ASSESSMENT OF DIAGNOSTIC FEATURES IN THE CORONARY ARTERY DISEASE (CAD) BY APPLICATION OF STATISTICAL METHODS AND NEURAL NETWORKS

KRYSTYNA STANISZ-WALLIS[1], ANDRZEJ IZWORSKI[3],
ALDONA DEMBIŃSKA-KIEĆ[2], RYSZARD TADEUSIEWICZ[3]
AND TOMASZ LECH[3]

[1] *Department of Bioinformatics and Telemedicine,*
Collegium Medicum Jagiellonian University,
Kopernika 17, 31-501 Cracow, Poland
mywallis@cyf-kr.edu.pl

[2] *Department of Clinical Biochemistry,*
Collegium Medicum Jagiellonian University,
Kopernika 15, 31-501 Cracow, Poland

[3] *Laboratory of Biocybernetics, Dept. of Automatics,*
AGH University of Science and Technology,
Al. Mickiewicza 30, 30-059 Cracow, Poland

(Received 12 September 2003; revised manuscript received 18 February 2004)

Abstract: The present work is aimed at comparing the effectiveness of two different methods of risk factor assessment used for prediction of the CAD (coronary artery disease): the logistic regression method and the application of artificial neural networks. The former is widely used in medical research, while the latter is relatively rare. In the logistic regression method hierarchical analysis was employed to select the significant variables of the classification process. In the neural network approach several strategies were proposed for selection of the discriminative variables, all based on weight analysis in the constructed networks. Both methods have produced a consistent set of discriminative variables (Glu0, Ins0, Ins30, BMI, apoA1 and HDL-Ch), belonging to three groups of risk factors associated with insulin resistance, obesity and lipid disorders.

Keywords: coronary artery disease, logistic regression method, neural networks

1. Introduction

Diseases of the cardiovascular system, in particular the coronary artery disease, are among the most frequently observed diseases. An essential component in the diagnosis and therapy of these diseases is the early detection of the so-called risk factors, *i.e.* internal and external symptoms directly related to the occurrence of the disease. In a mathematical description, the risk factors play the role of diagnostic variables.

A widely applied tool for pre-selection of the diagnostic variables is logistic regression. One of its possible applications is variable discrimination between groups of healthy and ill patients on the basis of data concerning the concentration of lipids and proteins in the serum, or in isolated protein subfractions, and the concentration of glucose and insulin in the patient's blood. A competitive method to determine the diagnostic variables is the application of artificial neural networks. At present it is already a mathematically formalized technique for data processing and analysis, mainly for classification purposes.

Logistic discriminative functions are widely applied in statistical analyses. They are commonly applied in the elaboration of epidemiological data in the process of risk assessment of coronary artery disease development. It happens, however, that the collected data do not fulfill the mathematical conditions providing the basis for logistic regression, what may lead to erroneous results. Therefore, the application of other methods is beneficial.

Artificial neural networks have been used, with good results, as a basic tool in medical diagnosis. The CAD diagnostics seem to be particularly suitable for automation, because the decision-making process is strongly based on quantitative data (results of biochemical analyses), with minor contributions from descriptive data and symptoms that can be expressed in qualitative categories only. An advantage of neural networks is also their reference to models that can easily describe non-linear dependencies. Neural networks offer tools to control complex problems of multidimensionality, which considerably hinder attempts at modelling non-linear functions with a great number of independent variables when other methods are applied. Therefore, the artificial neural network technique has been used as a method complementary to logistic regression.

In the present work the authors compare the results of both methods obtained from their application to the selection of diagnostic features which can be used for prediction of the coronary artery disease.

2. Research material

A database collected at the Department of Clinical Biochemistry, Collegium Medicum, Jagiellonian University, has been used in the study. For all the patients (both healthy and ill) the composition of serum and lipoprotein subfractions has been analyzed.

Two groups have been distinguished:

Group I (the ill group) consisted of 95 men, patients of the Department of the Coronary Artery Disease, Collegium Medicum, Jagiellonian University, for whom a coronarography examination has confirmed the presence of the CAD, indicating the presence of vasoconstriction in one, two or three coronary veins;

Group II (the healthy group/reference group) has been formed of 133 men without clinical symptoms (electrocardiography) of coronary vein failure.

The study included a composition analysis of the serum and lipoprotein subfractions. The concentration of cholesterol (Ch), free cholesterol (fCh), cholesterol esters and triglycerides (Tg) has been determined in the serum and the isolated lipoprotein subfractions: VLDL, LDL and HDL. The protein contents has been also determined in the VLDL, LDL and HDL subfractions, as well as the concentration of B apolipoprotein in the serum and the LDL subfraction and the concentration of A1 apolipoprotein in the serum and the HDL subfraction. The determination of glucose and insulin concentrations in the blood's serum has been done four times during oral glucose tolerance tests (OGTT). The body mass index (BMI [kg/m^2]) has also been calculated for the examined persons. The determined features have been collected in Table 1.

Table 1. The set of analyzed diagnostic variables

Tg	LDL-tg	HDL-tg	VLDL-tg	Glu0	Height
Ch	LDL-Ch	HDL-Ch	VLDL-Ch	Glu30*	Mass
fCh	LDL-fCh	HDL-fCh*	VLDL-fCh*	Glu60*	BMI
eCh	LDL-eCh	HDL-eCh	VLDL-eCh	Glu120	
apoB*	LDL-B	HDL-B	VLDL-B*	SumGlu*	
apoA1	LDL-apoB	HDL-apoA1*	VLDL-apoB	Ins0*	
				Ins30	
				Ins60*	
				Ins120	
				SumIns	

The data has been divided at random into two groups: a learning set, consisting of 160 cases, and a test set of 68 cases.

3. The logistic regression model

Logistic regression is a good tool to predict disease occurrence on the basis of biochemical data (concentrations of lipids and proteins in the serum and isolated subfractions). In such analysis the dependent variable can take two values: 0 – absence of the disease, or 1 – its occurrence.

The logistic regression model is described by the following equation:

$$p(Y=\text{CDA occurrence})=\exp(b_0+b_1x_1+\ldots+b_nx_n)/(1+\exp(b_0+b_1x1+\ldots+b_nx_n)),$$

where $x_1,\ldots,x_n$ are independent variables, b_0 – the regression constant, $b_1,\ldots,b_n$ – regression coefficients for individual independent variables, Y – a dependent variable, $p(Y)$ – determines the probability of the CDA's occurrence.

The independent variables used were the concentrations of lipids and proteins and the BMI index.

The statistical significance of these features (independent variables) is usually determined by testing the regression coefficient, b_i (Wald's test). The b_i coefficient for a particular feature provides an interpretation of the feature's influence on the disease's manifestation.

During the application of the logistic regression method, the data from the learning set have been used to construct hierarchical models and select features which exhibit the strongest influence on the CAD's development. Consecutive variables were added to the model and then the quality of the new model was tested by checking whether it better fitted the data. The final model was verified by its application to the test set. The variables selected in these models played a discriminative role. The sensitivity and specificity of individual variables were calculated for both the learning and the testing sets.

Statistically significant, or discriminative, variables obtained for each model have been collected in Table 2.

Table 2. The most significant diagnostic variables

Model 1	Model 2	Model 3	Model 4	Model 5
Glu0	Glu0	Glu0	Glu0	Glu0
Ins120	Ins120	Ins30	Ins0	Ins120
HDL-apoA1	apoA1	apoA1	VLDL-tg	HDL-Ch
BMI	BMI		apoA1	BMI

The final classification results (sensitivity and specificity) for the five models have been collected in Table 3.

Table 3. Classification quality for the learning and testing sets

set		Model 1	Model 2	Model 3	Model 4	Model 5
learning	Sensitivity	73.21%	81.13%	75.47%	81.13%	68.15%
	Specificity	84%	85.92%	91.36%	88.89%	76.25%
testing	Sensitivity	47.04%	93.94%	92.59%	77.78%	59.26%
	Specificity	93.94%	81.48%	90.63%	94.28%	87.88%

4. The neural networks model

In the present study a triple-layer, unidirectional network with a sigmoidal activation function was applied, which was learned by the error back-propagation method. Various combinations of the network's structure and the parameters of the learning process were considered within the accepted network architecture. The number of neurons in the input layer was equal to the number of input features of the network, which was $N = 38$. An additional threshold value signal (bias) was fed to the network's input and the input data were normalized before feeding to the network's input layer.

The number of neurons in the hidden layer was variable (mostly 2–6); the simulations allowed its reduction to two neurons, without a reduction in recognition quality. The output layer consisted of merely one neuron, transmitting only two possible signals: 1 or 0 (where 0 means "no" with respect to the presence of the CAD, while 1 means "yes"). The consecutive layers of the network were connected according to the full connection method.

In the course of the study the parameters of the learning process were also varied. Classification error at the level of $\varepsilon = 2\%$ was taken as the criterion for completeness of the network's learning. The criterion for the correctness of response of the output neurons was error of less than 30%; the respective level defines the classification of the neuron's responses to the binary signals (which means that the output level above 0.7 should be treated as 1, while the output level below 0.3 should be treated as 0).

The network chosen for the final analysis was the network model exhibiting the best classification features (both for ill and healthy patients) and the highest sensitivity and specificity. The recognition quality was 94.12%. The system used 38 input variables. The learning set consisted of 160 observations. The network was tested on the set of 68 observations: 93.55% of ill patients (sensitivity) and 94.59% of healthy patients (specificity) were classified correctly

The analysis of the neural network's performance has shown that one of the hidden neurons is activated more strongly than the others, thus indicating the network's decisions. It can be said that connections with higher weight values between the previous layer and the given neuron have a stronger influence on its state than connections with lower weight values. Consequently, by determining connection weights one can detect the neurons of the previous layer which mostly affect the state of the output neuron and then, by backtracking, find the input signals which have the greatest influence on the network's decisions. In the described network model, each of the two hidden neurons (labeled as n1 and n2) seems to influence the single neuron of the output layer. The extent of this influence depends on the weight values of connections between the hidden layer neurons and the output layer neuron. Positive values should be treated as an indication that the influence activates the output neuron, while negative values represent an influence that blocks the neuron. A reduction in dimensionality can be achieved by eliminating the input features for which the weights of the connections with the hidden neurons have small absolute values.

Initially, absolute values of weights were calculated for connections with neurons n1 and n2. Then the maximum and minimum absolute weight values were determined (min|weight|, max|weight|) and the sum of absolute weight values was calculated for a particular feature.

Three criteria have been defined and practically introduced, that can be used to remove some input features. One is the criterion taking into account the minimum weight values for neurons n1 and n2. The absolute values of both weights were analyzed for a given input feature and the lower of them was selected. Such an analysis was carried out for all the input features; next, the values of the smallest weight were arranged in the ascending order – in short, the method has been tagged as the MIN method.

The results obtained from the MIN method have been collected in Table 4. In the first column of the table the absolute values of the minimum weight are listed for individual features; the whole table has been ordered according to this value. In further columns one can find the feature's name, its ordinal number (location in the input record), and the results obtained for the network's recognition quality after elimination of the considered feature (and with all features included – the first row).

Table 4. Recognition quality after elimination of the consecutive features according to the MIN criterion

Minimum of the two weights		Para-meter number	Reference set					Testing set				
			Reco-gnition [%]	Number of correct healthy recognitions to the total number of healthy patients		Number of correct ill recognitions to the total number of ill patients		Reco-gnition [%]	Number of correct healthy recognitions to the total number of healthy patients		Number of correct ill recognitions to the total number of ill patients	
	Whole network		100	96/96	100%	64/64	100%	94.12	35/37	94.59	29/31	93.55
9.65115	HDL-fCh	27	78.75%	96/96	100%	30/64	46.88%	82.35%	36/37	97.3%	20/31	64.52%
7.56459	Glu0	1	42.5%	4/96	4.167%	64/64	100%	50%	3/37	8.108%	31/31	100%
6.93517	Height	37	78.12%	86/96	89.58%	39/64	60.94%	76.47%	31/37	83.78%	21/31	67.74%
6.58936	Ins30	6	65.62%	92/96	95.83%	13/64	20.31%	58.82%	35/37	94.59%	5/31	16.13%
5.56136	HDL-eCh	28	48.12%	22/96	22.92%	55/64	85.94%	54.41%	6/37	13.22%	31/31	100%
5.23277	LDL-eCh	23	40%	0/96	0%	64/64	100%	45.59%	0/37	0%	31/31	100%
4.22019	apoA1	33										
3.74515	VLDL-eCh	18										

Table 5. The most significant features selected by the MIN2, MAX2 and SUM2 methods (features typed in boldface are not common)

The most significant features for the respective criteria		
MIN	MAX	SUM
Ins30	Ins30	Ins30
HDL-eCh	LDL-eCh	LDL-eCh
Height	eCh	eCh
Glu0	HDL-eCh	HDL-eCh
HDL-fCh	VLDL-B	VLDL-B
	Ins0	Ins0
	Glu30	Glu30
	Glu120	Glu120
	HDL-fCh	HDL-fCh
	Glu0	Glu0
	LDL-fCh	**apoA1**
	Height	Height

In the second criterion, the maximum values of weights for the n1 and n2 neurons are taken into account. The absolute values of both weights are analyzed for a given feature and the higher of them is selected. Such analysis has been carried out for all the features and then the values of the highest weight are arranged in the descending order - in short, the method has been tagged as MAX.

In the third criterion, the algebraic sum of maximum absolute values of both weights for neurons n1 and n2 is taken into account. This operation is carried out for all the features and then the sum values are ordered from the lowest to the highest value – in short, the method has been tagged as SUM.

By application of the original MIN, MAX and SUM methods to validate diagnostic features, described in [1, 2], the authors have able to determine sets of features used by the network in the classification process. The obtained results are presented in Table 5.

In the final stage of the analysis, the set of input variables was considerably reduced with no significant loss in the level of recognition quality for both healthy and ill patients. For the reduced number of input variables (12 variables) a new input variable network model was built. Then one more model was built – a "discriminative" model, in which some of the input variables were replaced by well-known risk factors. The replacement took place because the new variables were highly correlated with the values taken into account in the original network model.

In Table 6, the most important diagnostic variables are listed, obtained as a result of the described strategy (Strategy I) and by application of the "discriminative" model.

Final results for the classification quality of both strategies have been collected in Table 7.

5. Summary and discussion

The input variables that have turned out to be the significant features discriminating between the groups of healthy and ill patients can be divided into three groups of features for both methods (Figure 1).

Table 6. The most important diagnostic variables determined by application of Strategy I and the "discriminative" model

Strategy I	"Discriminative" model
Glu0	Glu0
Glu30	Glu30
Glu120	Glu120
Ins0	Ins0
Ins30	Ins30
eCh	Ch
VLDL-apoB	VLDL-apoB
LDL-eCh	LDL-Ch
HDL-fCh	HDL-fCh
HDL-eCh	HDL-Ch
apoA1	apoA1
Height	BMI

Table 7. Classification quality for the testing set for Strategy I and the "discriminative" model

	Strategy I	"Discriminative" model
The number of input variables	12	12
Recognition quality	92.65%	94.12%
Sensitivity	93.55%	96.77%
Specificity	91.1%	91.89%

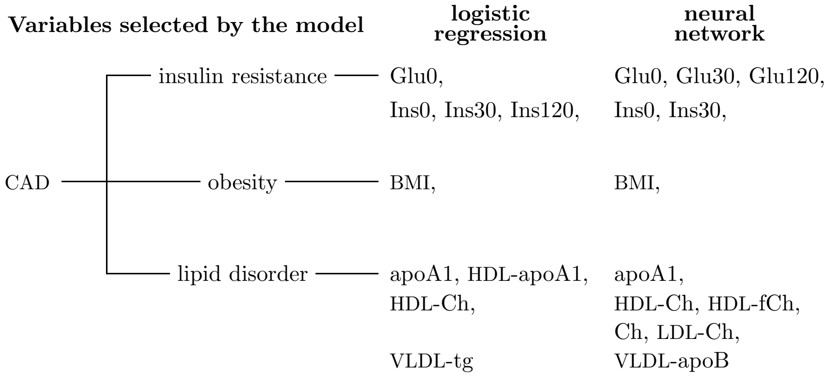
Variables selected by the model
logistic regression
neural network
CAD
insulin resistance
Glu0, Ins0, Ins30, Ins120,
Glu0, Glu30, Glu120, Ins0, Ins30,
obesity
BMI,
BMI,
lipid disorder
apoA1, HDL-apoA1, HDL-Ch, VLDL-tg
apoA1, HDL-Ch, HDL-fCh, Ch, LDL-Ch, VLDL-apoB

Figure 1. Schematic division of risk factors for the development of the coronary artery disease in both methods

In the results obtained from the *logistic regression* model, there is a variable in the lipid disorder group determining the cholesterol concentration in the HDL subfraction (HDL-Ch). This variable is an essential, independent indicator of a high risk of the coronary artery disease's development. This is confirmed by the Framingham study [3, 4].

In the present study, the apoA1 level in combination with HDL-Ch concentration is an important factor associated with a lowered risk of CAD development [5, 6]. The selection of the concentrations of HDL-Ch and the apoA1 protein as highly discriminative variables (CAD risk factors) has been obtained from the artificial neural network model. This is an important element confirming the effectiveness of that method.

The LDL-Ch feature has been included among the discriminative variables of the neural network model, what may indicate that this method is more sensitive, as it can also recognize non-linear dependencies. The simultaneous selection of both HDL-Ch and LDL-Ch concentrations is consistent with the results of recent works, in which the authors have noticed that medicines lowering the LDL-Ch concentration without raising the HDL-Ch concentration are not effective in CAD prevention [7].

The results obtained from the **neural network method** prove that in addition to such well-known discriminating factors as glucose concentration (Glu0, Glu120), insulin concentration (Ins0, Ins30), the total cholesterol concentration (Ch) and its concentrations in the HDL (HDL-Ch) and LDL (LDL-Ch) sub-fractions, the concentration of apoA1 and the BMI index, some additional input variables should be taken into account in the construction and learning process of a properly classifying network, like the concentration of apoB apolipoprotein in the VLDL sub-fraction (VLDL-apoB) and free cholesterol in the HDL sub-fraction (HDL-fCh). The (HDL-fCh) variable is highly correlated with HDL-Ch, and a simultaneous selection of these two variables by the network may indicate that by doing so the network finds a way to enhance the information signal for particularly significant variables.

The study described in the present work was the first to employ both the logistic regression method and the artificial neural networks technique for coronary artery disease risk assessment based on the concentrations of lipids and proteins in the blood's serum and its isolated subfractions. The obtained results confirm the usefulness of artificial neural networks for determination of the CAD's metabolic predictors.

A properly designed and learned neural network is able to evaluate which of the biochemical variables (the concentrations of lipoproteins and their proteins, glycaemia, insulaemia) are important for a correct diagnosis.

References

[1] Stanisz-Wallis K, Izworski A, Lech T and Dembińska-Kieć A 2001 *Proc. 12th Conf. Biocybernetics and Biomedical Engineering*, Warsaw, Poland, pp. 800–804 (in Polish)

[2] Stanisz-Wallis K, Lech T, Izworski A, Kwaśniak M, Dembińska-Kieć A 2000 *Proc. 5th Conf. Polish Neural Network and Soft Computing*, Zakopane, Poland, pp. 574–579

[3] Gordon T, Kannel W B and Castelli W P 1981 *Arch. Intern. Med.* **141** 1128

[4] Gotto A M 2001 *Circulation* **1** 59

[5] Assman G 2001 *Am. J. Cardiol.* **87** (5A) 2B

[6] Boden W E *Am. J. Cardiol.* **86** (12A) 19L

[7] Wierzbicki A S and Mikhailidis D P 2002 *Curr. Med. Res. Opin.* **18** (1) 36