Machine Learning for Assessment of Cardiometabolic Risk Factors Predictive Potential and Prediction of Obstructive Coronary Arteries Lesions

Karina Shakhgeldyan[1], Boris Geltser[2], Vladislav Rublev[2(✉)], Andrey Vishnevskiy[1], Elena Emtseva[1], and Mikhail Tsivanyuk[2]

[1] Vladivostok State University of Economics, Vladivostok, Russia
[2] Far Eastern Federal University, School of Medicine, Vladivostok, Russia

Abstract. The aim of this study was searching and validation of new obstructive coronary arteries lesions predictors and prognostic models development for its verification in patients with ischemic heart disease prior to invasive coronary angiography. Research included a step-by-step algorithm for predictors selection and validation as well as thresholds measurements with filtering and wrapping techniques. Cross-validation of predictive models based on multivariate logistic regression, support vector machine and random forest were made by averaging of 4 quality metrics. Based on selected predictors in continuous and categorical forms the best developed predictive model was logistic regression models ensemble with the following quality metrics: area under the ROC curve 0.85, accuracy - 0.80, sensitivity - 0.82, and specificity - 0. 73, which is higher than the existing CAD Consortium scale.

Keywords: Machine learning · Categorization of variables · Ensembles of models · Coronary arteries · Predictive models

1 Introduction

Coronary artery disease (CAD) takes one of the leading places in the morbidity and mortality structure among majority countries of the world [1]. Thus, annual CAD death rate is around 9,5 million people or more than 17% of all world death. Strategy of early CAD diagnostics is used for mortality reduction and according to the clinical recommendations of cardiological community is aimed to improve the technologies of risk stratification and prophylactics.

Coronary insufficiency is main known mechanism of CAD development due to an imbalance between myocardial oxygen demand and its actual delivery. The narrowing of coronary arteries (CA) lumen by more than 50% is the most common reason for this

This work was carried out with partial support of grants from the Russian Foundation for Basic Research within the framework of research projects No. 20-37-90081, No. 19-29-01077.

S. M. Kovalev et al. (Eds.): RCAI 2021, LNAI 12948, pp. 1–15, 2021.
https://doi.org/10.1007/978-3-030-86855-0_8

condition development. The exact degree of narrowing is specified during and invasive diagnostic procedure – coronary angiography (ICA). The latter is the gold standard of coronary blood flow functional anatomical status diagnostics and surgical treatment indications identification. Recently there are more and more studies indicating an increase number of persons with non-obstructive (less than 50% of the lumen) lesions of coronary artery (NOCAD) among patients with suspected CAD [2, 3]. Thus, hemodynamically significant CA stenoses in patients with clinical signs of CAD according to the results of ICA were detected only in 40% of cases [2]. According to the US national registry data, among of patients with suspected CAD NOCAD occurred in 58%. Registry data from Brazil, Finland and Switzerland describes NOCAD in 76%, 57% and 32% of patients, respectively [4].

Despite ICA's high diagnostic value, usage is associated with a certain risk of surgical complications. So, ICA is described as cause of death in 0.1–0.14% of patients. ICA-associated myocardial infarction is diagnosed in 0.06–0.07% cases. Allergic reactions to the contrast introduction and local post-puncture vascular complications were recorded in 0.23% and 2%, respectively, cerebrovascular complications - in 0.07–0.14% of patients.

Methods for assessing the pretest probability (PTP) of obstructive coronary arteries damage (OCAD) (before ICA) for patients with suspected CAD were firstly introduced in routine clinical practice around 40 years ago by American cardiologists George Diamond and James Forrester. Their article "Analysis of Probability as an Aid in the Clinical Diagnosis of Coronary Artery Disease", published in The New England Journal of Medicine in 1979 were presenting the Bayesian classifier model allowing to calculate OCAD probability for patients with suspected CAD before functional and laboratory tests [5]. Gender, age (from 30 up to 69 years) and clinical symptoms of CAD (typical and atypical angina pectoris, cardialgia) were described as predictors in this model based on ICA results of 4952 patients. For several decades, the Diamond-Forrester (DF) scale has been one of the most popular PTP OCAD methods [6]. Years after its usage showed a significant overestimation of CAD likelihood among surveyed, especially in the female population. In 2011 Genders TS, et al. modified the DF scale, adapting it for modern cohort of patients with age limit extension up to 80 years [7]. For a new CAD Consortium scale development, EuroAIM registry data of 2,260 patients from 14 Europe and United States medical centers were used. All included subjects were complaining chest pain, had no CAD history and were underwent ICA for its verification. Update and expansion of PTP models showed a significant increase of predictive power. In 2012, same authors improved calculator and proposed the CAD Consortium clinical model (base model + risk factors for cardiovascular diseases (CVD), including diabetes mellitus (DM), hypertension (HTN), smoking, hyperlipidemia, and body mass index (BMI)) and an extended model (clinical model + coronary calcium index according to multislice computed tomography). The latest model showed accuracy increase based on C-statistics indicators from 0.77 to 0.79, and the reclassification - by 35%. Advantages of this method was proven by results of Bittencourt MS, et al. (2016) study, where straight comparison of DF scale with two CAD Consortium models were made based on 2,274 patients clinical data [8]. Also, this study reaffirms the overestimation of CAD prevalence among the surveyed by DF scale. Withal, both clinical and extended CAD Consortium models provided higher prediction accuracy for OCAD detection: area under the ROC

- curve (AUC) for DF was 0.713, and for CAD Consortium 1 and 2 - 0.752 and 0.791, respectively. Besides, with these models, significantly more patients were attributed to the low OCAD probability group (24.6% and 30.0% vs 8.3% - by DF), and the persons proportion with a high OCAD risk was only 1.1% vs 18% by DF. Authors suggested that widespread of these methods in routine clinical decisions can tremendously reduce the need for invasive CAD diagnosis [9]. Juarez-Orozco LE, et al. (2019), developed new method for PTP OCAD determination based on the Bayesian classifier and data analysis of three large-scale studies results, describing patients with suspected coronary artery disease (n = 15815, mean age 59 ± 11 years) [10]. This method was included to the European Society of Cardiology recommendations for the diagnosis and treatment of chronic coronary syndrome [11, 12].

Recently, modern machine learning (ML) methods are more commonly used for predictive research in clinical cardiology, allowing to increase the forecast accuracy by identifying non-obvious patterns. At the same time, there is only a small number of articles where these methods were used for OCAD prediction.

The aim of this study was searching and validation of new OCAD predictors with determination of their threshold values and prognostic models development for its pre-test diagnostics in patients with CAD prior to ICA, based on ML methods.

2 Materials and Methods

2.1 Patient Characteristics

A prospective cohort study included 496 patients (314 men and 182 women) aged 30 to 80 years with a median of 62 years and a 95% confidence interval (CI) [60; 64], who were proceed to the emergency cardiology department of the Vladivostok Clinical Hospital No. 1 in 2017–2020. All patients underwent invasive ICA. Among the surveyed cohort, 2 groups were identified. The first included 345 (69.6%) patients with hemodynamically significant coronary artery narrowing (≥50%) according to ICA results, the second included 151 (30.4%) with NOCAD (<50%).

Before ICA patients clinical and functional status was evaluated by 29 indicators containing anamnestic, anthropometric, clinical and laboratory data associated with cardiometabolic risk (CMR). Measurements of height (Ht), weight, waist circumference (W), hips (H), calculation of body mass index (BMI), WH ratios (WHR) (indexed to gender), WHt ratio (WHtR) were carried out. The levels of glucose, total cholesterol (TC), high density lipoprotein cholesterol (HDL cholesterol) and low (LDL cholesterol) density, triglycerides (TG), creatinine, uric acid (UA) were determined.

The indices of visceral adiposity (VAI), lipid accumulation product (LAP), atherogenicity (AIP) were calculated using the well-known formulas [13]. The insulin resistance index (IRI) was determined by the ratio of TG/HDL cholesterol [14], and the glomerular filtration rate (GFR) was determined using the CKD EPI formula.

2.2 Data Processing

The end point of the study was presented by the OCAD in the binary form feature ("absence" or "presence"). Input signs - a subgroup of potential predictors was expressed in the form of continuous and categorical variables. Filtration and wrapping methods were used for data processing and analysis [15]. Filtration was performed by mathematical statistics, and ML methods were used as a "wrapper". The first were represented by the Fisher, Mann-Whitney, Chi-square tests and one-way logistic regression (LR) with the calculation of weights for a normalized sample. The second - by ML methods: multifactorial LR (MLR), random forest (RF) and support vector machine (SVM).

Significance of features and testing of hypotheses was confirmed by a p-value <0.05. The quality of the models developed on training samples was assessed using a cross-validation test procedure, by averaging 4 metrics: area under the ROC curve (AUC), accuracy (ACC), sensitivity (Sen), and specificity (Spec). Cross-validation was performed using a k-fold approach on 10 stratified samples.

2.3 Study Design

The study design included 5 stages.

1. On the first, the probability of the presence of OCAD was calculated using the CAD Consortium method [7] and its predictive value was assessed.
2. On the second stage, in order to identify potential predictors linearly related to OCAD, 29 CMR factors were analyzed in the comparison groups. We did not use the most significant predictor of the CAD Consortium scale (pain syndrome), which is the most significant for diagnostics, in order to identify other, previously unused predictors. For continuous variables, the Mann-Whitney test was used, and for categorical variables - the chi-square test. The odds ratio (OR) and their 95% CI were assessed by the Fisher test.
3. At the third stage, using one-factor LR-models, the weighting coefficients of individual indicators were determined.
4. At the fourth stage, based on LR results, threshold values of factors with the highest predictive potential were identified.
5. At the fifth stage, the predictive models of OCAD were developed using the MLR, RF and SVM. Data analysis and model development were performed in R-studio and Python by R languages.

3 Results

During the first stage of the study, the probability of OCAD was assessed in accordance with the CAD Consortium scale recommended by the European and American Society of Cardiology. The quality metrics of the obtained assessment were: ACC - 0.7, AUC - 0.75, Sen - 0.68, Spec - 0.71, which confirmed the need of more advanced models development based on new predictors.

At the second stage of the study, an intergroup analysis of the factors characterizing the clinical and functional status of patients was carried out, which showed the presence of statistically significant differences in 15 parameters (Table 1). At the same time, the maximum level of reliability was recorded for indicators of gender (male sex), HDLC, AIP, and Creatinine (p-value <0.0001). The highest OR values were associated with males (OR = 2.5) and active smoking (OR = 2.3). A less noticeable, but statically significant likelihood of OCAD was associated with a family history of CVD (OR = 1.6). It should be noted that DM and hypertension (HTN) were recorded with the same frequency in patients with OCAD and NOCAD. According to the preliminary analysis, age, height, WHtR, weight and BMI of the surveyed, the concentration of glucose and CRP, systolic (SBP), diastolic (DBP) and pulse (PAP) blood pressure also did not affect the likelihood of OCAD.

Table 1. Patients clinical and functional characteristics (ME, 95% CI).

Parameter	1 group (OCAD), n = 345	2 group (NOCAD), n = 150	OR, 95% CI	p-value
Age, years	62 [61;64]	62 [59;64]		0,57
Male, (%)	241 (69,9%)	72 (48,3%)	2,5 [1,6; 3,7]	<0,0001
Smokers, (%)	135 (40%)	34 (23%)	2,3 [1,5; 3,7]	0,0001
Family history of CVD, (%)	102 (30%)	32 (21%)	1,6 [1,0; 2,6]	0,048
WHR, c.u	1,08 [1,05; 1,1]	1,05 [1,0; 1,1]		0,005
TC, mmol/l	5,6 [5,5; 5,9]	5,3 [5,1; 5,55]		0,012
TG, mmol/l	1,4 [1,35; 1,6]	1,3 [1,15; 1,4]		0,005
HDLC, mmol/l	1,2 [1,2; 1,25]	1,35 [1,3; 1,4]		<0,0001
LDLC, mmol/l	3,6 [3,4; 3,8]	3,3 [3,0; 3,4]		0,0003
IRI, c.u	1,2 [1,1; 1,35]	0,9 [0,8; 1,1]		0,0005
VAI, c.u	1,7 [1,4; 2,1]	1,2 [0,9; 1,6]		0,011
LAP, cm × μmol/l	49,6 [42,0; 60,3]	36,3 [30,7; 48,5]		0,04
AIP, c.u	3,55 [3,4; 3,9]	2,8 [2,4; 3,2]		<0,0001
Creatinine, μmol/ml	90 [88; 93]	79 [76; 83]		<0,0001
GFR, ml/min/1,73 m^2	73,2 [71,35; 74,8]	76.9 [73,2; 82,5]		0,012
UAA, μmol/l	379 [366; 393]	338 [320; 361]		0,007

At the third stage of the study, using standardized data, univariate LR models were constructed with the calculation of weight coefficients. This approach expands the possibilities for data processing and analysis due to a more detailed assessment of the degree and vector influence of potential predictors on the resulting variable.

Table 2. Patients Weights for univariate LR models for OCAD probability estimation (ME, 95% CI).

Parameter	Weights	P-value
Age	0,4 [−0,7; 1,5]	0,5
Male gender	0,9 [0,5; 1,3]	<0,0001
Smokers	0,8 [0,4; 1,3]	0,0002
Family history of CVD	0,5 [0,03; 0,9]	0,04
W	2,5 [0,3; 4,9]	0,03
WHR	3,4 [1,3; 5,8]	0,0029
TC	1,9 [0,4; 3,4]	0,013
TG	3,8 [0,1; 7,8]	0,05
HDLC	−3,9 [−6,3; −1,7]	0,0007
LDLC	2,1 [1,0; 3,3]	0,0003
IRI	1,6 [0,4; 2,9]	0,01
AIP	3,5 [1,9; 5,2]	<0,0001
Creatinine	4,7 [2,8; 6,8]	<0,0001
GFR	−3,35 [-5,7; −1,1]	0,004
CRP	6,1 [1,1; 13,8]	0,06
UA	1,7 [0,3; 3,1]	0,016

According to the results of the analysis statistically significant level of weighting coefficients took place in 13 variables (Table 2). The highest values of the weighting factors were associated with the Creatinine level (4.7; $p < 0.0001$), HDLC (−3.9; $p = 0.0007$) and AIP (3.5; $p < 0.0001$). The less valued indicators were WHR (3.4), GFR (−3.35), LDLC (2.1), TC (1.9), IRI (1.6), male sex (0.9), smoking status (0.8) and family history of CVD (0.5). At the same time, the weight coefficients of factors such as age, height, WHtR, weight, SBP, DBP, PAP, the presence of hypertension, diabetes, CRP and glucose levels, VAI and LAP were statistically insignificant. In the developed univariate models, most of the weighting coefficients had a positive value, which indicated an increase in the likelihood of OCAD in the presence of these signs or their levels increasing. On the contrary, negative values of the weights of HDLC and GFR indicate an increase in the risk of OCAD with a decrease in the level of these indicators.

At the fourth stage of the study, among the indicators selected at the previous stages, using one-factor LR, their threshold values with the highest predictive potential were identified (Table 3). To accomplish this task, indicators in a continuous form were transformed into categorical ones. It is known that categorization in continuous space leads to the loss of some information and the appearance of "quantization noise". However, in medical research, it is customary to operate with the concepts of norms and thresholds that are associated with risk factors that contribute to the development of diseases and

their complications. Within a certain "normative" range, the values of the indicator are not interrelated with the development of the disease, but, starting from a certain threshold, this indicator can act as a risk factor for an adverse event. To confirm this hypothesis, the threshold values of the indicators were verified, which were used for further analysis.

Table 3. The range of threshold values for potential OCAD predictors based on univariate LR-models.

Thresholds	1 group (OCAD), n = 345	2 group (NOCAD), n = 150	OR, 95% CI	p-value
Age, years male ≥55 female ≥65	235 (68%)	84 (56%)	1,7 [1,1; 2,5]	0,01
W, см male ≥105 female ≥115	76 (22%)	13 (9%)	3,05 [1,15; 8,1]	0,025
WHR, c.u Female and male ≥0,9	255 (74%)	75 (50%)	2,9 [1,5; 5,7]	0,0017
WHtR, c.u. ≥ 0,69	48 (14%)	4 (3%)	5,7 [1,25; 26,35]	0,025
TC ≥ 5,9 mmol/l	148 (43%)	43 (29%)	2,0 [1,3; 3,0]	0,001
TG ≥ 1,6 mmol/l	157 (45,5%)	48 (32%)	1,75 [1,2; 2,6]	0,006
HDLC ≤ 1,1 mmol/l	129 (37,5%)	30 (20%)	2,4 [1,5; 3,9]	0,0002
LDLC > 3,5 mmol/l	181 (52,5%)	52 (35%)	2,1 [1,4; 3,1]	0,0004
IRI ≥ 1,5 c.u	134 (39%)	40 (27%)	1,75 [1,15; 2,7]	0,009
LAP ≥ 38,5 см*mmol/l	238 (69%)	67 (45%)	2,7 [1,3; 5,45]	0,003
AIP ≥ 3,4 c.u	190 (55%)	56 (38%)	2,0 [1,3; 3,1]	0,0006
Creatinine, μmol/ml female ≥94 male ≥87	190 (55%)	47 (31,5%)	2,6 [1,7; 4,05]	<0,0001
GFR < 75 ml/min/1,73 m^2	193 (56%)	69 (46%)	1,5 [0,1; 2,2]	0,049
UA ≥ 356 μmol/l	210 (61%)	61 (41%)	2,2 [1,35; 3,6]	0,0008

The analysis results allowed us to identify age ranges for male (≥55 years old) and female (≥65 years old), which increased the likelihood of OCAD (OR = 1.7, p = 0.01). In men with WC ≥ 105 cm and women with WC ≥ 115 cm, the probability of detecting hemodynamically significant CA lesions increased 3 times (p = 0.025). Increase in

WHtR ≥ 0.69 c.u. (OR = 5.7, p = 0.025) and WHR ≥ 0.9 c.u. (OR = 2.9, p = 0.0017) also increased the likelihood of OCAD regardless subjects gender. Comparable chances of having OCAD were associated with lipid metabolism disorders, manifested by an increase in TC concentration ≥5.9 mmol/l (OR = 2.0, p = 0.001), LDLC > 3.5 mmol/l (OR = 2.1, p = 0.0004) and TG ≥ 1.6 mmol/l (OR = 1.75, p = 0.006), as well as a decrease in HDLC level ≤1.1 mmol/l (OR = 2.4, p = 0.0002). Similar values of OR were correlated with indicators IRI ≥ 1.5 c.u. (OR = 1.75, p = 0.009), LAP ≥ 38.5 cm * mmol/l (OR = 2.7, p = 0.003) and AIP ≥ 3.4 c.u. (OR = 2.0, p = 0.0006). The risk of OCAD increased with a serum concentration of UA ≥ 356 μmol/L (OR = 2.2, p = 0.0008) and Creatinine ≥ 87 μmol/mL in men and ≥94 μmol/L in women (OR = 2, 6, p < 0.0001). At the same time, the GFR index <75 ml/min/1.73 m^2 increased the probability of OCAD by 1.5 times (p = 0.049). It should be noted that testing the predictive potential of individual factors in different numerical ranges allowed us to identify predictively significant threshold values even among indicators (WHtR and LAP), the intergroup differences in median values of which at the previous stages of the study were insignificant.

At the fifth stage of the study, based on the methods of MLR, RF and SVM, prognostic models were developed to assess the likelihood of OCAD prior to performing invasive ICA (Table 4). Developing the models, we tested both continuous and categorical forms of predictors. Those shapes that provided the best accuracy were included in the final version of the models.

Table 4. Evaluation of the predictive models accuracy for pretest OCAD verification.

№		AUC	ACC	Sen	Spec
MLR					
1	TC* + WHR*	0,65	0,56	0,60	0,57
	TC + WHR + HDLC	0,72	0,63	0,64	0,62
2	TC* + WHR* + HDLC*	0,75	0,67	0,54	0,82
3	TC* + WHR + HDLC*	0,75	0,66	0,56	0,68
4	TC* + WHR* + HDLC	0,75	0,7	0,67	0,68
5	TC* + WHR* + HDLC + TG*	0,75	0,7	0,7	0,69
6	**TC* + WHR* + HDLC + IRI**	0,8	0,73	0,74	0,71
7	**Ensemble of models (TC* + LDLC* + WHR* + WHtR* + AIP* + LAP* + UA* + HDLC + IRI)**	0,85	0,80	0,82	0,77
SVM					
8	TC* + WHR*	0,64	0,63	0,82	0,41
	TC + WHR + HDLC	0,68			
9	TC* + WHR* + HDLC*	0,73	0,65	0,65	0,65
10	TC* + WHR + HDLC*	0,75	0,71	0,7	0,72
11	TC* + WHR* + HDLC	0,7	0,65	0,65	0,64

(*continued*)

Table 4. (*continued*)

№		AUC	ACC	Sen	Spec
12	TC* + WHR* + HDLC + TG*	0,69	0,63	0,64	0,62
13	**TC* + WHR* + HDLC + IRI**	0,73	0,67	0,67	0,67
14	**Ensemble of models (TC* + LDLC* + WHR* + WHtR* + AIP* + LAP* + UA* + HDLC + IRI)**	0,74	0,65	0,64	0,68
RF					
15	TC* + WHR*	0,67	0,6	0,62	0,57
	TC + WHR + HDLC	0,69			
16	TC* + WHR* + HDLC*	0,71	0,63	0,53	0,77
17	TC* + WHR + HDLC*	0,7	0,66	0,66	0,65
18	TC* + WHR* + HDLC	0,72	0,67	0,65	0,69
19	TC* + WHR* + HDLC + TG*	0,64	0,56	0,56	0,56
20	**TC* + WHR* + HDLC + IRI**	0,69	0,65	0,67	0,61
21	**Ensemble of models (TC* + LDLC* + WHR* + WHtR* + AIP* + LAP* + UA* + HDLC + IRI)**	0,77	0,7	0,7	0,69

During models construction TC $\geq$ 5.9 mmol/l was determined as the basic predictor by the direct selection method (Forward Selection procedure). The step-by-step inclusion of other factors in their structure led to an increase in only certain quality metrics. Their noticeable rise was recorded in the MLR model (6) with a combination of 4 factors: TC $\geq$ 5.9 mmol/l, WHR $\geq$ 0.9 c.u., as well as HDLC and IRI in continuous form. At the same time, the predictive algorithm based on the ensemble of MLR models (7), developed and trained by the Bagging - Bootstrap aggregating method, had the optimal ratio of Sen (0.82) and Spec (0.77) indicators, as well as the maximum AUC value (0,85), corresponding to a high forecast accuracy. This model included a combination of 6 predictive algorithms selected from the list of logistic regressions ranked by the Akaike criterion. The coefficients of the aggregated model were calculated by the weighted averaging method using the formula:

$$\hat{\beta}_{ig} = \frac{\sum_{k=1}^{r} \hat{\beta}_{ik} s_k \Gamma_{ik}}{\sum_{k=1}^{r} s_k \Gamma_{ik}} \tag{1}$$

Where $\hat{\beta}_{ik}$ - coefficient of i-th variable in k-th model; s_k - weight of the k-th model accounting strength of its validity based on the Kullback-Leibler information loss ratio; Γ_{ik} - binary indicator values characterizing the presence of i -th predictor in k-th model.

The number of ensemble model components was selected according to predictive ability calculations depending on regression models number of included inside ensemble (Fig. 1).

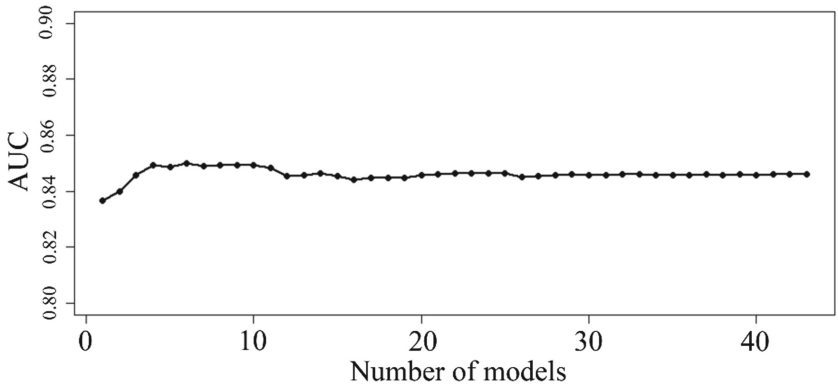

Fig. 1. Accuracy (AUC) dependence by the number of ensemble models

In this ensemble, a combination of 7 factors in categorical form was used as predictors (TC ≥5.9 mmol/l, LDLC > 3.5 mmol/l, WHR ≥0.9 c.u., WHtR ≥ 0.69 c.u., AIP ≥ 3.4 c.u., LAP ≥ 38.5 cm * mmol/l, UA ≥ 356 μmol/l) and 2 - in continuous (HDLC and IRI). Various ensemble model predictors on resulting variable influence degree are shown in the graph (Fig. 2).

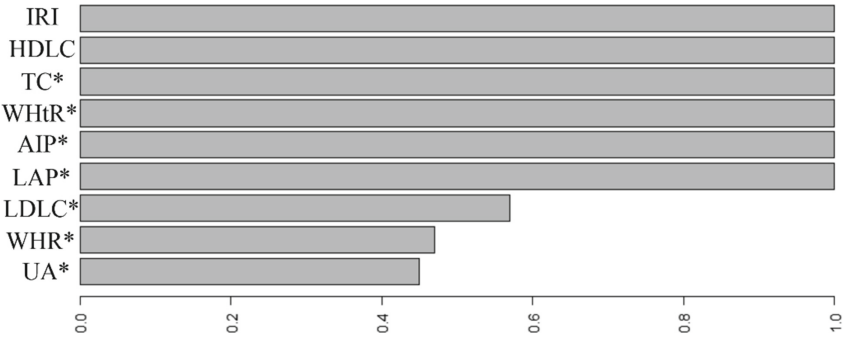

Fig. 2. Relative predictors contribution to the resulting variable

Greatest and equivalent impact on resulting variable (OCAD) were showed by 6 predictors: IRI, HDLC, TC, WHtR, AIP, LAP. Less visible contributions were shown by LDC, WHR и UA. It should be noted that the predictive accuracy of models based on SVM and RF methods was insufficient for any combination of potential predictors.

4 Discussion

Recently, ML methods have been used increasingly as clinical research predictive tools. Their application allows output variables modeling based on input factors, which are characterizing patient clinical and functional status with various diseases, therapy and

surgical treatment options. Modern technologies of collection, storage and processing info allowed to create big volumes repositories of biomedical data, including through the use of electronic medical history and patient records including anamnesis of diseases and their outcomes. This data contains certain knowledge about the causal relationships between patient current state, its dynamic changes during disease process - on the one hand, and the likelihood of developing different outcomes - on the other. The indicators for which such relationships were identified are classified as predictors of the corresponding events. If there are indicators threshold values, which are enchanting the predictive potential, they can be attributed to risk factors for predicted events. In contrast to some other areas of knowledge, where the main forecasting goal is the high developed models accuracy, in clinical medicine, in addition to this criterion, evidence of the predictors validity is being used and required along with predictors threshold values specification, degree of influence on the resulting variable assessment and analyzed factors relationships formalization. The presence of such knowledge increases the "explainability" of the ML models and, therefore, increases the confidence in the developed predictive models.

Accuracy is an objective function that determines the models application effectiveness in various fields of knowledge. At the same time, for medical practice not only the accuracy of predicting events is important, but also the ability to explain the causes, conditions and mechanisms of their development. This, in turn, is an important condition for personalizing prevention and therapy programs. Such approach implementation should be based on the algorithms development for predictors search and validation which are allowed to show clinical interpretation of their relationships with the endpoints of observation, provide a higher prognosis quality and create an evidence base for the predictive models use in clinical practice as support medical decisions tools.

It is generally accepted that as the accuracy of predictive models increases, so does their opacity. In medical research, the most explainable widely represented models are logistic models. At the same time, the best forecast quality can be obtained by other ML methods, for example, random forest, support vector machine and ensemble of models, which are the most problematic from the explainability point of view.

In recent years, definitions of ML methods explainability, causality, interpretability, and "confidence" have been given, and several approaches have been proposed to "whiten" the black box of ML models and develop "responsible" artificial intelligence [16]. One of the important explainable artificial intelligence parts is ensuring that in the developed model only predictors with proven effect on the final variable are being used. This task is most easily solved by linear or logistic regression models. In these cases, several approaches are proposed to improve explainability: hypothesis testing for individual predictors, overall assessment of model quality and forecast accuracy [17, 18]. These approaches can be used for models developed using other ML methods: rule-based learning, decision trees, Bayesian classifiers. Despite the fact that these models have the properties of transparency in construction, decomposition and algorithmization, in most cases, it is difficult for clinicians to interpret the relationship between the input data and the endpoint. For example, if continuous predictors in such models directly or inversely affect the resulting variable, then for physician, in addition to this fact, it is necessary to set a threshold, above (or below) which the variable would be describing as the risk

factor of a particular event. The absence of such threshold reduces explainability level of even simple models. According to other authors, the complexity of models interpretation is also associated with the consideration lack of the predictors mutual influence on the resulting variable [19, 20].

The high prevalence of NOCAD among persons with suspected CAD intensifies the development of prognostic models, which allows to assess the coronary arteries anatomical status before ICA [3]. It is assumed that these technologies usage would reducing the unnecessary risks of ICA and lowering irrational health care costs. In our study, no hemodynamically significant CA lesions were found in 30.4% of patients during ICA, which prompted the authors to assess the CMR factors predictive potential for OCAD at the stage of pretest diagnosis. The reason for this analysis was the well-known key role concept of these factors in pathogenesis [21]. It was previously shown that combined indices (VAI, LAP), including lipid varieties, differ from isolated indicators of lipid metabolism in a more reliable relationship with CA damage [22]. The data from other paper showed that CAD has a closer association with WHtR and VAI indicators than with LAP [13].

In our work, during a multistage selection procedure, we identified potential predictors of OCAD, including anthropometric and metabolic indices which characterizing patients metabolic status.

Obtained results showed that the predictive value of WHR, WHtR and LAP was higher than VAI, which made it possible to use them in predictive models. Insulin resistance is one of the leading pathogenetic factors of the arterial pool atherosclerotic remodeling. The surrogate markers of this syndrome include IRI [3], which demonstrated significant predictive potential in our study (models 6 and 7). A lot of papers indicate the relationship between CAD and hyperuricemia, which is also one of the informative indicators of CMR [23]. In our study, the serum UA level was linearly and nonlinearly related to OCAD, and its concentration $\geq$356 μmol/L increased the probability of verifying hemodynamically significant CA lesions by 2.2 times (Table 3). Which made it possible to use this indicator as a predictor in the ensemble of MLR models (Table 4). A lipid spectrum imbalance with an increased concentration of LDL cholesterol and a decrease in the level of HDL cholesterol has a proven causal relationship with atherosclerotic CA remodeling [24]. In our research the indicators of atherogenic dyslipidemia at the selection stage demonstrated a high predictive potential and were subsequently used to construct predictive algorithms (Tables 1, 2 and 3). Wherein, the predictive properties of HDL cholesterol were manifested in all developed models, while and LDL cholesterol and AIP - only in the model (7).

A comparative analysis of the algorithms predictive accuracy based on modern ML methods demonstrated the advantages of a 6 models ensemble developed by MLR. The quality metrics of this model had maximum values (AUC – 0,85, ACC – 0,80, Sen – 0,82, Spec – 0,77), which corresponded with high forecast accuracy. In our study, it was shown that the CAD Consortium model provided a forecast accuracy by AUC 0.75. In the DISCHARGE 2020 pilot study, the predictive accuracy of the CAD Consortium scale was AUC 0.73 [25]. The elevation of AUC up to 0.85 was achieved by categorizing individual indicators, new predictors usage and models ensemble application. The results

obtained in our study are competitive and comparable to the best predictive indicators in this area.

5 Conclusion

Based on a comprehensive analysis of data characterizing the functional and metabolic patients' status with acute coronary syndrome, the CMR factors were identified and verified as predictors of OCAD and their threshold values were evaluated (TC $\geq$ 5.9 mmol/l, LDLC > 3.5 mmol/l, WHR $\geq$ 0.9 c.u., WHtR $\geq$ 0.69 c.u., AIP $\geq$ 3.4 c.u., LAP $\geq$ 38.5 cm * mmol/l, UA $\geq$ 356 μmol/l).

- The MLR model's ensemble demonstrated the highest prediction accuracy (AUC – 0,85, ACC – 0,80, Sen – 0,82, Spec – 0,77) based on 6 MLR predictive algorithms combination.
- In this study, the models based on SVM and RF had significantly lower predictive accuracy (AUC – 0,74, Acc 0,65, Spec – 0,68, Sen – 0,45 и AUC – 0,77, Acc и Sen – 0,7, Spec – 0,69, respectively).

Prospects for further research in this area are associated with the improvement of predictive models based on expanding the range of predictors and methods of ML, including multilayer artificial neural networks. The limitations of the study associated with an insufficient sample size, limited range of analyzed factors and methods of ML. Conflict of Interest: All authors declare no potential conflict of interest.

References

1. The World Health Organization. The top 10 causes of death (2020). https://www.who.int/ru/news-room/fact-sheets/detail/the-top-10-causes-of-death
2. Sumin, A.N.: The assessment of pretest probability in obstructive coronary lesion diagnostics: unresolved issues. Russ. J. Cardiol. **11**(151), 68–76 (2017). https://doi.org/10.15829/1560-4071-2017-11-68-76
3. Geltser, B.I., Tsivanyuk, M.M., Shakhgeldyan, K.I., et al.: Machine learning for assessing the pretest probability of obstructive and non-obstructive coronary artery disease. Russ. J. Cardiol. **25**(5), 3802 (2020). https://doi.org/10.15829/1560-4071-2020-3802
4. Wang, Z.J., Zhang, L.L., Elmariah, S., et al.: Prevalence and Prognosis of nonobstructive coronary artery disease in patients undergoing coronary angiography or coronary computed tomography angiography: a meta-analysis. Mayo Clin Proc. **92**(3), 329–346 (2017). https://doi.org/10.1016/j.mayocp.2016.11.016
5. Diamond, G.A., Forrester, J.S.: Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. N. Engl. J. Med. **300**, 1350–1358 (1979). https://doi.org/10.1056/NEJM197906143002402
6. Fihn, S.D., Blankenship, J.C., Alexander, K.P., et al.: 2014 ACC/AHA/AATS/PCNA/SCAI/STS focused update of the guideline for the diagnosis and management of patients with stable ischemic heart disease. J. Am. Coll. Cardiol. **64**(18), 1929–1949 (2014). https://doi.org/10.1016/j.jacc.2014.07.017

7. Genders, T.S., Steyerberg, E.W., Alkadhi, H., et al.: A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension. Eur. Heart J. **32**, 1316–1330 (2011). https://doi.org/10.1093/eurheartj/ehr014
8. Bittencourt, M.S., Hulten, E., Polonsky, T.S., et al.: European Society of Cardiology-recommended coronary artery disease consortium pretest probability scores more accurately predict obstructive coronary disease and cardiovascular events than the Diamond and Forrester Score: The Partners Registry. Circulation **134**, 201–211 (2016). https://doi.org/10.1161/circulationaha.116.023396
9. Baskaran, L, Danad, I., Gransar, H., et al.: A comparison of the updated Diamond-Forrester, CAD Consortium, and CONFIRM history-based risk scores for predicting obstructive coronary artery disease in patients with stable chest pain. JACC: Cardiovasc. Imaging 12(7 Pt 2), 1392–400 (2019).https://doi.org/10.1016/j.jcmg.2018.02.020
10. Juarez-Orozco, L.E., Saraste, A., Capodanno, D., et al.: Impact of a decreasing pre-test probability on the performance of diagnostic tests for coronary artery disease. Eur. Heart J. Cardiovasc. Imaging **20**, 1198–1207 (2019). https://doi.org/10.1093/ehjci/jez054
11. Knuuti, J., Wijns, W., Saraste, A., et al.: 2019 ESC Guidelines on the diagnosis and management of chronic coronary syndromes: the task force for diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). Eur. Heart J. **41**, 407–477 (2020). https://doi.org/10.1093/eurheartj/ehz425
12. Montalescot, G., Sechtem, U., Achenbach, S., Members, T.F., et al.: ESC guidelines on the management of stable coronary artery disease: the task force on the management of stable coronary artery disease of the European Society of Cardiology. Eur. Heart J. **34**(38), 2949–3003 (2013). https://doi.org/10.1093/eurheartj/eht296
13. Shalnova, S.A., Deev, A.D., Muromtseva, G.A., et al.: Relation of anthropometric indexes and coronary heart disease. Cardiovasc. Ther. Prevent. **17**(3), 11–16 (2018). https://doi.org/10.15829/1728-8800-2018-3-11-16
14. Geltser, B.I., Orlova-Ilyinskaya, V.V., Vetrova, O.O., et al.: Assessment of cardiometabolic risk factors in various phenotypes of masked hypertension. Cardiovasc. Ther. Prevent. **19**(4), 2422 (2020). https://doi.org/10.15829/1728-8800-2020-2422
15. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**, 16–28 (2014). https://doi.org/10.1016/j.compeleceng.2013.11.024
16. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
17. Bursac, Z., Gauss, C.H., Williams, D.K., et al.: Purposeful selection of variables in logistic regression. Source Code Biol. Med. **3**, 17 (2008). https://doi.org/10.1186/1751-0473-3-17
18. Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, p. 398. Wiley, Hoboken (2013)
19. Mood, C.: Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Sociol. Rev. 26(1), 67–82 (2010). www.jstor.org/stable/4060247
20. Lundberg, S.M., Erion, G., Chen, H., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9
21. Guo, F., Moellering, D.R., Garvey, W.T.: The progression of cardiometabolic disease: validation of a new cardiometabolic disease staging system applicable to obesity. Obesity **22**(1), 110–118 (2014)
22. Amato, M.C., Giordano, C., Galia, M., et al.: AlkaMeSy Study Group: Visceral Adiposity Index: a reliable indicator of visceral fat function associated with cardiometabolic risk. Diab. Care **33**, 920–922 (2010). https://doi.org/10.2337/dc09-1825

23. Biscaglia, S., Ceconi, C., Malagu, M., et al.: Uric acid and coronary artery disease: an elusive link deserving further attention. Int. J. Cardiol. **213**, 28–32 (2016). https://doi.org/10.1016/j.ijcard.2015.08.086
24. Mach, F., Baigent, C., Catapano, A.L., et al.: ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur. Heart J. **41**, 111–188 (2019). https://doi.org/10.1093/eurheartj/ehz455
25. Feger, S., et al.: Clinical pre-test probability for obstructive coronary artery disease: insights from the European DISCHARGE pilot study. Eur. Radiol. **31**, 1–11 (2020). https://doi.org/10.1007/s00330-020-07175-z

Author Queries

Chapter 8

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author and email address have been identified as per the information available in the Copyright form.	
AQ2	As Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city and country names in affiliations 1 and 2. Please check and confirm if the inserted city and country names are correct. If not, please provide us with the correct city and country names.	
AQ3	The ORCID ID "0000-0002-9250-557X" of authors "B. Geltser and V. Rublev" seems to be identical, we have deleted the duplicate. And also provide ORCID ID for author 'V. Rublev', if possible.	
AQ4	Please note that as per Springer guideline private email address of the author will not be displayed.	