outcomes research.

Artificial neural networks, or connectionist systems, are being used increasingly to represent and process information by means of networks of interconnected processing elements, similar to neurons. Several emerging global properties of connectionist systems, such as associative memory and distributed parallel processing, have favored its application in a wide variety of tasks involving pattern classification and recognition. Connectionist systems constitute a new and interesting para-

From the Centro de Cardiologia Não Invasiva, São Paulo and *Center for Biomedical Informatics, State University of Campinas, Campinas, Brazil.

Manuscript received June 27, 1994; revised manuscript received July 7, 1995; accepted July 13, 1995.

Address for correspondence: Dr. Juarez Ortiz, Centro de Cardiologia Não Invasiva, Rua Cubatão, 726, São Paulo, SP, CEP 04013-002, Brazil.

0735-1097/95/$9.50
0735-1097(95)00385-H

digm for the use of artificial intelligence in medicine and have found applications in processing and interpretation of biologic signals and images and decision support systems (2–4).

Artificial neural networks can be used, among other techniques, to find approximate maps between input and output patterns of a quantitative nature. This end is achieved by exemplar-based learning, that is, by presenting to the network data about patients with known outcomes of the disease process (e.g., death or survival after a given period of observation). Using the proper learning algorithm, the network may be able to predict outcome after reaching a minimal classification error, whenever input (predictor) variables are related in an orderly manner to outcome. One characteristic property of a certain kind of multilayer artificial network is that nonintuitive, complex nonlinear separation between patient classes can be achieved. Neural networks often have the ability to find hidden features in input space, where none are visible by conventional statistical methods or by human decision alone. In medical applications connectionist models have been used for predicting outcome in patients with coronary heart disease (5), patients in intensive care (6,7), wound infections (8), the onset of diabetes (9), breast cancer (10) and colon cancer recurrence (11), among others. In some recent applications of medical decision making, such as diagnosis of acute myocardial infarction, the performance of artificial neural networks has been impressive compared with conventional quantitative methods and human judgment (12).

The present study sought to evaluate the usefulness and accuracy of a particular artificial neural network in a task of medical prognosis, using a complex mixture of predictor variables, based on echocardiographic data, namely the prediction of 1-year mortality in patients with heart failure.

Methods

Clinical and echocardiographic data from 95 consecutive patients with dilated myocardial disease and heart failure were studied. Inclusion criteria were 1) clinical diagnosis of heart failure, and 2) echocardiographic demonstration of dilated left ventricle and diffuse impairment of myocardial contractility. Patients with arrhythmias and heart failure resulting from primary valvular disease were excluded.

The time of examination in relation to the onset of extant disease was not controlled. At the echocardiographic examination and during the follow-up period, treatment was determined solely by each patient's cardiologist, in accordance with their own criteria. Drug therapy was not homogeneous and included digitalis (91% of patients), diuretic drugs (73%), vasodilator agents (35%) and antiarrhythmic agents (12%).

Echocardiographic examination. M-mode, two-dimensional and Doppler examinations were performed using a commercially available Aloka SSD-870 ultrasound system with a 2.5- or 3.5-MHz transducer. The variables analyzed are shown in Table 1, and echocardiographic measurements were made according to the American Society of Echocardiography and Penn convention recommendations.

Table 1. Clinical and Echocardiographic Variables

Aortic diameter (mm)
Left atrial diameter (mm)
Left atrial diameter/body surface area (mm/m^2)
Left ventricular systolic diameter (mm)
Left ventricular systolic diameter/body surface area (mm/m^2)
Left ventricular diastolic diameter (mm)
Left ventricular diastolic diameter/body surface area (mm/m^2)
Left ventricular posterior wall thickness (mm)
Interventricular septal diastolic thickness (mm)
Left ventricular diastolic diameter/myocardial thickness ratio
Mitral valve E point minus septal separation (mm)
Mitral valve flow deceleration time (ms)
Left ventricular mass (g)
Left ventricular fractional shortening (%)
Left ventricular ejection fraction (%)
Heart rate (beats/min)
Cardiac index (ml/m^2)
Left ventricular filling pattern

Left ventricular mass (LVM) and volumes (Teichholz) were calculated using the following formula:

$$\mathrm{LVM} = [(\mathrm{LVDD} + \mathrm{LVPWTh} + \mathrm{IVSTh})^3 - (\mathrm{LVDD})^3] \times 1.05 - 14;$$

$$\mathrm{Volume} = \frac{\dfrac{7.0 \times (\mathrm{LVDD})^3}{2.4 + \mathrm{LVDD}} - \dfrac{7.0 \times (\mathrm{LVSD})^3}{2.4 + \mathrm{LVSD}}}{\dfrac{7.0 \times (\mathrm{LVDD})^3}{2.4 + \mathrm{LVDD}}},$$

where LVDD = left ventricular diastolic diameter; LVPWTh = left ventricular posterior wall thickness; IVSTh = interventricular septal wall thickness; and LVSD = left ventricular systolic diameter.

Left ventricular filling pattern (flow) was coded as follows: 0 = normal or "normalized"; 1 = rapid filling predominant pattern; 2 = atrial filling predominant pattern.

Clinical data. The following data were obtained by patient interview at the time of the initial examination: age (in years); gender (coded as 1 = male, 2 = female); New York Heart Association functional class (coded as 1 = class I; 2 = class II; 3 = class III; 4 = class IV).

Basic presumed underlying etiology was classified according to the following types: 1 = Chagas' disease; 2 = hypertensive cardiomyopathy; 3 = coronary artery disease; 4 = idiopathic cardiomyopathy; 5 = myocarditis; 6 = alcoholism; 7 = diabetes.

During the follow-up period, information regarding survival or circumstances of death were obtained by letter or telephone interview. This information produced the prognostic variable.

Outcome after a 12-month observation period was coded as follows: 0 = survivor; 1 = nonsurvivor.

Data preparation. The full data matrix was analyzed statistically using the EPI-INFO 5.0 system (developed by the Centers for Disease Control and Prevention) for the following: 1) frequency, mean value and standard deviation for each variable: 2) chi-square measures of association between cate-

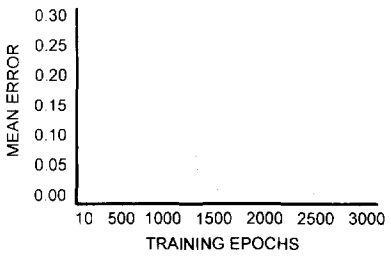


Figure 1. Example of learning curve for an artificial neural network trained to make prognosis for 1-year mortality in congestive heart disease. Each training epoch represents the mean squared error of classification achieved by the artificial neural network after running through all input case examples.

goric variables, such as gender, etiology, flow pattern, functional class and 12-month mortality; 3) analysis of variance or the Student *t* test of significance of differences between survivor and nonsurvivor groups for all scalar variables; 4) linear correlation and regression between selected scalar variables; 5) scattergrams between selected scalar variables and column diagrams between all scalar variables and patient group (survivors or nonsurvivors).

In addition, a linear discriminant analysis was performed with the full data set of scalar variables to evaluate its separability power on the original patient data, as well as to further point out the linear predictive power of individual variables. The results of this analysis were used to derive an overall picture of data association and causal structure, and to guide the selection of input variables to the neural network.

Neural network approach. Finally, the full relation data base was divided randomly into a *training data* base (47 cases, 8 deaths) and a *testing data base* (48 cases, 7 deaths) and each one converted to the ASCII sequential files required by the neural network and linear discriminant analysis programs. For the linear discriminant analysis, only scalar variables were used, and, where appropriate, body surface area–corrected indexes only. For use by the neural network, all input variables were automatically standardized into the interval [0, 1]. The maximal and the minimal values used for standardizing each variable were obtained from the univariate data analysis of the full data set. Categoric variables, when binary (e.g., gender), were allocated to a single node using 0 and 1. When the variable was ordinal (e.g., functional class), it was allocated a sequential integer and then standardized to the range [0, 1]. When the variable was multinominal (e.g., etiology), it was allocated to a suite of binary nodes, each one holding a [0/1] value for each possible alternative (e.g., ET1, ET2, ET3).

After extensive experimentation with several network configurations, reported elsewhere (13), we centered our investigation on the prognostic ability of artificial neural networks in dilated heart disease with the following configurations (Fig. 1): 1) Three-layer feed-forward network with the plain backpropagation learning rule. 2) Input variables: Two sets of variables were investigated (Table 2): a) *complete* = almost all recorded variables, with 29 nodes for 18 variables; b) *reduced* = only the 11 most significant variables, indicated by the univariate and linear discriminant statistical analysis, giving 13 input nodes (number of hidden nodes = 2, 5 or 10). 3) One output node (survival/death). 4) Learning method: backpropagation of er-

Table 2. Prognosis of Dilated Myocardial Disease: Input Variables

Input Variable	Variable Type	Analysis Set		
		LDA	Complete Set ANN	Reduced Set ANN
Age (yr)	S	×	×	×
Aortic diameter (mm)	S	×		
Cardiac index (ml/m^2)	S	×	×	
Etiology (1–7)*	C		×	
Mitral valve flow pattern (0–2)*	C		×	×
Heart rate (beats/min)	S	×	×	
IVS (mm)	S	×	×	×
Left atrial index (mm/m^2)	S	×	×	×
Left ventricular diameter/thickness ratio	S	×	×	×
Left ventricular diastolic diameter index (mm/m^2)	S	×	×	×
Left ventricular ejection fraction (%)	S	×	×	×
Left ventricular mass (g)	S	×	×	
Left ventricular systolic diameter index (mm/m^2)	S	×	×	×
Mitral E point IVS separation (mm)	S	×	×	×
Mitral valve insufficiency (0–4)*	C		×	
Mitral flow deceleration time (ms)	S	×	×	×
Initial NYHA index functional class (1–4)*	O		×	
Left ventricular posterior wall thickness (mm)	S	×	×	×
Gender (1–2)*	C		×	

*See Table 4 for explanation of codes. ANN = artificial neural network; C = categoric; IVS = interventricular septal; LDA = linear discriminant analysis; NYHA = New York Heart Association; O = ordinal; S = scalar.

Table 3. Prognosis of Dilated Myocardial Disease: Univariate Analysis of Association for Scalar Variables

Scalar Variable	Survival		Death		Statistics		
	Mean	SD	Mean	SD	Diff	F*	p Value†
Age (yr)	50.10	14.73	56.87	17.36	6.77	2.44	0.1173
Aortic diameter (mm)	32.20	4.57	32.60	3.91	0.40	0.10	0.7499
Interventricular septal thickness (mm)	9.27	1.33	8.87	1.19	0.40	1.23	0.2702
Left ventricular posterior wall thickness (mm)	8.91	1.29	8.40	1.18	0.51	2.06	0.1511
Left ventricular diameter/thickness ratio	3.71	0.69	4.20	0.54	0.49	6.49	0.0120
Left ventricular fractional shortening (%)	19.07	4.69	14.48	3.31	4.59	11.81	0.0012
Mitral E point IVS separation (mm)	23.37	6.36	26.73	5.23	3.36	3.70	0.0542
Mitral flow deceleration time (ms)	183.16	57.92	157.40	51.61	25.76	2.58	0.1077
Heart rate (beats/min)	78.56	14.16	86.93	19.47	8.37	3.89	0.4850
Left ventricular diastolic diameter index (mm/m^2)	38.31	5.84	42.52	6.13	4.21	6.47	0.0121
Left ventricular systolic diameter index (mm/m^2)	31.08	5.51	36.59	5.83	5.53	12.42	0.0010
Left atrial index (mm/m^2)	24.27	4.20	28.83	4.04	4.56	15.04	0.0004
Left ventricular mass (g)	326.67	88.86	364.40	141.61	37.73	1.85	0.1739
Left ventricular ejection fraction (%)	38.10	8.69	29.74	6.16	8.36	12.62	0.0009
Cardiac index (ml/m^2)	49.29	14.52	48.91	18.71	0.38	0.01	0.9270

*Ratio statistics. †Two-tailed probability. Diff = difference between mean values; IVS = interventricular septal.

rors. 5) Transfer function: sigmoid. Internal bias of 0.0, learning rate of 1.0 to 1.5 and momentum of 0.0; 6) Weight initialization: randomized in the range [−0.5, +0.5]. Method of presentation of examples during training: randomized. Method of weight updating: continuous.

Evaluation of artificial neural network performance was carried out using the original data set for each network, as well as its complementary test data set, containing patient data not used for training the network. For each patient in these data sets, the program compared actual with predicted outcome, generating a file of comparative results. Finally, this file was analyzed and test variables were computed, on the basis of a 2×2 contingency table constructed from expected or obtained statistics (accuracy, sensitivity, specificity and predictivity).

Because the output of the artificial neural network and linear discriminant analysis were graded, that is, a number in the inclusive continuous range [0, 1] was produced for each test record, it was necessary to define a cutoff point to count the number of true and false predictions to build the 2×2 contingency table. Because this definition is essentially arbitrary, the best procedure is to study the effect of different cutoff points on the performance statistics. For each cutoff point, sensitivity and specificity were calculated and plotted one against the other in a two-dimensional line graph, producing a receiver operating characteristic curve (14).

Clinical judgment. An additional proposition was to compare the results of the artificial neural network classification with those resulting from clinical judgment and conventional heuristic-based programs. We achieved this by estimating the classification accuracy of heuristic rules derived previously by a group of experienced physicians for the present data set (15), as well as by a commercially available program for automatic induction, KnowledgeSEEKER (First Mark Technology).

Results

The study group included 95 patients (57 male [47 survivors], 38 female [33 survivors]; range 13 to 86 years old, mean [±SD] 51.25 ± 15.27). Presumed etiologies were distributed in descending order of prevalence as follows: idiopathic cardiomyopathy (46.3%), hypertensive cardiomyopathy (23.2%), Chagas' disease (13.7%), coronary artery disease (12.6%) and other cardiomyopathies, including those due to diabetes and alcoholism (4.2%).

For nonsurvivors, time of survival ranged from 23 to 367 days (mean 166.07 ± 123.7).

Statistical analysis. The results of the univariate statistical analysis of study variables are shown in Tables 3 and 4). Scalar variables significantly associated with 12-month mortality were left ventricular diastolic diameter/thickness ratio, mitral E point to interventricular septal separation, left ventricular diastolic diameter index, left ventricular systolic diameter index, left atrial diastolic index, all directly proportional to death risk; left ventricular fraction shortening, mitral valve flow deceleration time and left ventricular ejection fraction, all inversely proportional to death risk. The variables not associated with death risk were age, diameter of aorta, interventricular septal thickness, left ventricular posterior wall thickness, heart rate, left ventricular mass and cardiac index. Furthermore, all categoric variables, such as gender, disease etiology, type of mitral flow and functional class were not significantly associated with mortality risk. Several of the significant variables were statistically correlated to a larger or lesser degree, but only left ventricular fractional shortening was removed from the variable set because it had a linear correlation of 1.0 with left ventricular ejection fraction.

The linear discriminant analysis results are shown in Table 5. As judged from the final discriminant weights, the most important variable was left ventricular systolic diameter index.

Table 4. Prognosis of Dilated Myocardial Disease: Univariate Analysis of Association for Categoric Variables

Categoric Variable and Code/Definition	Survival (%)	Mortality Rate (%)	Chi-Square	p Value
Gender				
1/male	58.75	66.67	0.33	0.5657
2/female	41.25	33.33		
Etiology				
1/Chagas' disease	11.25	26.67	5.57	0.3499
2/hypertension	25.00	13.33		
3/coronary disease	13.75	6.67		
4/idiopathic	46.25	46.67		
5/myocarditis	1.25	6.67		
6/alcoholism	1.00	0.00		
7/diabetes	1.00	1.00		
Mitral flow				
0/normal	46.25	26.67	2.93	0.2310
1/restrictive	31.25	53.33		
2/atrial dominant	22.50	20.00		
Initial NYHA class				
I	30.00	13.33	3.92	0.2698
II	35.00	26.67		
III	20.00	40.00		
IV	15.00	20.00		

NYHA = New York Heart Association.

Other important variables were left ventricular posterior wall thickness, left ventricular diastolic diameter/thickness ratio, left ventricular diastolic diameter index and left ventricular ejection fraction.

The linear discriminant analysis was not efficient enough to distinguish survivors from nonsurvivors on the basis of input scalar variables. The distributions of the discriminant function for both patient groups overlapped considerably (Fig. 2), and

Table 5. Prognosis of Dilated Myocardial Disease: Results of Linear Discriminant Analysis*

Variable	Discriminant Coefficient
Age (yr)	0.0595
IVS thickness (mm)	−0.7998
Left ventricular posterior wall thickness (mm)	−2.1362
Left ventricular diameter/thickness ratio	1.3513
Mitral E point IVS separation (mm)	0.0719
Mitral flow deceleration time (ms)	0.0098
Heart rate (beats/min)	0.0910
Left ventricular diastolic diameter index (mm/m^2)	−1.2668
Left ventricular systolic diameter index (mm/m^2)	17.5586
Left atrial index (mm/m^2)	0.5296
Left ventricular mass (g)	−0.0762
Left ventricular ejection fraction (%)	−2.4943
Cardiac index (ml/m^2)	−0.0320

*Mean value for discriminant function (death): 495.85. Mean value for discriminant function (survival): 380.72. Difference between mean values (Mahalanobis distance): 115.13. Cutoff point of discriminant functions: 438.28; z = −5.36, F ratio 142.40; degrees of freedom 13/93. IVS = interventricular septal.

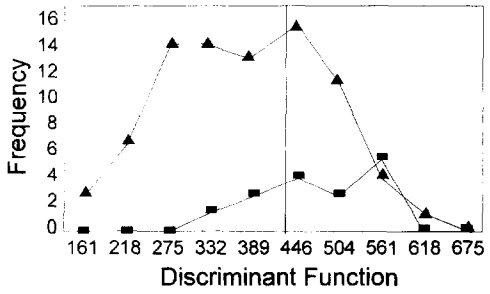


Figure 2. Distribution of linear discriminant function calculated for two groups of patients, according to the 1-year mortality prognosis: patients with congestive heart disease who died **(squares)** and patients who survived **(triangles)**. **Vertical line** indicates the cut point for best separation of the two statistical classes.

accuracy at the ideal cutoff value was only 67.4%, with a sensitivity of 67.5%. The weakest feature of the linear discriminant analysis solution was positive predictivity, which was only 27.8% (i.e., linear discriminant analysis predicted far more deaths than actually occurred in the patient sample), whereas negative predictivity (91.5%) was excellent. These results show that a statistical linear model is not able to perform class separation in multidimensional space and that a nonlinear approach is justified.

Neural network analysis. The artificial neural network results are summarized in Table 6. In the discussion that follows, artificial neural network configurations are denoted by three numbers, meaning the number of nodes in the input, hidden and output layers, respectively (e.g., 13-5-1).

All networks were trained to convergence within 3,000 training cycles, and many of them were able to converge to an acceptable mean error just below 1,000 cycles. Global mean squared error was <0.01 in all cases, reaching up to 0.004 for the best cases. The distribution of network output gave evidence that the solutions converged to the near proximity of 0 and 1 in all cases (Fig. 3). Only ~9% of all outputs were in the range 0.1 to 0.8.

All networks were able to predict outcome (Table 6, Fig. 3), with accuracy ranging from 93.33% (13-10-1) to 100% (several nets) for the training data set and from 72% (29-2-1) to 90% (13-5-1) for the test data set. In general, the networks with the reduced set of variables (11 variables) fared better than those with the the full set of variables (18 variables); and within each of these two groups of artificial neural networks, those with the intermediate number of hidden nodes had the best performance. However, differences were not large, except for the best network (13-5-1), with 11 variables and 5 hidden nodes, which yielded 90% accuracy, 71.4% sensitivity and 93% specificity. The worst variable for all networks was always sensitivity, denoting a relative inability to predict correctly the number of nonsurvivors.

Clinical judgment analysis. When applied to the full data set, the deterministic rule produced by the clinicians yielded an accuracy of 82%, a sensitivity of 46.7% and a specificity of 88%. The rule produced by the clinicians was as follows:

Table 6. Prognosis of Dilated Myocardial Disease: Comparative Results of Classification Performance

Method	Data Set	Accuracy	Sensitivity	Specificity	Positive Predictivity	Negative Predictivity
Linear discriminant analysis	Full set	67.37	66.67	67.50	27.78	91.53
Heuristics rule by clinician	Full set	82.00	46.67	88.75	43.75	89.87
Heuristics rule by program	Full set	91.50	46.67	100	100	90.90
Neural network 13-2-1	Training set	95.60	75.00	100	100	94.87
	Test set	88.00	42.90	95.30	60.00	91.10
Neural network 13-5-1	Training set	100	100	100	100	100
	Test set	90.00	71.42	93.00	62.50	95.24
Neural network 13-10-1	Training set	93.33	75.00	97.33	85.70	94.70
	Test set	88.00	28.60	97.70	50.00	89.10
Neural network 29-2-1	Training set	97.80	87.50	100	100	97.40
	Test set	72.00	42.90	82.20	33.33	84.60
Neural network 29-5-1	Training set	100	100	100	100	100
	Test set	80.00	42.90	86.00	33.33	90.20
Neural network 29-10-1	Training set	100	100	100	100	100
	Test set	72.40	42.86	76.74	23.08	89.19

If MVDT ≤ 140 ms and LADI ≥ 25 mm/m^2

and LVDDTh ≥ 3.4, then Outcome = Death,

where MVDT = mitral valve flow deceleration; LADI = left atrial diastolic index; and LVDDTh = left ventricular diastolic diameter/thickness ratio. The rule produced by the heuristic methods was as follows:

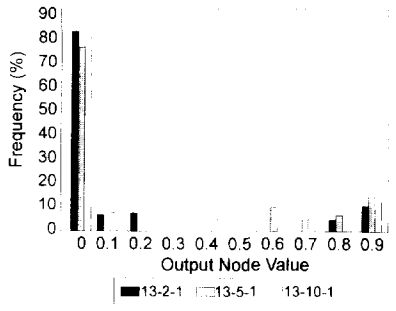
If LADI ≥ 28.571 mm/m^2 and HR ≥ 75 beats/min

and LVEF ≤ 31.121, then Outcome = Death (100%),

where HR = heart rate; and LVEF = left ventricular ejection fraction.

Using these rules, the actual mortality rate was 10.1% in the low risk group versus 43.8% in the high risk group (statistically significant, chi-square 11.31, 1 degree of freedom, p = 0.00077). The decision rule formulated by physicians had unexpected characteristics because it ignores other variables with a higher univariate association with survival. The rules produced by conventional heuristic methods, besides being different from the rule produced by the clinicians, is also surprising because it uses a variable (heart rate) that was not found to be significantly associated with outcome in the univariate analysis. The rule achieved an accuracy of 91.5%, a sensitivity of 46.7% and a specificity of 100% (Table 6). Using these criteria, the actual mortality rate was 10% for the low risk group and 100% for the high risk group (statistically significant, chi-square 40.3, 1 degree of freedom, and p < 0.000001), which was better than human judgment.

Figure 3. Distribution of prediction outputs for three completely trained artificial neural networks. Each **vertical bar** represents the relative frequency of cases for which the neural network produced a graded output of the value indicated on the **horizontal axis.** A value near 0 indicates a low probability of death within 1 year. A value near 1 indicates a high probability of death. The **network number** indicates the number of input nodes (e.g., 13), the number of hidden nodes (e.g., 2) and the number of output nodes (e.g., 1). Data refer to the test data set.



Discussion

Comparison of neural network performance, linear discriminant analysis and heuristic rules. In analyzing the performance of neural networks, linear discriminant analysis and heuristic rules in outcome prognosis of dilated myocardial disease it is clear that all the methods systematically predicted more deaths than actually occurred.

When all 95 patients in the training and test data sets are considered, the best artificial neural network had only five erroneous predictions: three falsely positive (Patients 19, 26 and 38; the latter died within 24 months) and two falsely negative (Patients 44 and 48). Because the number of nonsurvivors is considerably lower than that of survivors, the impact of three errors is large, leading to a low positive predictivity for all neural networks tested in the present study. A highly inflated false positive error was also observed for the linear discriminant analysis, which had a very low positive predictivity (27.8%). Simple heuristic rules derived by clinical expertise or automatic induction also encountered the same problem, that is, high specificities and low to medium sensitivities.

This result could be explained by the fact that patients underwent uncontrolled therapeutic intervention after the initial examination. Assuming that therapy had been conducted effectively for most patients, it is reasonable to expect

JACC Vol. 26, No. 7
December 1995:1586–93

that many patients who were predicted to die, did not. Therefore, adding therapeutic variables to the predictor variables could lead to an improvement in decision specificity. Another difficulty occurs in diseases where the proportions of outcomes differ too much (in the present study 10.5% of nonsurvivors vs. 84.2% of survivors within 12 months). In this situation, it is harder to build a model in which the lowest risk is predicted with the same accuracy as the highest risk.

Performance of artificial neural networks. The performance of artificial neural networks in the prognosis of outcome for patients with dilated myocardial disease can be rated as very good if the following are considered: 1) The time of measurement of variables was not fixed in relation to onset of disease; 2) a large number of input variables were associated with outcome; 3) input variables presented large variability among patients of the same outcome class; 4) the number of training examples was too low in relation to problem dimensionality; 5) subsequent therapeutic intervention was not controlled or accounted for; and 6) several etiologic factors were present. A nearly optimal combination of high sensitivity and specificity was achieved with the 13-5-1 network, which was the decision method that outperformed all others. The impressive accuracy of 90% achieved with the test data demonstrates that this neural network was able to find a decision surface with acceptable prognostic power by using only the clinical and echocardiographic variables provided to input and that the aforementioned factors are not really necessary.

We can therefore support the proposition that artificial neural networks are valid and reliable tools for carrying out quantitative medical decision making in the area of prognosis of cardiac diseases.

Whereas both heuristic solutions have good to excellent accuracies and specificities, similar to those obtained by neural networks, sensitivities are unacceptably low for good clinical decision support (46.7%). Furthermore, they were derived with knowledge of the full data set, so the comparison with artificial neural networks, which had much more stringent criteria of test performance, cannot be made with fairness. In fact, all artificial neural networks with 5 to 10 hidden neurodes (processing elements) tested in the present study had accuracies, sensitivities and specificities of 100% when tested with the full data set. Even when tested with the separate test data set, receiver operating characteristic artificial neural network curves lie above the point of performance of heuristic rules (Fig. 4).

In the present work, we found that the best network contained five hidden neurodes and a reduced set of variables. Unfortunately, the number of cases available for training and testing was not large, and some of the classification inaccuracies may stem from this. Performance statistics (accuracy, sensitivity, specificity) become very unstable and unreliable when the number of cases is small in any given class; that is, one false negative or false positive error has a disproportionately larger impact on them. Thus, generalization seems to have been achieved, but larger sample sizes would provide a firmer ground for this statement.

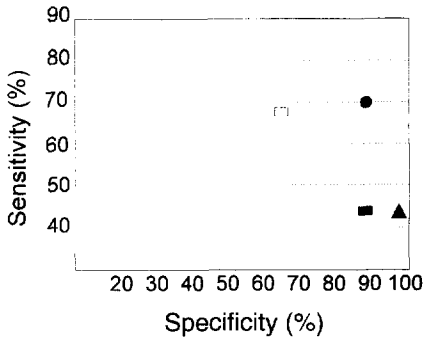


Figure 4. Comparison of classification performance of decision methods for the prognosis of 1-year mortality in patients with heart failure. **Symbols** depict the best sensitivity/specificity pairs for the 13-5-1 neural network (**circle**), logical rules produced by a clinician (**solid square**), by an automatic induction program (**triangle**) and by linear discriminant analysis (**open square**).

Limitations of clinician rules. In the present study, we observed that expert clinicians have a strong tendency to reason in terms of simple, easy to remember rules for estimating prognosis for individual patients and that they often use them to justify their own prescience. Moreover, the "black box" nature of neural networks and the need to use a computer to apply their decisions to everyday prognostic problems will surely preclude their wider use in medicine for some time to come (15). In this context, artificial neural networks could be more useful if used to extract the dominant rules that are able to achieve highest classification performances. Work along these lines is already taking place (16), and hybrid decision support systems, which combine artificial neural networks and logical reasoning in different ways seem to have a promising future for medical application (8,17).

Another important lesson extracted from this study is that accuracy alone is not a good indicator of performance for medical decision-making tools. Because of the disproportion of outcome rates in the patients studied, specificity and sensitivity tend to assume rather disparate values, usually lower than the overall accuracy rate.

Artificial neural networks and medical decision making. The use of artificial neural networks proved useful for implementing medical decision making in the area of outcome prognosis because 1) a reference standard (a definitive decision for outcome in the training and test data sets) exists, is clear cut and also avoids problematic procedures of appraisal by human experts. 2) Clinical medicine has a tradition of using quantitative analyses and graded estimates of outcome, thus making the addition of artificial neural networks easily acceptable. 3) The complex nonlinear mixture of predictor variables is always present and cannot be solved satisfactorily by conventional linear multivariate statistical analysis. Recent work (18) has combined artificial neural network technology with conventional statistical estimators, such as logistic regression and log-likelihood analysis, with promising results. A final advantage of artificial neural networks is that the prognostic model derived by them is closely dependent on the casuistics

gathered in a particular institution or geographic area. In consequence, artificial neural networks will perform better than "global" indexes (such as the well known APACHE index for outcome prognosis of critically ill patients) (19).

Clinical considerations. This study shows that echocardiography can be considered an excellent tool for diagnosis and evaluation of the prognosis of dilated myocardial disease (14) despite the recommendation by some investigators that other invasive and noninvasive diagnostic studies should be carried out to better evaluate pertinent variables in the disease's natural history (20). It is clear that many other potentially important variables for prognosis in dilated heart disease were not collected in the present work, such as angiographic and hemodynamic variables (21,22), myocardial thallium perfusion study results (23), myocardial biopsy results (24,25), electrocardiographic signals (26) and cardiovascular stress test results (27). Perhaps, adding them to the study would improve prognosis (and artificial neural network technology would make this easy to do), but this is uncertain, as evidenced by the results of using artificial neural networks with all input variables.

Finally, we are convinced that many uncontrolled variables, such as therapeutic intervention and individual organic predisposition, will make it impossible to achieve higher prognostic performance than that reported here. However, further studies with larger numbers of patients are required to better understand the mechanisms involved in systolic and diastolic function changes affecting the prognosis of patients with dilated myocardial disease.

References

1. Avanzolini G, Barbini P, Gnudi G. Unsupervised learning and discriminant analysis applied to identification of high risk postoperative cardiac patients. Int J Bio-Med Comput 1990;25:207–21.
2. Reggia JA, Sutton GG III. Self-processing networks and their biomedical implications. Proc IEEE 1988;76:580–92.
3. Reggia JA. Neural computation in Medicine. Artif Intell Med 1993;5:143–57.
4. Sabbatini RME. Aplications of connectionist systems in Biomedicine. Proceedings of the 7th World Congress on Medical Informatics (MEDINFO 92). International Federation of Medical Informatics. Amsterdam: North Holland, 1992:418–26.
5. Jayaweers A, Drake KC, Abbott R, Kaul S. Determination of long-term outcome in patients with coronary artery disease using an artificial neural network. J Am Coll Cardiol 1993;21:7A.
6. Felipe P Jr, Sabbatini RME, Carvalho-Júnior P. Beseggio RE, Terzi RGG. Outcome prediction for critical patients under intensive care, using backpropagation neural networks. Anais I Fórum Nacional de Ciência e Tecnologia em Saúde; 1992 Nov; Caxambu, Brazil, 1992:344–7.
7. Tu JV, Guerriere MRJ. Use of a neural network as a predictive instrument of length of stay in the intensive care unit following cardiac surgery. Proceedings of the 16th Annual Symposium on Computerized Applications in Medical Care. Washington, DC: AMIA; 1992.
8. Hudson DL, Cohen ME, Lammers RK. Use of a hybrid expert system to predict wound infections. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, editors. Proceedings of the 7th World Congress on Medical Informatics (MEDINFO 92). Amsterdam: North Holland, 1992:546–51.
9. Smith JW, Everhat JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the 12th Annual Symposium on Computerized Applications in Medical Care. New York: IEEE Press, 1988:261–5.
10. Ravdin PM, Clark GM. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. Breast Cancer Res Treat 1992;22:285–93.
11. Ferrer Salvans P, Alonso Valles, Osorio Gullon A, Saures Menendez M, Vilaplana Birba J, Rubio Garcia R, Valino Blanco J. An empirical comparison of backpropagation and multivariate discrimanant analysis to the prediction of colon cancer recurrence after surgical treatment. Proceedings of the International Neural Network Conference (INNC 90); 1990 July; Paris. Dordrecht: Kluwer, 1990:348.
12. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115:843–8.
13. Sabbatini RME. Artificial neural networks in biology and medicine [dissertation]. Campinas, Brazil: State Univ. of Campinas, 1993 June.
14. Beck JR, Schultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Path Lab Med 1986;110:13–20.
15. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. Med Informat 1990;15:229–36.
16. Omlin CW, Giles CL, Miller CB. Heuristics for the extraction of rules from discrete-time recurrent neural networks. Proceedings of the International Joint Conference on Neural Networks. New York: IEEE Press, 1992;1:33–8.
17. Cios KJ, Liu N. A machine learning method for generation of a neural network architecture: a continuous ID3 algorithm. IEEE Trans Neural Networks 1992;3:280–91.
18. Spackman KA. Combining logistic regression and neural networks to create predictive models. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. Washington (DC): AMIA, 1992.
19. Knaus WA, Draper EA, Wagner D, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818–29.
20. Romeo F, Pelliccia F, Cianfrocca C, et al. Determinants of end stage idiopathic dilated cardiomyopathy: a multivariate analysis of 104 patients. Clin Cardiol 1989;12:387–92.
21. Kuroda T, Shiina A, Suzuki O, et al. Prediction of prognosis of patients with idiopathic dilated cardiomyopathy: a comparison of echocardiography with cardiac catheterization. Jpn J Med 1989;28:180–8.
22. Zanchetta M, Pedon L, Carlon R, Franceschetto L, Maiolino P. Cardiomiopatia dilatativa. Analisi discriminante multivariata dei principali indici emodinamico-angiografici. G Ital Cardiol 1990;20:15–9.
23. Doi YL, Chikamori T, Tukata J, et al. Prognostic value of thallium-201 perfusion defects in idiopathic dilated cardiomyopathy. Am J Cardiol 1991;67:188–93.
24. Otsuka H. Factors to determine the survival of idiopathic dilated cardiomyopathy: selecting candidates for cardiac transplantation. Nippon Kioku Geka Gakkai Zasshi 1989;37:318–25.
25. Tanganelli P, Di Lenarda A, Bianciardi G, et al. Correlation between histomorphometric findings on endomyocardial biopsy and clinical findings in idiopathic dilated cardiomyopathy. Am J Cardiol 1989;64:504–6.
26. Ohnishi Y, Ynoue T, Fukuzaki H. Value of the signal-averaged electrocardiogram as a predictor of sudden death in myocardial infarction and dilated cardiomyopathy. Jpn Circ J 1990;54:127–36.
27. Ferrer Salvans P, Alonso Valles L, Osorio Gullon A, et al. An empirical comparison of backpropagation and multivariate discriminant analysis to the prediction of colon cancer recurrence after surgical treatment. Proceedings of the International Neural Network Conference (INNC 90); 1990 July; Paris. Dordrecht: Kluwer, 1990:348.