







Association of Cardiovascular Events and Blood Pressure and Serum Lipoprotein Indicators Based on Functional Data Analysis as a Personalized Approach to the Diagnosis

N. G. Plekhova^{1,2} , V. A. Nevzorova^{1,2} , T. A. Brodskay¹,
K. I. Shakhgeldyan^{3,4} , B. I. Geltser^{3,4} , L. G. Priseko¹,
I. N. Chernenko¹, and K. L. Grunberg¹

¹ Pacific State Medical University,
2 Ostryakova Ave, Vladivostok 690002, Russia
pl_nat@hotmail.com.com

² Institute of Chemistry, Far Eastern Branch of the Russian Academy
of Sciences, 159 Pr. 100th Anniversary of Vladivostok,
Vladivostok 690022, Russia

³ Vladivostok State University of Economics and Service, 41 Gogolya St.,
Vladivostok 690014, Russia

⁴ Far East Federal University, Sukhanova St. 8, 690091 Vladivostok, Russia

Abstract. The development of trends and practice-oriented approaches to personalized programs for the diagnosis and correction depending on the clinical and phenotypic variants of the person is relevant. A software application was created for data mining from respondent profiles in a semi-automatic mode; libraries with data preprocessing were analyzed. The anthropometric measurements and serum lipoprotein spectrum of 2131 volunteers (average age 45.75 ± 11.7 years) were studied. To estimate the association of blood pressure and cardiovascular events markers was carried out by means of multivariate analysis of data by the methods of selection and classification significant signs. The machine learning was used to predict cardiovascular events. Depends on gender there was found the significant difference in atherogenic index of plasma (AIP) ($F < 0.05$). In young women (20–30 y.o.), the lipoproteins did not correlate with the presence of hypertension, whereas for older women the statistically significant markers were higher, such as cholesterol (CH, $F = 0.03$), low-density lipoproteins (LDL, $F = 0.03$) and AIP ($F = 0.02$). In men for identifying the risk of hypertension developing lipoproteins should be considered depending on age. Accuracy of the risk recognition for the cardiovascular disease (CVD) model was more than 89% with an average confidence of the model in each forecasted case of 90%. The markers for diagnosing the risk of CVD, the following indicators can be used according to their degree of significance: AIP, CH and LDL. Thus, the data obtained indicate the importance of risk factor phenotyping using anthropometric markers and biochemical profile for determining their significance in the top 17 predictors of CVD. The machine learning provides CVD prediction according to standard risk assessments.

Keywords: Machine learning · Cardiovascular diseases · Arterial hypertension

1 Introduction

Cardiovascular disease (CVD) associated with atherosclerosis is the main cause of adult mortality in both economically developed and developing countries. In the development and progression of CVD, the accompanying criteria, called risk factors (RF), play a leading role. Today, more than 200 RF of the CVD are known, and their number annually increases [1, 2]. RFs are divided into two subgroups: non-modifiable, impossible to influence, and modifiable, amenable to both multimodal behavioral interventions and medical therapy. Moreover, it is necessary to determine the total cardiovascular risk that means – the probability of developing a cardiovascular event connected with atherosclerosis over a specific period. This is the key to selecting preventive strategies and specific interventions for patients. The prevention and management of CVD increasingly demand effective diagnostic testing. Consensus defines a diagnostic as a method and an associated device that performs a physical measurement from a patient or associated biological sample and produces a quantitative or descriptive output, known as a biomarker. The definition of a biomarker, in turn, encompasses “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [3]. Diagnostics, because of their strategic position at the intersection between patients and their clinically actionable data, directly affect the patient experience and the quality of care that individuals receive. The methods of the biochemical and cellular biofluid analysis advanced, the portfolio of available tests expanded and central laboratories emerged to standardize sample acquisition and measurement [4]. Today, technology is expanding the number of diagnostic tests that can reach beyond the walls of centralized laboratories and back to the point of care for use across a broad range of clinical settings.

Established risk factors for CVD – such as AH, high levels of low-density lipoprotein cholesterol (LDL-C), low levels of high-density lipoprotein cholesterol (HDL-C), smoking, male gender, and old age do not entirely account for CVD risk [5]. Since treating modifiable risk factors is known to reduce the risk of CVD [6, 7], improving CVD risk stratification would enable better allocation of prevention resources [8]. And one approach to improving risk prediction is to consider the risk of CVD associated with the size distribution of a patient’s lipoprotein particles. The Atherogenic Index of Plasma (AIP) is easily calculated from a standard lipid profile. It is a logarithmically transformed ratio of molar concentrations of triglycerides (TG) to high-density lipoprotein cholesterol (HDL). The strong correlation of AIP with lipoproteins may explain its high predictive value [9]. However, the determination of TC, HDL, and low-density lipoprotein cholesterol (LDL) concentrations is not sufficient for appropriate medical therapy. LDL and HDL should be sub-fractionated to measure concentrations of large, anti-atherogenic HDL [10]; less atherogenic HDL [11]; and small, atherogenic particles of HDL [12], as well as less atherogenic LDL [9] and atherogenic LDL lipoproteins component – ApoA component – ApoB [13].

Again, the basis of a specialized medical information system consists of the instrument-computer complexes. Such use of a computer in combination with measuring technology in clinical and laboratory diagnostics allows creating new effective

means for providing automated information, collection on a patient's condition and processing of this data in real-time. Data used for medical diagnosis have several features, such as the qualitative nature of information, the presence of data gaps, a large number of variables with a relatively small number of observations [14, 15]. Moreover, the significant complexity of the object of observation (disease) often does not allow building even a verbal description of the doctor's diagnosis procedure. The creation of medical device-computer complexes allows us to approach from the new positions to the understanding of instrumental diagnostic methods in the cumulative consideration of all parameters to establish an accurate diagnosis [16]. Traditionally, when modeling the course of the disease, probabilistic prediction of the value of binary variables is used [15], which is made based on regression analysis or using automated systems built based on neural network analysis. The optimal approach is the use of a scoring scale, such as, for example, the SCORE cardiovascular risk assessment system, the Framingham scale or the mathematical model of PROCAM [1, 17, 18]. The disadvantage of this approach is that the scales known today have been developed regarding a certain population and nosological forms of the disease. The important direction in this area is developing a more universal scale, with a specific application for analyzing any feature, interested to the researcher, described by the binary variable.

Objective: To assess the prospects of using artificial intelligence technologies in predicting the outcomes and risks of cardiovascular diseases in patients with hypertension.

2 Materials and Methods

2.1 Study Design

A cross-sectional study of the impact of factors on the development of CVD is a prospective population study made on a representative sample of the population of Vladivostok in the multicenter observational Russian program Epidemiology of Cardiovascular Diseases (ESSE-RF) in Primorsky Krai to identify the prevalence and frequency of chronic noncommunicable Diseases and their attendant risk factors. The data were published earlier [19]. The procedures were approved by the Ethics Committee of the Pacific State Medical University (agreement no. 46/23.11.2014). Written informed consent was obtained from all the subjects. Beginning with 2014, three basic surveys (2012–2019) with follow up 3 observation phases were conducted at intervals of 2 years. Since people from 24 to 65 years old were included in the ESSE-RF study, volunteers aged 20–23 ($n = 245$) were additionally examined whose survey protocol was closely match to the ESSE-RF. The 901 volunteers (502 women and 399 men) 692 healthy individuals of them were included in the first representative sample and patients in the second, with a diagnosed arterial hypertension (AH, 209 people). The age of the volunteers was from aged 20 to 44 years, which according to the WHO classification corresponded to young age.

In research, the method of questioning and clinical research, including anthropometric, instrumental (investigation of arterial pressure, pulse, ECG recording, were used. In this group anthropometric parameters such as height, body weight, body mass

index (BMI), and waist circumference were monitored. To determine the body weight standard stand-on scales were used. The height measurement was performed using altimeter, the measured person was always without shoes. BMI was calculated as the ratio of body weight in kg to the squared height in meters. The waist circumference was measured at half the distance between the bottom edge of the lower rib and iliac crest of the hip bone at a horizontal level. Waist circumference values were defined according to the cardio-metabolic risk. There is a moderate risk in waist circumference >94 cm, possibly 80 cm (risk level 1), and a high risk (risk level 2) in waist circumference higher than 102 cm in men and 88 cm in women.

For the biochemical examination, following an overnight fast, blood samples were drawn into tubes and centrifuged the same day to separate serum, which was stored frozen (-80°C) for subsequent analysis. A venous blood sample was withdrawn on an empty stomach and parameters were determined in certified laboratories using standard laboratory methods. From biochemical parameters the following values were monitored: glycaemia, uric acid, total cholesterol, LDL-cholesterol, and components of low (LDL) and high-density lipoproteins (HDL), and triglycerides (TG) was carried out in a colorimetric method using an automatic biochemical analyzer Mindray BS-200 (Shenzhen Mindray Bio-Medical Electronics, Chine) and reagents from Alpha Diagnostics (San Antonio, TX, US). The atherogenic index of plasma (AIP) and the atherogenic coefficient (AC) were computed as $\log(\text{TG}/\text{HDL})$ and $\text{non-HDL}/\text{HDL}$, respectively.

2.2 Machine Learning

For the machine learning the Neural network data processing was carried out using the NeuralNetworkTool software package, which is part of the Matlab R2010b software (Mathworks, USA). The software product data was selected due to its modernity, accuracy of results and user support policy. Due to their capability to solve complex problems by manipulation of high volume data the designation, training and usage of NeuralNetworkTool requires computer environment. The network was trained according to the Bayesian regularization algorithm, since it gave the smallest error equal to 0.01. Bayesian regularization minimizes the linear combination of quadratic errors and weights. Modification is carried out in such a way that as a result a network with high generalizing properties is obtained.

2.3 Statistical Analysis

All statistical analyses were performed using SPSS Statistics 22 (IBM, Armonk, NY, United States). Testing for normality was performed by the Kolmogorov Smirnov test. Differences between group means were calculated using a two-sample t-test, assuming or not assuming equal variances (based on Levene's Test for Equality of Variances). The strength of the linear relationship between the two variables was expressed by the Pearson correlation coefficient; a p-value of <0.05 was assumed to be statistically significant. Means and standard deviations are reported in terms of the original distributions.

Step 1. χ^2 test was used to analyze the associations between data (the sex of the subjects (0, 1), the body mass index (BMI), the presence of a smoking factor (yes, no), arterial hypertension (AH, yes, no) and interval variables were included; age, systolic pressure, serum cholesterol, TG, HDL, LDL, ApoA, ApoB, C reactive protein (CRP), thyroid-stimulating hormone (TSH), leptin, adiponectin, and insulin.

Step 2. Analyze the relation of smokers and AH, gender status, age status and another. Generate receiver-operating characteristic (ROC) curve, calculate the Youden index, and found cutoff points for AH, gender status, age status.

Step 3. Binary logistic regression with backward: conditional method was performed. Dependent variables were AH, gender status, age status, and systolic pressure. Covariates were the related clinicopathological data from step1 and the AH cut-off points from step 2. Borrmann type and grading were exceptional because some cases had no available Borrmann type or grading data.

Step 4. If Borrmann type or grading factor was one of the related factors, binary logistic regression was then performed again using AH, gender status, age status or systolic pressure from step 2 as dependent variables and the related clinicopathological data from step 1 and the AH.

Step 5. If Borrmann type or grading factor was not one of the related factors obtained through binary logistic regression from step 3, the results in step 2 were the final result. However, if Borrmann type or grading factor was the related factor, the results in step 3 were the final result.

3 Result

3.1 Patient Characteristics

The average age in the group of patients with AH was 34.5 ± 2.8 years (from 25 to 44 years), the same indicator in the group of healthy individuals was 32 ± 1.7 years (from 20 to 44 years). In the group of patients with AH, there was a significant increase of indicators as blood pressure (BP), weight, and triglyceride levels compared with the control group (Table 1). Significant differences in indicators of total cholesterol, HDL, and LDL between the subjects with normal pressure and those with the presence of AH was not detected ($P \leq 0.007$ for overall trend for each variable).

3.2 Analyses of Biochemical Parameters

For a more accurate study of the distribution of the lipid spectrum, we separated the indicators of healthy subjects and patients with AH, respectively, of age and sex (Tables 2 and 3). The indicators of concentration of total cholesterol increase with age and reaches a level of 5.3 ± 0.7 mmol/L for the older age group (over 40 years) of healthy women, and patients with AH – 5.5 ± 0.5 mmol/L (Table 3). For men, a similar trend was noted (Table 2). In all age groups and regardless of gender, the average index of HDL was higher than 1.0 mmol/L for men and 1.2 mmol/L for women, which corresponds to the normal values accepted in the population (>1.0 for men and >1.2 for women). Significant differences of HDL surveyed with normal

Table 1. Comparisons of risk factors, and biochemical parameters between groups

Indicators	Group I (healthy volunteers) (n = 692)	Group II (patients with AH) (n = 209)
Age (years)	34,5 ± 2,8	32 ± 1,7
Weight (kg)	69 ± 5,8	87* ± 7,5
AP (ASP and ADP), mm Hg	118 ± 14,5	151 ± 11,5*
Heart rate, beats per minute	72 ± 5,4	80 ± 7,5
Smokers, n (%)	55 ± 4,2	92 ± 6,5
Smoking Person Index, pack/years	4,75	5,25
Total cholesterol, mmol/L	4,95 ± 0,6	5,0 ± 0,4
ApoA, gr/L	1,79 ± 0,3	1,81 ± 0,6
ApoB, gr/L	0,92 ± 0,03	0,83 ± 0,05
LDL, mmol/L	3,13 ± 0,47	3,35 ± 0,27
HDL, mmol/L	1,31 ± 0,6	1,28 ± 0,4
TG, mmol/L	1,21 ± 0,09	1,87* ± 0,06
AIP	2,26 ± 0,06	3,0 ± 0,2
AC	0,46 ± 0,02	0,55 ± 0,01

Note: values are represented as n (%) or mean ± SD/median (IQR).

Abbreviations: BMI – body mass index, AP – blood arterial pressure, ASP – arterial systolic pressure, ADP – arterial diastolic pressure (measured in sitting position using a mercury sphygmomanometer), ApoA – apolipoprotein A, ApoB – apolipoprotein B, LDL and HDL – low and high density lipoproteins, TG – triglycerides, AIP – atherogenic index of plasma, AC – atherogenic coefficient, conventional units.

* P < 0.05, compared with Group I.

pressure and individuals with the presence of AH was not detected ($P \leq 0.005$ for overall trend for each variable). Concerning LDL, there is a slight increase correlated with age. In healthy volunteers, the maximum value of the concentration LDL was found in persons of the age group 40 and more years: for men – 3.2 mmol/L, for women – 3.3 mmol/L. In persons with hypertension, this indicator did not exceed 3.8 mmol/L in men and 3.4 mmol/L in women.

The atherogenic index of plasma (AIP) and the atherogenic coefficient (AC) is not recognized by all researchers and in existing clinical guidelines, it is not included in asses of the risk of cardiovascular accidents [20]. At the same time, there are opinions about the possibility of its use to determine the risk of developing CVD. The AIP should not exceed 2.5 in healthy men between 20 and 30 years old, and 2.2 in healthy women of the same age, in people of both sexes 31–40 years old 3.0, and in people over 40 years old without clinical manifestations of atherosclerosis 3.5. We found that in men with increasing age AIP decreased. So, if in the age group from 20 to 30 years it was 2.4 ± 0.7 , then in the rest AIP had average values of 1.8 ± 0.4 and 1.1 ± 0.6 , respectively ($p < 0.05$, Table 2). While, AIP rates in men with AH in all age groups significantly exceeded those for healthy individuals ($p < 0.05$, Table 2). In women,

Table 2. Biochemical parameters of different age men's groups

Indicators	Group I (healthy volunteers, n = 298), years old			Group II (patients with AH, n = 92), years old		
	25–30 (n = 96)	31–40 (n = 112)	40 - more (n = 90)	25–30 (n = 20)	31–40 (n = 48)	40 - more (n = 24)
Total cholesterol, mmol/L	4,7 ± 0,5*	5,0 ± 0,7	5,1 ± 0,4	4,5 ± 0,5	5,4 ± 0,5	5,6 ± 0,6
LDL, mmol/L	1,3 ± 0,03	1,3 ± 0,05	1,3 ± 0,03	1,2 ± 0,04	1,2 ± 0,02	1,2 ± 0,1
HDL, mmol/L	3,1 ± 0,4	3,3 ± 0,5	3,2 ± 0,4	3,1 ± 0,35	3,7 ± 0,4	3,8 ± 0,4
AIP	2,4 ± 0,06	1,8 ± 0,06*	1,1 ± 0,6*	2,7 ± 0,05	3,3 ± 0,07	3,6 ± 0,03
AC	0,4 ± 0,03*	0,5 ± 0,15	0,5 ± 0,07	0,6 ± 0,07	0,6 ± 0,06	0,7 ± 0,5

Note: values are represented as n (%) or mean ± SD/median (IQR).

Abbreviations: LDL and HDL – low and high density lipoproteins, AIP – atherogenic index of plasma, AC– atherogenic coefficient.

* P < 0.05, compared with Group II.

Table 3. Biochemical parameters of different age women's groups

Indicators	Group I (healthy volunteers, n = 394), years old			Group II (patients with AH, n = 117), years old		
	25–30 (n = 127)	31–40 (n = 164)	40 - more (n = 103)	25–30 (n = 21)	31–40 (n = 64)	40 - more (n = 32)
Total cholesterol, mmol/L	4,6 ± 0,5*	5,0 ± 0,0	5,3 ± 0,3	4,3 ± 0,5	5,1 ± 0,4	5,5 ± 0,7
LDL, mmol/L	1,3 ± 0,04	1,4 ± 0,03	1,4 ± 0,03	1,3 ± 0,05	1,4 ± 0,03	1,4 ± 0,06
HDL, mmol/L	2,7 ± 0,5	3,2 ± 0,4	3,3 ± 0,25	2,8 ± 0,4	3,3 ± 0,35	3,4 ± 0,5
AIP	2,6 ± 0,05	2,8 ± 0,04	2,9 ± 0,05	2,7 ± 0,05	2,8 ± 0,05	2,9 ± 0,02
AC	0,4 ± 0,02	0,5 ± 0,05	0,5 ± 0,07	0,4 ± 0,07	0,5 ± 0,05	0,5 ± 0,09

Note: values are represented as n (%) or mean ± SD/median (IQR).

Abbreviations: LDL and HDL – low and high density lipoproteins, AIP – atherogenic index of plasma, AC– atherogenic coefficient.

* P < 0.05, compared with Group II.

both with normal blood pressure values and in the group with AH, the AIP values increased with age ($F < 0.05$, Table 3). Thus, with the indicator for healthy women, the rate of 2.2 in our study, AIP for this age group was 2.6 ± 0.05 , and in patients with AH 2.7 ± 0.05 . From our point of view, this fact testifies to the variability of this indicator for different regions. Probably, the “normal” value of the cholesterol coefficient for this group of women (20–30 years old) requires adjustment. As for the AIP values for older women, this indicator did not differ and did not exceed the established normal values, both in the group of healthy women and in the group of women with AH (Table 3). AC is a sensitive marker for detecting the risk of CVD and normally does not exceed 1.0

[21]. We found that this coefficient did not exceed a critical value for both healthy individuals and those with AH. The average AC values for all age groups of men and women ranged from 0.4 to 0.6 (Table 2, 3). Thus, our study showed that the values of lipid profile total cholesterol, HDL, LDL, AC in men and women do not have statistically significant differences in age and sex, whereas a significant difference between the indicators of men and women is determined only concerning the AIP ($P < 0.05$).

The direct relationship between indicators lipid profile and AH at women in the age group of 20–30 years no found ($P < 0.05$). For indicators, total cholesterol, HDL, LDL, AIP, and AC the Fisher's correlation coefficients were 0.39, 0.20, 0.32, 0.28 and 0.81, respectively. In women of 31–40 years old, a correlation was found between the indicators of LDL and AIP and the presence of AH ($F = 0.04$, $P < 0.05$). As for women in the age group of 40 years or more, significant values of the Fisher criterion were identified for three, except for HDL and AC, the values of total cholesterol, LDL and AIP and were 0.03, 0.03 and 0.02, respectively ($P < 0.05$). In men of the age group of 20–30 years, the AIP was determined as a significant marker of the presence of AH. The levels of total cholesterol, HDL and AC showed the value of the Fisher coefficient more critical. In men over 40 years old, a positive correlation was found between cholesterol ($F = 0.04$), LDL ($F = 0.04$), AC ($F = 0.05$) and AH.

3.3 Numerical Results

Optimal Scaling (CATREG) was chosen as the regression model, this model operates with categorical variables, all included interval and order predictors were categorized, taking into account categorization. The binary variable was the presence of AH, 12 out of 18 potential predictors were included in the regression model. The “importance” coefficients (importance) calculated by the regression analysis are presented in Table 4, their values are proportional to the degree of the predictor's contribution. The values of the dependent variable were calculated for each of the predictors, included in the regression model by multiplying the absolute value of the corresponding importance factor by 100 and rounding to the integers.

The two groups of healthy individuals and patients with AH were compared by the values of each of the 21 potential predictors. For nominal variables, analysis of contingency tables was used, for ordinal and interval tests the Kruskal-Wallis test was used. The result of the analysis is shown in Table 5. With a significance level of 0.05, they were reliably associated only with the dependent variable. Predictors that had a statistical relationship with the dependent variable with a significance level of $p = 0.15$ or more were then included in the regression model.

A threshold total score was determined, after which the dependent variable assumes a value with an empirical probability of developing unwanted development of CVD. To calculate the threshold score, a regression analysis was initially carried out, in which the total score of each patient served as a predictor, and the dependent variable remained the same. By definition, the dependent variable used binary logistic regression. The equation

Table 4. Statistical relationship of the dependent variable with potential predictors

n	Predictor	Group I (n = 692)	Group II (n = 209)	P
1	Gender (male 0, female 1)	n = 298 (0); n = 394 (1)	n = 92 (0); n = 117 (1)	0,148
2	Age, Med (HKВ, BKВ)	34,5 (20; 42)	32,0 (26; 44)	0,582
3	Smoking (no 0, yes 1)	n = 312 (0) n = 380 (1)	n = 17 (0); n = 192 (1)	0,763
4	Systolic pressure, mm Hg, Med (HKВ, BKВ)	118 (90; 140)	151,8 (120; 230)	0,2833
5	BMI, kg/m ² , Med (HKВ, BKВ)	22 (33; 39)	28 (32; 37)	0,099
6	Glucose, mmol/l Med (HKВ, BKВ)	5,3 (5,0; 6,0)	6,0 (4,9; 6,8)	0,072
7	Total cholesterol, mmol/L	4,9 ± 0,47	5,26 ± 0,15	0,645
8	HDL, mmol/L	1,365 ± 0,09	1,33 ± 0,085	0,047
9	LDL, mmol/L	3,27 ± 0,12	3,32 ± 0,28	0,051
10	TG, mmol/L	1,35 ± 0,38	1,4 ± 0,3	0,049
11	ApoA-I gr/L	1,785 ± 0,13	1,68 ± 0,15	0,067
12	ApoB gr/L	0,79 ± 0,01	0,82 ± 0,04	0,174
13	TH mmol/L	1,14 ± 0,1	1,39 ± 0,08	0,182
14	Leptin, ng/ml (HKВ, BKВ)	9,75 (6,7; 15,6)	14,7 (13,8; 16,7)	0,002
15	Adiponectin, µg/ml	8,38 ± 2,21	10,13 ± 3,26	0,020
16	CRP, gr/L, Med (HKВ, BKВ)	1,74 ± 0,6	1,64 ± 0,6	0,031
17	Insulin	2,97 ± 0,4	7,7 ± 0,7	0,079
18	TSH, mEd/L	1,8 ± 0,08	1,43 ± 0,2	0,161

Note: values are represented as n (%) or mean ± SD/median (IQR).

Abbreviations: BMI – body mass index, LDL and HDL – low and high density lipoproteins, TG – triglycerides, ApoA – apolipoprotein A, ApoB – apolipoprotein B, TH – thyroid hormone, CRP – C reactive protein, TSH – thyroid-stimulating hormone.

$$p = 1/1 + e^{3.698} * e^{-0.045x}$$

was obtained, where p is the theoretical probability of the presence of AH (dependent variable), and x is the value of the total score for a particular patient. We calculated with the help of this equation, the theoretical values of the probability of conjunction with AH. The scattering diagram, which reflects this dependence (see Fig. 1A).

When calculating the average probability values in the group of patients with the value of the dependent variable “no” and in the group of patients with the value of the dependent variable “yes” it turned out, that the diagnosis of AH was noted if the theoretical probability of its development was in the range from 0.124 to 0.151. The graph (see Fig. 1B) shows that the lower limit of this range (0.124) corresponds to an interval of 10 to 40 points. Moreover, the actual frequency of AH is noted at patients with a total score of 10 or less. It turned out to be equal to 4.93% (about 5%).

Table 5. The fragment of the resulting regression analysis table with the optimal scaling and the total score of the CVD risk

n	Predictor	Standardized coefficients		F	Correlations			Importance	Points
		Beta	Std. error		Zero order	Partial	Part		
1	Gender	9,823E-02	0,047	4,364	0,082	0,106	0,097	0,048	+5
2	BMI	2,298E-02	0,064	0,130	0,132	0,018	0,017	0,018	+2
3	Glucose	8,015E-02	0,049	2,709	0,129	0,084	0,077	0,062	+3
4	HDL	0,0239	0,051	2,209	0,300	0,233	0,219	0,0427	+4
5	LDL	0,111	0,047	0,544	0,144	0,118	0,109	0,015	+2
6	TG	3,514E-02	0,050	0,497	0,084	0,036	0,03	0,018	+2
7	ApoA	3,679E-02	0,063	0,346	0,102	0,030	0,027	0,022	+2
8	TH	0,141	0,051	0,683	0,099	0,014	0,013	0,017	+2
9	Leptin	6,955E-02	0,052	0,176	0,014	0,007	0,006	0,006	+1
10	ADP	0,14	0,013	0,005	0,588	0,016	0,012	0,013	+1
11	CRP	0,11	0,0013	0,043	0,088	0,02	0,009	0,011	+1
12	Insulin	0,139	0,048	0,712	0,095	0,009	0,008	0,019	+2

Abbreviations: BMI – body mass index, LDL and HDL – low and high density lipoproteins, TG – triglycerides, ApoA – apolipoprotein A, TH – thyroid hormone, ADP – adiponectin, CRP – C reactive protein.

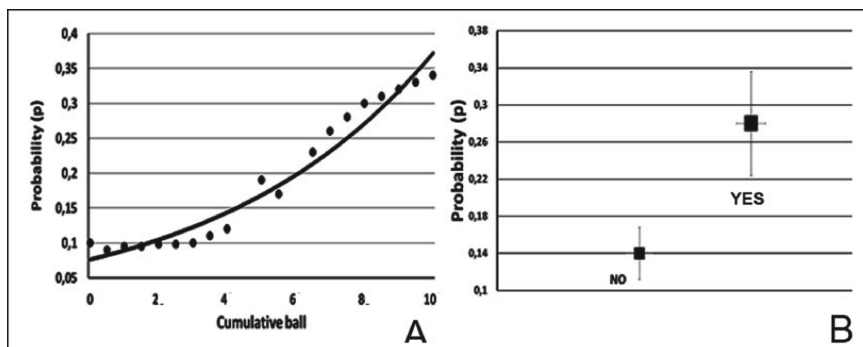


Fig. 1. The theoretical probability presence of AH. A. – the scattering diagram the dependence of the theoretical probability of the AH on the value total score (sensitivity 0.688, specificity 0.673 range from 0.124 to 0.151); B. – range of theoretical probability risk AH in healthy individuals with the absence (NO) and presence of this diagnosis (mean value ± error of mean).

The machine learning was carried out using NeuralNetworkTool software package, which is part of the Matlab R2010b software (Mathworks, USA). The network was trained according to the Bayesian regularization algorithm, since it gave the smallest error equal to 0.01. With the input-output method, a hidden layer and an output layer are created, which is a multilayer neural network. Next, the number of training set, test

set and test set is selected. The training set contains a set of data that have pre-classified target and predictor variables. To determine how well the model works with data outside the training set, a test data set or test set is used. The test set contains data obtained from the preliminary profile, but they are not used when the data from the test set passes through the model to the end, when the compared data is compared with the model results. Using the randomization function $X_{train}, X_{test}, y_{train}, y_{test} = \text{train_test_split}(X, Y, \text{test_size} = 0.40, \text{random_state} = 42)$, 2 samples were formed from the total data array: training (488 people) and test (245 people), which included data from patients with established the diagnosis of hypertension. Of all the subjects with hypertension ($n = 733$), the number of smokers was 144 people, 170 smoked and quit smoking, 41 non-smokers. As input, 17 of the most important variables were used, which constituted the input forecast layer model (Table 4, Fig. 2).

Hidden layers were determined empirically: the first layer includes 26 neurons (positions where multiplication weighting matrix and matrix input data of previous neurons), output layer consisted of 1 neuron, which corresponded with AH.

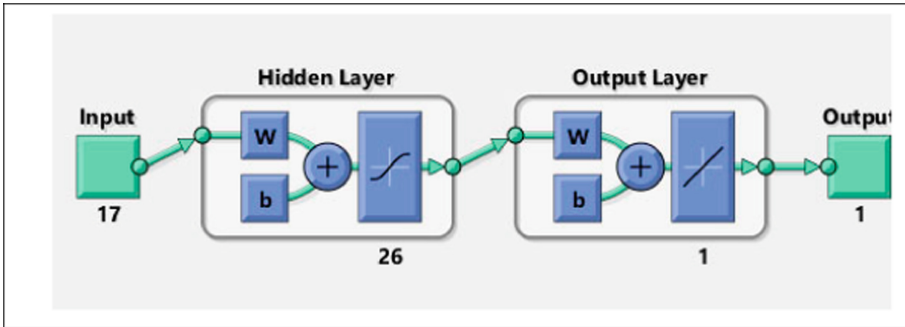


Fig. 2. Neural network model.

Training and optimization of the machine learning were carried out according to the Bayesian regularization algorithm, since it gave the smallest error equal to 0.01. Neural network testing was carried out on indicators of 288 people not included in the training sample, 144 of which were with the presence of hypertension, the rest without the presence of hypertension. The neural network predicts the presence of hypertension in young people at a given time with an accuracy of 76.06%, which is a satisfactory result. Since office blood pressure measurement can't determine availability disguised hypertension common in this cohort and in general population the probability of a correct assessment by a doctor is approximately 70% [20]. The change in the accuracy of the machine learning in the process of training and testing is presented in Fig. 3. The sample size for the machine learning was 66.6% of all subjects with hypertension. Training and optimization was carried out in 1000 eras, the volume of data submitted at a time amounted to 32 units. As a result of testing using the Bayesian regularization algorithm, the prognostic accuracy reached 97.9%, and the loss value was in the range 10^{-7} – 10^{-8} (Fig. 3). During testing, the accuracy of the network decreased to 95.5%.

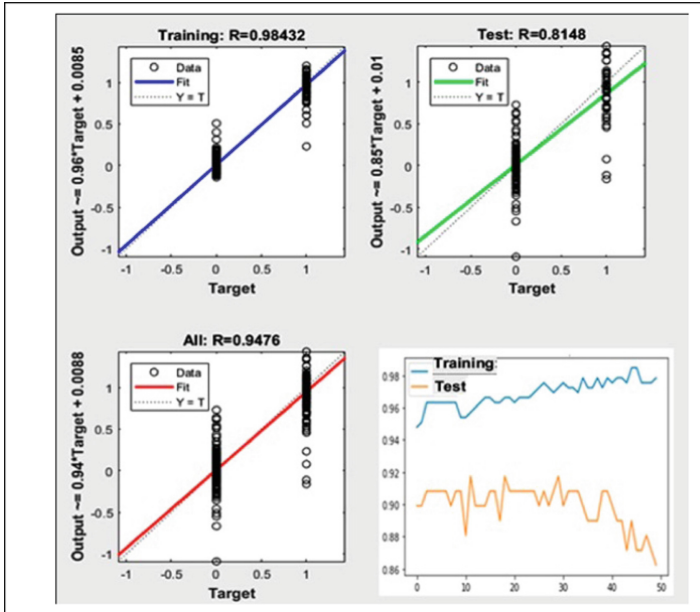


Fig. 3. Testing of neural network in the learning process.

4 Discussion

In 2012, the Ministry of Health of the Russian Federation initiated a multicenter observational study “Epidemiology of cardiovascular diseases in various regions of the Russian Federation (ESSE-RF)” to study the “traditional” and “new” risk factors CVD for implement preventive programs [22]. About 20 000 participants were included in the research: representative samples from the unorganized male and female population aged 20–64 years from 13 regions of the Russian Federation, including Primorsky Krai, in which Pacific State Medical University took part [23]. The results of this study will provide objective information on the prevalence of major CVDs in the population and predict the health of Russians. In the present study, we evaluated the significance of the atherogenic index’s blood spectrum and also revealed the degree correlation between their and AH. Depending on the gender and age of the examined persons, the most significant markers for diagnosing of AH were determined. As a result, it was found that in young men of the age group of 20–30 years, the AIP can be considered as a risk factor combined with AH. A statistically significant index of correlation is determined between LDL, AIP and AH for the age group of 31–40 years and for the age group of 41 and more years total cholesterol ($F = 0.04$), LDL ($F = 0.04$) and AIP ($F = 0,05$). In young women, no indicator of the lipoprotein blood spectrum correlated with the existence of AH. Probably, additional extended studies of this group of persons are needed. Whereas, in the age group of 31–40 years, as well as for men, LDL, AIP can serve as marker conjunction with the AH. Also, a statistically significant correlation

between the indicators of total cholesterol ($F = 0.03$), LDL ($F = 0.03$), AIP ($F = 0.02$) and AH for the age group of women 41 years and more is proved.

Thus, as markers for diagnosing the risk of CVD, in our opinion, it is possible to use the following indicators, according to their degree of significance: AIP, LDL, and total cholesterol. Moreover, even though when calculating the AIP, the values of total cholesterol and HDL were used and the latter does not correlate with the presence of AH, this indicator, is of fundamental importance concerning the serum level of cholesterol. Similar data on the selective diagnostic significance of individual indicators lipoprotein spectrum to assess the risk of developing CVD were also obtained by other researchers [24, 25].

During our further work, by constructing a regression model of risk, in which additional parameters, such as the high-density lipoprotein cholesterol level, the body mass index, the adiponectin content and another we can make individual characteristics of the specific patients to improve the accuracy of predictions [26]. These functions are multivariate algorithms that combine the information in CVD risk factors such as sex, age, systolic blood pressure, total cholesterol, high-density lipoprotein cholesterol, and smoking behavior to estimate risk developing CVD over a fixed time [27, 28].

5 Conclusion

Taking into attention about of the content certain substances norms in the body are the average values characteristic for the majority of healthy people, their correction is necessary for each case. So, patients suffering from diabetes, obesity and other diseases that usually accompany the change in lipid metabolism, are recommended to maintain the level of total cholesterol at the lowest level for the prevention of CVD, while for healthy people, these values may be slightly increased. Besides, when evaluating research data, not only the figures obtained for different indicators are important, but also their ratio among themselves. The conditional norm of the total cholesterol content is 2.97–8.79 mmol/L (for middle-aged people - up to 5.2 mmol/L) stays in a rather wide range. So, for individuals younger than 40 years, this indicator should be considered in conjunction with other factors, namely age, sex, smoking status and the values of systolic pressure. Such an approach from the position of multifactorial analysis allows diagnosing the state of lipid metabolism in healthy individuals more accurately, considering an objective assessment of the developing CVD risk for earlier treatment of lipid-regulating therapy.

The value of the coefficient for assessing the overall risk of developing CVD specific for the Russian population, allows us to estimate also the relative risk (RR) since it establishes a monotonous numerical scale, low values of which indicate low relative risk values, and high scales indicate a high relative risk. The advantage of the study is the consideration of a set of anthropometric data, the results of laboratory tests and other important predictors of CVD development. Thus, the machine learning in combination with extended phenotyping increases the accuracy of predicting cardiovascular events in the population of subjects with the presence of such developmental RF as hypertension. The developed approaches allow us to approach a more

accurate understanding of the markers of subclinical diseases without a priori assumptions about the causality of their occurrence.

Acknowledgments. The study was supported by a grant from the Russian Foundation for Basic Research 19-29-01077 and is part of the Ministry Health the Russian Federation state task «Clinical and phenotypic variants and molecular genetic features of vascular aging in people of different ethnic groups».

Declaration of financial and other relationships. All authors participated in the development of the concept, the design of the study and the writing of the manuscript. The final version of the manuscript was approved by all authors.

References

1. Boersma, E., Pieper, K.S., Steyerberg, E.W., Wilcox, R.G., Chang, W., Lee, K.L., Akkerhuis, K.M., Harrington, R.A., Deckers, J.W., Armstrong, P.W. et al.: Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation. Results from an international trial of 9461 patients. *Circulation* 101(22), 2557–2567 (2000)
2. Pollack Jr., C.V., Sites, F.D., Shofer, F.S., Sease, K.L., Hollander, J.E.: Application of the TIMI risk score for unstable angina and non-ST elevation acute coronary syndrome to an unselected emergency department chest pain population. *Academic. Emergency Med.* 13(1), 13–18 (2006)
3. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Biomarkers definitions working group. Downing GJ, ed. *Clin. Pharmacol. Ther.* 69, 89–95 (2001)
4. Rosenfeld, L.: Clinical chemistry since 1800: growth and development. *Clin. Chem.* 48, 186–197 (2002)
5. Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B.: Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18), 1837 (1998)
6. Thompson, P.D., Buchner, D., Pina, I.L., Balady, G.J., Williams, M.A., Marcus, B.H., et al.: Exercise and physical activity in the prevention and treatment of cardiovascular disease. *Circulation* 107, 3109–3116 (2003)
7. Stone, N.J., Robinson, J.G., Lichtenstein, A.H., Bairey Merz, C.N., Blum, C.B., Eckel, R.H., et al.: American college of cardiology/American heart. *Circulation* 25(Suppl 2), S1–45 (2014)
8. Superko, H.R., King, S. Lipid management to reduce cardiovascular risk: a new strategy is required. *3rd. Circulation* 117(4), 560–568 (2008)
9. Dobiášová, M.: AIP-atherogenic index of plasma as a significant predictor of cardiovascular risk: From research to practice. *Vnitr. Lek.* 52, 64–71 (2006)
10. Pearson-Stuttard, J., Bandosz, P., Rehm, C.D., Afshin, A., Peñalvo, J.L., Whitsel, L., Danaei, G., Micha, R., Gaziano, T., Lloyd-Williams, F., et al.: Comparing the effectiveness of mass media campaigns with price reductions targeting fruit and vegetable intake on US cardiovascular disease mortality and race disparities. *Am. J. Clin. Nutr.* 106, 199–206 (2017)

11. Perk, J., De Backer, G., Gohlke, H., Graham, I., Reiner, Z., Verschuren, M., Albus, C., Benlian, P., Boysen, G., Cifkova, R., et al.: European guidelines on cardiovascular disease prevention in clinical practice (version 2012). The fifth joint task force of the European society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). *Eur. Heart J.* **33**, 1635–1701 (2012)
12. Roth, G.A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S.F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., et al.: Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**, 1–25 (2017)
13. Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., McQueen, M., Budaj, A., Pais, P., Varigos, J., et al.: Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART Study): Case-control study. *Lancet* **64**, 937–952 (2004)
14. Perez, L., Dragicevic, S.: An agent-based approach for modeling dynamics of contagious disease spread. *Int. J. Health Geogr.* **8**(1), 50–54 (2009)
15. Hernández, A.I., Le Rolle, V., Defontaine, A., Carrault, G.A.: Multiformalism and multiresolution modelling environment: application to the cardiovascular system and its regulation. *Philos. Transact. Math. Phys. Eng. Sci.* **367**(1908), 4923–4940 (2009)
16. Antman, E.M., Cohen, M., Bernink, P.J., McCabe, C.H., Horacek, T., Papuchis, G., Mautner, B., Corbalan, R., Radley, D., Braunwald, E.: The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. *J. Am. Med. Assoc.* **284**(7), 835–842 (2000)
17. Eagle, K.A., Lim, M.J., Dabbous, O.H., Pieper, K.S., Goldberg, R.J., Van de Werf, F., Goodman, S.G., Granger, C.B., Steg, P.G., Gore, J.M.: A validated prediction model for all forms of acute coronary syndrome. Estimating the risk of 6-month postdischarge death in an international registry. *J. Amer. Medical Assoc.* **291**(22), 2727–2733 (2004)
18. Conroy, R.M., Pyörälä, K., Fitzgerald, A.P.: Estimation of ten-year risk cardiovascular disease in Europe: the SCORE project. *Eur. Heart J.* **24**, 987–1003 (2003)
19. Sakovskaia, A., Nevzorova, V., Brodskaya, T., Chkalovec, I.: Condition aortic stiffness and content of adipokines in the serum of patients with essential hypertension in young and middle-aged. *J. Hypertension* **33**(N e-suppl.1), 182–187 (2015)
20. Ni, W., Zhou, Z., Liu, T., Wang, H., Deng, J., Liu, X., Xing, G.: Gender-and lesion number-dependent difference in “atherogenic index of plasma” in Chinese people with coronary heart disease. *Sci Rep.* **16**,7(1), 13207 (2017)
21. Gunay, S., Sariaydin, M., Acay, A.: New predictor of atherosclerosis in subjects with COPD: atherogenic indices. *Respir Care.* **61**(11), 1481–1487 (2016)
22. Scientific and Organizing Committee of the ESSE-RF project. Epidemiology of cardiovascular diseases in various regions of Russia (ESSE - RF). Justification and design of the research. *Preventive medicine.* **6**, pp. 25–34 (2013)
23. Nevzorova, V.A., Shumatov, V.B., Nastradin, O.V.: The state of the function of the vascular endothelium in people with risk factors and patients with coronary heart disease. *Pacific Med. J.* **2**, 37–44 (2012)
24. Odden, M.C., Tager, I.B., Gansevoort, R.T., Bakker, S.J.L., Fried, L.F., Newman, A.B., Katz, R., Satterfield, S., Harris, T.B., Sarnak, M.J., Siscovick, D., Shlipak, M.G.: Hypertension and low HDL cholesterol were associated with reduced kidney function across the age spectrum: a collaborative study. *Ann. Epidemiol.* **23**(3), 106–111 (2013)
25. Al-Naamani, N., Palevsky, H.I., Lederer, D.J., Horn, E.M., Mathai, S.C., Roberts, K.E., Tracy, R.P., Hassoun, P.M., Girgis, R.E., Shimbo, D., Post, W.S., Kawut, S.M.: Prognostic significance of biomarkers in pulmonary arterial hypertension. *Ann. Am. Thorac. Soc.* **13**(1), 25–30 (2016)

26. Fowkes, F.G., Murray, G.D., Butcher, I., Heald, C.L., Lee, R.J., Chambless, L.E.: Ankle brachial index combined with Framingham risk score to predict cardiovascular events and mortality: a meta-analysis. *JAMA* **300**, 197–208 (2008)
27. D'Agostino, R.B., Pencina, M.J., Massaro, J.M., Coady, S.: Cardiovascular disease risk assessment: insights from Framingham. *Global Heart* **8**(1), 11–23 (2013)
28. Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W.: Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**(1), 128–138 (2010)