https://russjcardiol.elpub.ru
doi:10.15829/1560-4071-2020-3-3751

ISSN 1560-4071 (print)
ISSN 2618-7620 (online)

Machine learning for predicting the outcomes and risks of cardiovascular diseases in patients with hypertension: results of ESSE-RF in the Primorsky Krai

Nevzorova V. A.[1], Plekhova N. G.[1], Priseko L. G.[1], Chernenko I. N.[1], Bogdanov D. Yu.[2], Mokshina M. V.[1], Kulakova N. V.[1]

Aim. To assess the prospects of using artificial intelligence technologies in predicting the outcomes and risks of cardiovascular diseases (CVD) in patients with hypertension (HTN).

Material and methods. A software application was created for data mining from respondent profiles in a semi-automatic mode; libraries with data preprocessing were analyzed. We analyzed the main and additional parameters (35) of CVD risk factors in 2131 people as a part of ESSE-RF study (2014-2019). To create a forecasting model, a high-level language Python 2.7 was used using object-oriented programming and exception handling with multithreading support. Using randomization, learning (n=488) and test (n=245) samples were formed, which included data from patients with an established diagnosis of HTN.

Results. The prevalence of HTN among subjects was 34,39%. There were following significant factors for predicting CVD: anthropometric parameters, smoking, biochemical profile (total cholesterol, ApoA, ApoB, glucose, D-dimer, C-reactive protein). As a result of a 5-year follow-up, CVD was found in 235 people (32,06%) with HTN and 187 people (13,38%) without HTN; mortality rates were 1,27% in subjects with HTN and 1,12% — without HTN. The absolute mortality risk among participants with HTN (0,037) was significantly higher ($p<0,05$) than in patients without HTN (0,017). To create a neural network (NN), the basic Sequential model from the Keras library was used. During machine learning, 26 variables important for the CVD development were used as input and 9 neurons — as output, which corresponded to the number of established cardiovascular events. The created NN had a predictive value of up to 97,9%, which exceeded the SCORE value (34,9%).

Conclusion. The data obtained indicate the importance of risk factor phenotyping using anthropometric markers and biochemical profile for determining their significance in the top 20 predictors of CVD. The Python-based machine learning provides CVD prediction according to standard risk assessments.

Key words: cardiovascular risk factors, hypertension, artificial intelligence.

Relationships and activities: the study was supported by the grant of Russian Foundation for Basic Research (№ 19-29-01077).

[1]Pacific State Medical University, Vladivostok; [2]Vladivostok Clinical Hospital № 1, Vladivostok, Russia.

Nevzorova V. A.* ORCID: 0000-0002-0117-0349, Plekhova N. G. ORCID: 0000-0002- 8701-7213, Priseko L. G. ORCID: 0000-0002-3946-2064, Chernenko I. N. ORCID: 0000-0001-5261-810X, Bogdanov D. Yu. ORCID: 0000-0002-8388-5566, Mokshina M. V. ORCID: 0000-0003-3663-1560, Kulakova N. V. ORCID: 0000-0001-6473-5653

*Corresponding author: nevzorova@inbox.ru

Received: 13.02.2020
Revision Received: 21.02.2020
Accepted: 12.03.2020

For citation: Nevzorova V. A., Plekhova N. G., Priseko L. G., Chernenko I. N., Bogdanov D. Yu., Mokshina M. V., Kulakova N. V. Machine learning for predicting the outcomes and risks of cardiovascular diseases in patients with hypertension: results of ESSE-RF in the Primorsky Krai. *Russian Journal of Cardiology*. 2020;25(3):3751. (In Russ.)
doi:10.15829/1560-4071-2020-3-3751

Most often, to predict the risk of cardiovascular diseases (CVD), multivariate regression analysis models are developed, which combines data on a limited number of established risk factors (RF). Such an algorithm assumes that all included RF are linearly associated with CVD outcomes and are characterized by limited interaction between each other or its absence. Due to such a limitative approach to modeling and predictors, these algorithms, in particular, the Framingham, SCORE, and DECODE equations, demonstrate insufficient prognostic efficiency [1]. In various areas, including in medicine, the most effective prognostic approach is data mining, especially, deep neural networks (DNN). For now, there are many libraries ready for use, on the basis of which it is possible to use DNN in practice. Such a methods based on machine learning (ML) increase the efficiency of risk prediction through the use of data warehouses with the independent identification of new risk predictors and complex interactions between them. There is a small number of studies on the prospects of using ML to predict CVD risk. Some studies showed that, compared with the above equations, ML significantly increases the accuracy of CVD risk prediction and, as a result, the number of patients who could benefit more from preventive measures before the onset of severe manifestations [2-4].

The current study presents the potential value of using ML to develop a model for CVD risk prediction using blood pressure (BP) data. We prospectively analyzed data obtained by a cross-sectional examination of 2800 residents of the Primorsky Krai without CVD. This examination was conducted from 2014 to 2019 as a part of ESSE-RF study. To develop a risk prediction model, we used the modern automated high-level language Python and open-source neural-network library Keras. Learning and optimization of DNN were carried out using the Adam algorithm. The prognostic value of DNN in healthy general population, including a clinically significant subgroup of patients with hypertension (HTN), was assessed.

The aim of the study was to assess the prospects of using artificial intelligence technologies in predicting the outcomes and risks of CVD in patients with HTN.

Material and methods

As a part of the ESSE-RF study (2014-2019), cross-sectional examination of Primorsky Krai residents was performed [5]. This study was performed in accordance with the Helsinki declaration and Good Clinical Practice standards. To form a representative sample, we used the continuous method by an individual invitation of participants. There were following inclusion criteria: signed informed consent, age 24-65 years, full completion of the questionnaire, available data on cardiovascular RF. The exclusion criteria were refusal to participate and active cancer. A total of 2,800 people were included in the study; 2131 of them (76,1%) completed the program by 2019. Patterned sampling using the data adjustment algorithm was carried out in a computer program for extracting data from respondents' questionnaires in a semi-automatic mode (Figure 1).

We analyzed the prevalence of main RF: overweight with body mass index (BMI) calculation, waist circumference (WC), BP and pulse pressure (PP) levels; heart rate (HR), smoking, sedentary lifestyle; SCORE 10-year mortality risk (in individuals ⩾40 years old and ⩽65 years old) based on gender, age, systolic blood pressure (SBP), total cholesterol (TC) and smoking status. BP levels were evaluated in accordance with the guidelines [6], where BP ⩾140/90 mm Hg belongs to HTN. Family history of heart

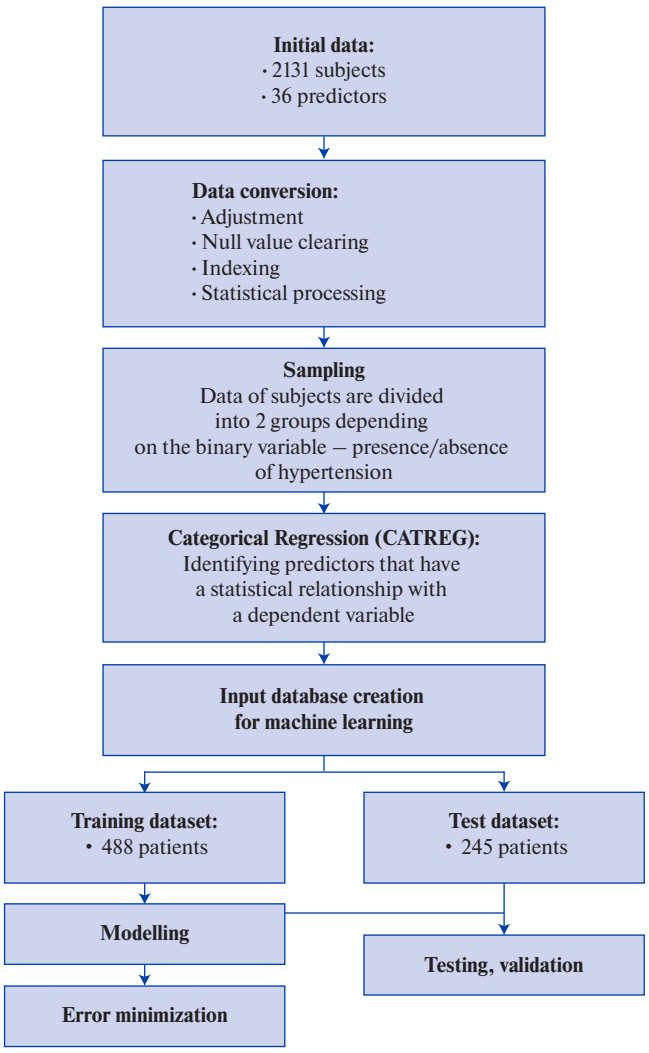


Figure 1. Study design flowchart.

Table 1

Clinical and laboratory characteristics of the subjects

Parameters (M±m)	Group of healthy individuals (n=1398)	Group of hypertensive patients (n=733)
Mean age, years	42,68±11,45	51,56±9,82*
Height, cm	168,82±0,25	168,02±0,36
Weight, kg	75,62±0,44	85,44±0,63*
Body mass index, kg/m^2	26,47±0,14	30,35±0,22*
Waist circumference, cm	85,97±0,39	96,71±0,53*
Mean SBP, mm Hg	123,87±0,27	156,48±0,58*
Mean DBP, mm Hg	75,39±0,22	89,19±0,39*
Mean PP, mm Hg	48,49±0,22	67,29±0,50*
Mean HR, bpm	74,91±0,32	77,75±0,68*
Total cholesterol, mmol/L	5,49±0,03	5,87±0,05*
LDL, mmol/L	3,49±0,03	3,76±0,04*
HDL, mmol/L	1,45±0,01	1,4±0,01*
Triglycerides, mmol/L	1,24±0,02	1,67±0,04*
LP(a), mg/dl	20,19±0,65	20,62±0,92
ApoA, g/l	1,76±0,01	1,81±0,02*
ApoB, g/l	0,82±0,01	0,89±0,01*
Glucose, mmol/L	5,23±0,03	5,86±0,08*
Creatinine, μmol/L	69,12±0,44	71,55±0,77*
Uric acid, μmol/L	315,87± 2,71	356,38±4,01*
D-dimer, μg/L	212,30±7,16	186,05±4,93*
C-reactive protein, mg/L	2,63±0,16	3,78±0,25*

Note: differences are significant at * — p<0,05.

Abbreviations: PP — pulse pressure, LP(a) — Lipoprotein(a), ApoA — apolipoprotein A, ApoB — apolipoprotein B, DBP — diastolic blood pressure, HDL — high density liproproteins, LDL — low density liproproteins, SBP — systolic blood pressure, HR — heart rate.

disease, smoking and alcohol status was determined by anamnesis collection. The parameters of lipid profile (TC, triglycerides (TG), low density lipoproteins (LDL) and high density lipoproteins (HDL), lipoprotein(a) (LP(a)), apolipoprotein A (ApoA), apolipoprotein B (ApoB)), glucose, creatinine, uric acid, D-dimer, C-reactive protein (CRP) levels were determined.

For neural network data analysis, a high-level language Python 2.7 (Python Software Foundation License) was used based on object-oriented programming with exception handling mechanism and multithreading support. After analysis of Python libraries (TensorFlor, Keras), Keras was used to initiate ML. Learning and optimization of the DNN were carried out according to the Adam algorithm (adaptive moment estimation) with the calculation of adaptive learning rate for each parameter. Adam also keeps an exponentially decaying average of past squared gradients AdaDelta and past gradients m_t, similar to momentum.

Statistical processing was carried out using the software package Stata 11.2 and R 3.2.1 (StataCorp LP, USA). Continuous variables are represented as medians with interquartile intervals; the comparison was carried out using the Student's t-test. To compare discrete variables, the Chi-squared test or the Fisher's exact test were used. The cumulative probabilities for CVD were estimated using the Kaplan-Meier method and compared using a log rank test. To assess the impact of various variables on the CVD risk, univariate and multivariate regression models (Cox proportional-hazards model) were used. Hazard ratios and 95% confidence intervals with corresponding p-values are presented. The differences were considered significant at p<0,05. The effectiveness of ML prediction algorithms was assessed using a validation coefficient.

The study was supported by the grant of Russian Foundation for Basic Research (№ 19-29-01077).

Results and discussion

Characteristics of the study population. Complete information on 2131 participants was determined using a randomized algorithm for computer data adjustment (Table 1). The mean age of participants at

Table 2

Parameters included in the machine learning algorithm (data from hypertensive patients)

Risk Factors (M±m)	CVD (n=293)	Without CVD (n=440)	P
Women, %	67,8	42,65	-
Mean age, years	52,67±0,85	52,16±0,50	0,61
Smoking, %	33,9	39,53	-
Height, cm	165,87±0,94	167,48±0,52	0,02*
Weight, kg	85,12±1,57	85,25±0,91	0,94
Body mass index, kg/m^2	31,03±0,55	30,42±0,30	0,33
Waist circumference, cm	97,33±1,35	97,47±0,79	0,92
Thigh circumference, cm	107,43±0,95	107,40±0,56	0,97
Mean SBP, mm Hg	156,29±1,59	157,55±0,85	0,48
Mean DBP, mm Hg	87,52±0,91	89,64±0,58	0,05
Mean PP, mm Hg	67,91±0,73	69,08±1,43	0,46
Mean HR, bpm	76,83±1,27	77,26±0,62	0,76
Glucose, mmol/L	5,77±0,13	5,96±0,13	0,31
Total cholesterol, mmol/L	5,92±0,12	6,01±0,07	0,51
HDL, mmol/L	1,40±0,03	1,40±0,02	1
LDL, mmol/L	3,83±0,10	3,86±0,06	0,79
Triglycerides, mmol/L	1,66±0,09	1,68±0,06	0,85
LP(a), mg/dl	20,09±0,4	20,22±0,2	0,77
ApoA, g/l	1,84±0,04	1,85±0,02	0,82
ApoB, g/l	0,89±0,02	0,92±0,01	0,18
C-reactive protein, mg/L	3,34±0,61	3,78±0,34	0,52
Creatinine, μmol/L	68,93±0,95	72,04±1,30	0,05
Uric acid, μmol/L	353,56±8,98	354,29±5,56	0,94
D-dimer, μg/L	178,99±9,82	185,92±6,16	0,55

Note: differences are significant at * — p<0,05.

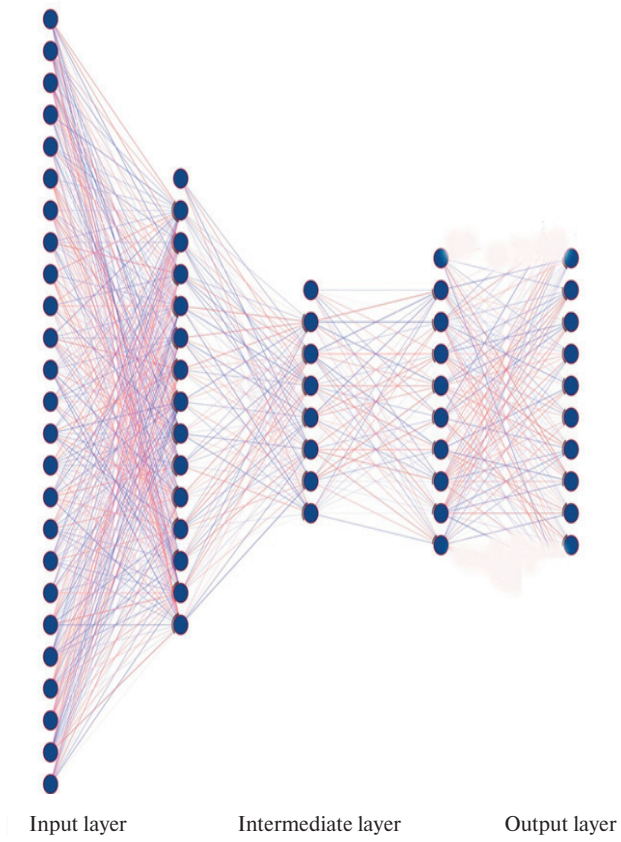
Abbreviations: PP — pulse pressure, LP(a) — Lipoprotein(a), ApoA — apolipoprotein A, ApoB — apolipoprotein B, DBP — diastolic blood pressure, HDL — high density liproproteins, LDL — low density liproproteins, SBP — systolic blood pressure, HR — heart rate.

the study beginning was 45,75 (11,7) years (men — 874 (41%)). During the 5-year follow-up (5-95th percentile: 3,4-4,7 years), 422 cases of CVD were detected in the age range of 60,2±5,6 years for men and 61,1±4,8 years for women. In the group of people without HTN (n=1398), CVD was established in 13,38% (n=187), while among HTN people (n=733) — in 32,06% (n=235). According to the International Classification of Diseases (ICD-10), angina was detected in 51,06% of HTN people; atrial fibrillation and flutter — in 11,06% and 14,44% of people with and without HTN, respectively; old myocardial infarction — in 5,53% and 9,09% of people with and without HTN, respectively; unspecified stroke — in 6,81% of people with HTN. In the HTN group, the absolute mortality risk was significantly higher than in individuals without HTN (0,037 vs 0,017, respectively; p <0,05); the relative mortality risk was 2,146.

CVD risk predictors. The revealed statistical differences in the studied RF between HTN (experimental) and non-HTN (comparison) groups are presented in Table 1.

All participants had excess body weight. Among people with HTN, the mean BMI was higher compared with non-HN individuals (p=0,00001). WC in men did not exceed the recommended value (highest value — 98,5±0,67 cm). Compared with comparison group, women of the experimental group had higher WC (95,13±0,78 cm vs 82,89±0,49 cm; p<0,0001).

The PP level exceeded the threshold level among HTN participants, while the maximum value (68,88±0,71 mm Hg) was observed in women. The



Fugure 2. Neural network model.

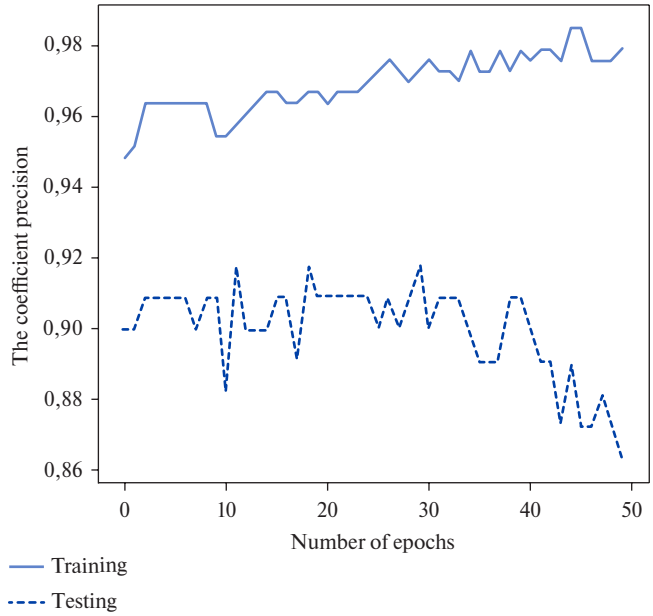


Figure 3. Changes of neural network accuracy in the learning and testing processes (fragment of 50 epochs).

mean HR in the groups was within the acceptable range.

The mean TC level exceeded the normal value in all subjects. The highest mean LDL value (3,88±0,05 mmol/L) was noted in women with HTN. A significant difference in HDL levels was found between HTN and non-HTN women (p=0,007). The mean TG level exceeded the norm only in HTN men (1,77±0,08 mmol/L).

Fasting glucose >5,6 mmol/L is considered to be the RF for diabetes and CVD. Significant differences between the groups were found (p<0,001). Exceeding the threshold level was observed in all HTN participants.

The mean creatinine level did not exceed acceptable values in 100% of cases. However, the studied groups had significant differences in terms of this RF (p=0,006). The highest mean creatinine level was 318,80±4,96 μmol/L in HTN women.

LP(a) is an atherogenic lipid variant, which has a high prognostic value for atherosclerosis and CVD, in particular, coronary artery disease. Acceptable values of LP(a) are in the range of 5-18 mg/dl. There were no significant differences of this RF between HTN and non-HTN groups (20,62±0,93 vs 20,19±0,65 mg/dl, respectively; p=0,704). Non-HTN men had slightly higher mean value of LP(a) (20,70±1,09 mg/dl) compared with HTN men (18,16±1,28 mg/dl), but this difference was not significant.

It is assumed that ApoA and ApoB levels may be decisive in determining the atherosclerosis risk, especially when other lipid parameters do not exceed the norm and/or there are no manifestations of vascular damage [7]. There were significant differences between experimental and comparison groups in ApoA (p=0,025) and ApoB (p=0,00001) levels.

Compared with the experimental group, non-HTN patients had higher levels of D-dimer (Table 1) (p=0,0026). Also, a significant (p=0,0001) difference was found between non-HTN (236,51±1,56 μg/L) and HTN women (190,51±5,72 μg/L).

The mean values of CRP were higher in HTN individuals compared with non-HTN individuals, regardless of gender. The differences between groups were significant (p=0,0001).

Thus, statistical processing revealed following significant factors: anthropometric parameters — height, weight, BMI, WC; blood biochemical parameters — levels of TC, fasting glucose, ApoA and ApoB, D-dimer and CRP.

ML model for predicting CVD outcomes in HTN patients. Various programming languages are used to create DNN, where basic mathematical operators and multidimensional arrays are supported. These include such interpreted C languages as Python, which we used for ML. To develop the DNN, we used the basic sequential model from the Keras

Table 3

Stratification of hypertensive subjects aged 24-65 years without CVD at the study beginning, depending on the presence of first cardiovascular event after a 5-year follow-up

№	№ ICD-10 code	Disease	Number of persons	Specific weight
1	I20.8	Other forms of angina pectoris	120	51,06%
2	I48	Atrial fibrillation and flutter	26	11,06%
3	I25.2	Old myocardial infarction	13	5,53%
4	I64.0	Unspecified stroke	16	6,81%
5	I70.2	Atherosclerosis of native arteries of the extremities	26	11,06%
6	I20.1	Angina pectoris with documented spasm	14	5,96%
7	I69.3	Sequelae of cerebral infarction	7	2,98%
8	I69.4	Sequelae of stroke, unspecified	6	2,55%
9	I20.0 + I21.9	Acute coronary syndrome (unstable angina and acute myocardial infarction)	7	2,98%

Abbreviation: ICD-10 — International Classification of Diseases.

library, which is represented by multiple layers combining to a Rumelhart multilayer perceptron. Using the randomization function X_train, X_test, y_train, y_test=train_test_split (X,Y,test_size=0,40, random_state=42), 2 samples were formed from the total data array: learning (n=488) and test (n=245, Figure 1). These included data from patients with established HTN. Of all the subjects with HTN (n=733), 144 participants were smokers, 170 — former smokers, 419 — non-smokers. Input layer of the prediction model included 26 most important variables (Table 2, Figure 2). Hidden layers were determined empirically: the first layer, where the matrix of weighting coefficients and the matrix of input data of previous neurons are multiplied (15 neurons); the second layer contained the result of minimizing the error (8 neurons), and the third layer was used to refine the prognosis (10 neurons). The output layer consisted of 9 neurons, each of which corresponded to the number of events belonging to the ICD-10 diagnosis (Table 3).

Learning and optimization of DNN were carried out according to the Adam algorithm, which calculates adaptive learning rates for each parameter. Adam keeps an exponentially decaying average of past squared gradients AdaDelta and past gradients m_t, similar to momentum. The Adam algorithm differs from other adaptive methods in the rapid learning rate and efficiency. Changes of the DNN accuracy in the learning and testing processes are presented in Figure 3.

The sample size for the ML was 66,6% of all HTN subjects. Learning and optimization of DNN was carried out in 1000 epochs. As a result of testing using the Adam algorithm, the DNN accuracy reached 97,9%, and the loss value was in the range 10-7-10-8 (Figure 3). During testing, accuracy decreased to 95,5% (Figure 3).

Classification analysis. To assess the clinical significance of our results, we compared our model with the SCORE model in predicting CVD risk. At this operating point, the basic SCORE model correctly predicted 145 CVD out of 465 cases (sensitivity — 61,7%, predictor coefficient — 1,5%). Our ML model correctly predicted 230 CVD out of 733 subjects (sensitivity — 97,9%). The resulting difference is 36,2% of increase in the accuracy of predicting CVD using ML methods.

Conclusion

The study showed that ML methods can be effectively used for cardiovascular risk prediction. The Python-based method provides CVD prediction using standard risk assessments. The use of the randomization function for selecting variables, followed by use of Cox regression methods allows improving prediction. The results also indicate the importance of advanced phenotyping of the subjects using anthropometric markers and blood biochemical parameters, when determining the top-20 predictors for CVD.

The Multi-Ethnic Study of Atherosclerosis (MESA) showed that indicators such as age, inflammation, and vascular diseases prevails in death prognosis. It is also indicated that impaired glucose metabolism an HTN is associated with stroke prognosis, and markers of subclinical atherosclerosis are central to the prognosis of various CVD [8]. The ML method used by us is unique in that it demonstrates the patterns of predictor changes that differ for specific disease outcomes. Relatively high accuracy values (from 86 to 98%) indicate the acceptability of

using this ML method in cardiovascular risk estimation. The advantage of current study is the consideration of anthropometric data, the results of laboratory tests and other important predictors of CVD. Thus, combination of ML and advanced phenotyping increases the accuracy of predicting cardiovascular events in HTN population. The developed approaches allow to more accurately understand markers of subclinical diseases without a priori guess on their nature.

Relationships and activities: the study was supported by the grant of Russian Foundation for Basic Research (№ 19-29-01077).

References

1. Siontis GC, Tzoulaki I, Siontis KC, et al. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318. doi:10.1136/bmj.e3318.
2. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. Published 2017 Apr 4. doi:10.1371/journal.pone.0174944.
3. Ahmad T, Lund LH, Rao P, et al. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association. 2018;7(8):e008081. doi:10.1161/JAHA.117.008081.
4. Plekhova NG, Nevzorova VA, Rodionova LV, et al. Scale of Binary Variables for Predicting Cardiovascular Risk Scale for predicting cardiovascular risk. Proceedings of the 2018 3rd Russian-Pacific Conf. on computer technology and applications (RPC). 2018. doi:10.1109/RPC.2018.8482216.
5. The Scientific and Organizing Committee of the project of the ESSE-RF. Epidemiology of cardiovascular diseases in various regions of Russia (ESSE-RF). Rationale and design of the study. Prophylactic medicine. 2013;6:25-34. (In Russ.)
6. Mancia G, Fagard R, Narkiewicz K, et al. Recommendations for the treatment of arterial hypertension. ESH/ESC 2013. Russian Journal of Cardiology. 2014;(1):7-94. (In Russ.) doi:10.15829/1560-4071-2014-1-7-94.
7. Plekhova NG, Nevzorova VA, Rodionova LV, et al. Indicators of lipoprotein metabolism in young patients with arterial hypertension. Bulletin of modern clinical medicine. 2019;4:44-51. (In Russ.) doi:10.20969/VSKM.2019.12(4).44-51.
8. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017;121(9):1092-101. doi:10.1161/CIRCRESAHA.117.311312.