Кардиоваскулярная терапия и профилактика. 2022;21(7):3222.
doi:10.15829/1728-8800-2022-3222
ISSN 1728-8800 (Print)
ISSN 2619-0125 (Online)

Верификация субклинического каротидного атеросклероза в рамках риск-стратификации при избыточном весе и ожирении: роль методов машинного обучения в формировании диагностического алгоритма

Дружилов М. А.[1], Кузнецова Т. Ю.[1], Гаврилов Д. В.[2], Гусев А. В.[2,3,4]

[1]ФГБОУ ВО "Петрозаводский государственный университет". Петрозаводск; [2]ООО "К-Скай". Петрозаводск; [3]ФГБУ "Центральный научно-исследовательский институт организации и информатизации здравоохранения" Минздрава России. Москва; [4]ГБУЗ "Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы". Москва, Россия

Цель. Сравнительный анализ математических моделей, полученных с помощью многофакторного логистического регрессионного анализа (МЛРА) с пошаговым включением предикторов и методов машинного обучения (МО), в отношении прогнозирования вероятности наличия субклинического каротидного атеросклероза у нормотензивных пациентов с избыточным весом и ожирением без сердечно-сосудистых заболеваний и/или сахарного диабета.

Материал и методы. Информация о пациентах извлекалась из базы данных платформы Webiomed, критериями включения являлись возраст ⩾18 лет, индекс массы тела ⩾25 кг/м^2, наличие результатов ультразвукового исследования брахиоцефальных артерий, критериями невключения — сахарный диабет и/или сердечно-сосудистые заболевания. Проводился МЛРА с пошаговым включением предикторов, для создания альтернативной модели использовали алгоритмы МО.

Результаты. Общий процент верных классификаций для математической модели, полученной методом МЛРА, составил 73,2%, процент верных отрицательных предсказаний — 80,1%, процент верных положительных предсказаний — 63,4%. Математические модели, созданные с помощью методов МО, характеризуются предсказательной способностью от 75 до 97% при чувствительности от 77 до 92% и специфичности от 80 до 98%.

Заключение. Выявлено существенное превосходство моделей, созданных с помощью методов МО, при изучении комплекса широкодоступных клинических и лабораторно-инструментальных параметров. Интеграция математической модели, созданной с помощью методов МО, в диагностический алгоритм принятия решения о направлении пациента на ультразвуковое исследование брахиоцефальных артерий в рамках проведения риск-стратификации пациенту с "невысоким" риском по шкалам-рискометрам, позволит значительно увеличить ее точность, оптимизируя при этом расходы на оказание медицинской помощи.

Ключевые слова: субклинический каротидный атеросклероз, сердечно-сосудистый риск, ожирение, методы машинного обучения.

Отношения и деятельность. Исследование выполнено на уникальной научной установке "Многокомпонентный программно-аппаратный комплекс для автоматизированного сбора, хранения, разметки научно-исследовательских и клинических биомедицинских данных, их унификации и анализа на базе центра обработки данных с использованием технологий искусственного интеллекта" (регистрационный номер 2075518), при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2021-665.

Поступила 15/02-2022
Рецензия получена 01/04-2022
Принята к публикации 13/06-2022

Для цитирования: Дружилов М. А., Кузнецова Т. Ю., Гаврилов Д. В., Гусев А. В. Верификация субклинического каротидного атеросклероза в рамках риск-стратификации при избыточном весе и ожирении: роль методов машинного обучения в формировании диагностического алгоритма. *Кардиоваскулярная терапия и профилактика.* 2022;21(7):3222. doi:10.15829/1728-8800-2022-3222. EDN WZFTKJ

Verification of subclinical carotid atherosclerosis as part of risk stratification in overweight and obesity: the role of machine learning in the development of a diagnostic algorithm

Druzhilov M. A.[1], Kuznetsova T. Yu.[1], Gavrilov D. V.[2], Gusev A. V.[2,3,4]

[1]Petrozavodsk State University. Petrozavodsk; [2]LLC K-Sky. Petrozavodsk; [3]Federal Research Institute for Health Organization and Informatics. Moscow; [4]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies. Moscow, Russia

Aim. Comparative analysis of mathematical models obtained using multivariate logistic regression (MLR) with stepwise inclusion of predictors and machine learning (ML) for assessing the probability of subclinical carotid atherosclerosis in normotensive overweight

*Автор, ответственный за переписку (Corresponding author):
e-mail: drmark1982@mail.ru
Тел.: +7 (911) 403-19-48

[Дружилов М. А.* — к.м.н., доцент кафедры факультетской терапии, фтизиатрии, инфекционных болезней и эпидемиологии медицинского института, ORCID: 0000-0002-3147-9056, Кузнецова Т. Ю. — д.м.н., зав. кафедрой факультетской терапии, фтизиатрии, инфекционных болезней и эпидемиологии медицинского института, ORCID: 0000-0002-6654-1382, Гаврилов Д. В. — руководитель медицинского направления, ORCID: 0000-0002-8745-857X, Гусев А. В. — к.т.н., директор по развитию бизнеса, эксперт по искусственному интеллекту, с.н.с., ORCID: 0000-0002-7380-8460].

and obese patients without cardiovascular diseases and/or diabetes.

Material and methods. We received data on patients from the Webiomed platform database. The inclusion criteria were age ⩾18 years, body mass index ⩾25 kg/m^2, extracranial artery ultrasound results, while the exclusion criteria included diabetes and/or cardiovascular disease. MLR analysis was carried out with stepwise inclusion of predictors. ML algorithms were used to create an alternative model.

Results. The overall percentage of true results for MLR model was 73,2%, while the proportion of true negative and positive predictions was 80,1% and 63,4%, respectively. Mathematical models created using ML methods are characterized by a predictive value from 75 to 97% with a sensitivity of 77 to 92% and a specificity of 80 to 98%.

Conclusion. A significant superiority of ML models was revealed in the study of available clinical and paraclinical parameters. Integration of ML mathematical models into a diagnostic algorithm for making a decision to refer a low-risk patient for extracranial artery ultrasound will significantly improve its accuracy and cost efficiency.

Keywords: subclinical carotid atherosclerosis, cardiovascular risk, obesity, machine learning methods.

Relationships and Activities. The study was carried out on an original scientific system “Multicomponent software and hardware system for automated collection, storage, markup of research and clinical biomedical data, their unification and analysis based on Data Center with Artificial Intelligence technologies” (№ 2075518) and financially supported by Ministry of Science and Higher Education of the Russian Federation within the Agreement № 075-15-2021-665.

Druzhilov M. A.* ORCID: 0000-0002-3147-9056, Kuznetsova T. Yu. ORCID: 0000-0002-6654-1382, Gavrilov D. V. ORCID: 0000-0002-8745-857X, Gusev A. V. ORCID: 0000-0002-7380-8460.

*Corresponding author:
drmark1982@mail.ru

Received: 15/02-2022
Revision Received: 01/04-2022
Accepted: 13/06-2022

For citation: Druzhilov M. A., Kuznetsova T. Yu., Gavrilov D. V., Gusev A. V. Verification of subclinical carotid atherosclerosis as part of risk stratification in overweight and obesity: the role of machine learning in the development of a diagnostic algorithm. *Cardiovascular Therapy and Prevention*. 2022;21(7):3222. (In Russ.) doi:10.15829/1728-8800-2022-3222. EDN WZFTKJ

БЦА — брахиоцефальные артерии, ДАД — диастолическое артериальное давление, ДИ — доверительный интервал, ИМТ — индекс массы тела, ЛНП — липопротеины низкой плотности, МЛРА — многофакторный логистический регрессионный анализ, МО — машинное обучение, САД — систолическое артериальное давление, СКА — субклинический атеросклероз, ССЗ — сердечно-сосудистые заболевания, ССР — сердечно-сосудистый риск, УЗИ — ультразвуковое исследование, ХС — холестерин.

Ключевые моменты

Что известно о предмете исследования?

- Совершенствование системы риск-стратификации является приоритетным направлением кардиоваскулярной профилактики, учитывая невысокую предсказательную способность классических шкал-рискометров.
- Математические прогностические модели, получаемые при выполнении многофакторного регрессионного анализа с включением различных предикторов, оцениваемых дополнительными лабораторно-инструментальными методами исследования, часто не находят широкого применения в клинической практике ввиду ограниченной доступности последних.

Что добавляют результаты исследования?

- Математические прогностические модели, создаваемые в результате обработки данных методами машинного обучения, характеризуются высокой предсказательной способностью целевого события при анализе комплекса широкодоступных клинических и лабораторно-инструментальных параметров.

Key messages

What is already known about the subject?

- Improving the risk stratification is a priority area in cardiovascular prevention, given the low predictive ability of conventional risk scores.
- Mathematical predictive models obtained using multivariate regression analysis with the inclusion of various predictors evaluated by additional paraclinical methods are often not widely used in clinical practice due to the limited availability of the latter.

What might this study add?

- Mathematical predictive models created using machine learning methods are characterized by a high predictive ability when analyzing a set of widely available clinical and paraclinical parameters.

Введение

В настоящее время сердечно-сосудистые заболевания (ССЗ) продолжают занимать лидирующие позиции в структуре общей смертности в большинстве стран мира, в т.ч. в Российской Федерации, несмотря на определенные успехи соответствующих национальных систем здравоохранения в отношении снижения распространенности неко-

торых основных факторов сердечно-сосудистого риска (ССР) [1]. В этой связи совершенствование системы риск-стратификации в рамках первичной и вторичной кардиоваскулярной профилактики остается, по-прежнему, одним из основных приоритетных направлений [2].

В большинстве используемых сегодня шкал-рискометров у пациентов без клинических симптомов и признаков ССЗ, сахарного диабета и/или хронической болезни почек III-V стадий предикторами величины ССР выступают "классические" факторы риска с доказанной результатами масштабных эпидемиологических исследований прогностической значимостью [2, 3]. Вместе с тем, нередко сердечно-сосудистые осложнения развиваются у лиц, отнесенных по данным шкалам к низкому или умеренному ССР, что свидетельствует об их невысокой предсказательной способности. Отчасти это связано с ограниченным числом предикторов сердечно-сосудистых событий и методами формирования прогностической модели, в связи с чем в существующих клинических рекомендациях определена необходимость использования дополнительных реклассификаторов риска [4, 5].

В качестве таких риск-реклассификаторов могут выступать как уровни "циркулирующих" биомаркеров крови, являющихся отражением протекающих процессов системного воспаления, фиброза, хронического миокардиального повреждения и атерогенеза, дисбаланса системы коагуляции/антикоагуляции [6], так и субклинические органные поражения, верифицируемые соответствующими методами исследования [2].

Среди последних особое место в системе стратификации ССР, согласно положениям европейских и отечественных рекомендаций, занимает субклинический каротидный атеросклероз (СКА), выявляемый при выполнении ультразвукового исследования (УЗИ) брахиоцефальных артерий (БЦА) [3]. Результаты многочисленных эпидемиологических, в т.ч. проспективных, исследований продемонстрировали высокое прогностическое значение СКА в отношении риска развития сердечно-сосудистых осложнений, а его предсказательная ценность сопоставима с таковой при использовании коронарного кальциевого индекса [7-9].

В свою очередь, пациенты с избыточным весом и ожирением и "невысоким" ССР по шкалам-рискометрам являются той группой лиц, для которых оптимизация системы риск-стратификации приобретает особенное значение. С одной стороны, избыточный вес и ожирение также относят к факторам, реклассифицирующим величину ССР [2], с другой — широко обсуждается феномен гетерогенности фенотипов ожирения в отношении ССР, предполагающий различную ассоциацию индекса массы тела (ИМТ) и риска развития сердечно-сосудистых событий [10, 11].

Одним из вариантов оптимизации риск-стратификации у данных пациентов может стать прямая визуализация абдоминальной и/или эктопической висцеральной жировой ткани с помощью ультразвуковых и томографических исследований для диагностики висцерального ожирения, определяющего фенотип ожирения с высоким ССР [12, 13]; вместе с тем, данные методики не являются широкодоступными в клинической практике.

В качестве альтернативного варианта представляется обоснованной тактика совершенствования диагностических алгоритмов по выявлению СКА, отражающего воздействие на сосудистую стенку целого спектра неблагоприятных факторов при избыточном весе и ожирении и принципиально изменяющего категорию риска у данных пациентов [3, 14, 15]. Данные алгоритмы, безусловно, должны базироваться на прогностических математических моделях, создаваемых при анализе совокупности антропометрических, клинических, лабораторных и инструментальных параметров и определяющих высокую вероятность получения положительного результата при направлении пациента на исследование.

Примерами таких прогностических математических моделей являются регрессионные уравнения, созданные в результате выполнения многофакторного логистического регрессионного анализа данных с пошаговым включением предикторов [14, 15]. Шенкова Н. Н. и др. (2017) [14] предложили математическую модель оценки вероятности выявления СКА у пациентов с избыточным весом и ожирением с предсказательной точностью в 89,7%, в которой предикторами выступают уровни грелина, лептина и С-реактивного белка крови и эхокардиографическая толщина эпикардиальной жировой ткани, Дружилова О. Ю. и др. (2016) [15] разработали аналогичное регрессионное уравнение с общим процентом верных предсказаний 91,7%, предикторами в котором являются среднесуточные скорость пульсовой волны и систолическое артериальное давление (САД) в аорте, уровни гликемии натощак и мочевой кислоты крови.

Однако определение вышеуказанных предикторов в приведенных математических моделях ограничено ввиду низкой доступности данных методов исследования в практическом здравоохранении, что обуславливает актуальность создания и внедрения в клиническую практику более простых в использовании, при этом не менее точных прогностических моделей.

Появление таких моделей сегодня стало возможным благодаря внедрению методов машинного обучения (МО) и обработки "больших данных", позволяющих значительно упростить и одновре-

менно улучшить систему риск-стратификации [16], а создаваемые таким способом прогностические математические модели существенным образом превосходят имеющиеся алгоритмы и шкалы в отношении точности оценки вероятности наличия/наступления того или иного события [17-20].

Целью данного исследования является сравнительный анализ математических моделей, полученных с помощью многофакторного логистического регрессионного анализа (МЛРА) с пошаговым включением предикторов и методов МО, в отношении прогнозирования вероятности наличия СКА у нормотензивных пациентов с избыточным весом и ожирением без ССЗ и/или сахарного диабета.

Материал и методы

Для формирования выборки с целью последующего анализа и создания математических прогностических моделей использована база данных платформы Webiomed, включающая деперсонифицированную формализованную информацию из электронных медицинских карт 2,9 млн пациентов, проходивших обследование и лечение в медицинских организациях различных регионов Российской Федерации. В связи с использованием обезличенной медицинской информации заполнение информированного добровольного согласия не предусматривалось.

Критериями включения информации о пациенте в формируемую выборку являлись возраст >18 лет, ИМТ ⩾25 кг/м^2, наличие результатов УЗИ БЦА, критериями невключения — сахарный диабет и/или ССЗ, в т.ч. артериальная гипертензия.

Был определен перечень возможных предикторов высокой вероятности наличия СКА в планируемых математических прогностических моделях, в него были включены клинические и лабораторно-инструментальные параметры, доступные для оценки в широкой клинической практике, при этом окончательный набор параметров устанавливался совместно со специалистом по анализу данных в зависимости от их частоты встречаемости в базе данных Webiomed, в частности заполненность должна была составлять не <40%. В случае частоты встречаемости предиктора в окончательной выборке <100%, допускалась замена отсутствующих значений признака средним значением данного параметра.

Первый этап обработки данных заключался в статистическом анализе окончательной выборки и проведении МЛРА с пошаговым включением предикторов методом последовательного отбора (stepwise), использовались программы Statistica 10 и SPSS 22. Ввиду нормального типа распределения данных результаты представлены средним арифметическим со стандартным отклонением (M±SD) и частотами, для сравнения подгрупп применялся двусторонний t-критерий Стьюдента и критерий χ^2 Пирсона. Величина порогового уровня статистической значимости (p) принята как 0,05.

На втором этапе для создания альтернативной прогностической математической модели использовали библиотеку Scikit-learn для языка программирования Python и 3 алгоритма машинного обучения: Random Forest (“случайный лес”) — ансамблевый алгоритм, основанный на объединении “деревьев решений” для задач классификации, где каждое отдельное “дерево решений” дает предсказание класса, а набравший наибольшее количество голосов класс, становится предсказанием модели; AdaBoostClassifier (“адаптивный бустинг”) — ансамблевый алгоритм, основанный на преобразовании в ходе итеративного процесса слабых классификаторов в сильные для решения проблем классификации; KNeighborsClassifier (“классификация с помощью алгоритма K-ближайшего соседа”) — алгоритм, позволяющий решать задачу классификации путем нахождения максимально похожих между собой объектов и сопоставления их целевых меток (классов)[1]. Данные методы МО являются классическими при работе с числовыми данными и выбраны на основании предварительного анализа имеющейся информации; при этом целевыми параметрами математической прогностической модели были заданы точность (accuracy) ⩾75% и площадь под ROC-кривой ⩾0,75. Данные значения были определены в соответствии с допустимым уровнем ошибок при работе алгоритмов МО, а также на основании предыдущего опыта разработки аналогичных моделей [21].

Исследование выполнено на уникальной научной установке “Многокомпонентный программно-аппаратный комплекс для автоматизированного сбора, хранения, разметки научно-исследовательских и клинических биомедицинских данных, их унификации и анализа на базе центра обработки данных с использованием технологий искусственного интеллекта” (регистрационный номер 2075518), при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2021-665.

Результаты

Из базы данных платформы Webiomed первоначально была получена информация о 5750 пациентах, соответствующих выбранным критериям включения/исключения, из которых СКА был выявлен у 385 (6,7%) человек. После обработки информации специалистом по анализу данных с учетом частоты встречаемости в выборке данных о клинических и лабораторно-инструментальных признаках, предполагаемых в качестве возможных предикторов в математических моделях (на этом этапе сформирована выборка из 1902 пациента), а также после применения функции случайного стратифицированного разделения выборки [22] с целью уравнивания числа лиц с целевым событием (атеросклеротическая бляшка в БЦА) и без такового, окончательная выборка включала информацию о 447 пациентах, из которых 197 (44,1%) имели СКА по результатам УЗИ.

В качестве признаков, рассматриваемых в последующем при создании математических прогностических моделей, отобраны 23 параметра. В таблице 1 приведены их средние значения в целом по

[1] Tavasoli S. Top 10 Machine Learning Algorithms for Beginners: Supervised, Unsupervised Learning and More. https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article. (3 March 2022).

Таблица 1

Клинические и лабораторно-инструментальные параметры окончательной выборки для создания математических прогностических моделей (M±SD, %)

Параметр	СКА+ (n=197)	СКА- (n=250)	Выборка в целом (n=447)
Возраст, лет	56,6±11,9***	44,8±12,1	49,7±13,4
Мужчины, %	36,6**	24,5	29,5
Рост, см	164,0±8,1	162,4±12,8	163,3±10,3
Вес, кг	79,4±11,5**	71,5±11,3	76,1±12,1
ИМТ, кг/м^2	29,5±3,6*	28,6±3,4	29,1±3,5
Окружность талии, см	87,3±7,0**	82,9±8,9	85,9±9,0
Курение, %	14,0	13,0	13,4
САД, мм рт.ст.	123,5±15,4**	120,0±10,7	121,4±12,9
ДАД, мм рт.ст.	77,6±8,3	77,3±7,0	77,4±7,6
Частота сердечных сокращений, в мин	71,9±11,7	72,6±11,8	72,3±11,8
Частота дыхательных движений, в мин	17,6±6,2	17,8±7,0	17,7±6,7
Гликемия натощак, ммоль/л	4,9±0,7*	4,2±0,8	4,4±1,0
Общий ХС, ммоль/л	5,9±3,2	5,7±3,3	5,8±3,6
ХС ЛНП, ммоль/л	3,8±1,1***	3,3±0,9	3,6±1,1
ХС ЛВП, ммоль/л	1,6±0,5	1,6±0,5	1,6±0,5
Триглицериды, ммоль/л	1,7±1,2	1,7±1,6	1,7±1,4
Аспарагиновая аминотрансфераза, Ед/л	25,3±9,6	24,1±13,4	24,5±12,1
Аланиновая аминотрансфераза, Ед/л	22,1±17,3	21,5±11,6	21,7±14,0
С-реактивный белок, мг/л	4,6±2,8	4,0±2,4	4,4±2,7
Креатинин крови, мкмоль/л	87,4±30,2	85,0±26,6	86,4±25,7
Скорость клубочковой фильтрации, мл/мин	78,7±20,0*	87,7±21,4	84,0±22,9
Гипертрофия левого желудочка, %	21,0	15,3	17,7
Индекс массы миокарда левого желудочка, г/м^2	93,6±20,6*	82,8±22,7	88,2±23,6

Примечание: различия между группами СКА+ vs СКА-: * – p<0,05, ** – p<0,01, *** – p<0,001. ДАД — диастолическое артериальное давление, ИМТ — индекс массы тела, ЛВП — липопротеины высокой плотности, ЛНП — липопротеины низкой плотности, САД — систолическое артериальное давление, СКА — субклинический атеросклероз, ХС — холестерин.

выборке, а также в подгруппах с наличием/отсутствием СКА.

Средний возраст пациентов выборки составил 49,7±13,4 лет, средний ИМТ — 29,1±3,5 кг/м^2, среднее САД/диастолическое артериальное давление (ДАД) — 121,4±12,9/77,4±7,6 мм рт.ст., мужчины составили 29,5% случаев, анамнез курения имели 13,4% пациентов.

Сравнительный анализ подгрупп с наличием/отсутствием СКА показал, что пациенты с атеросклеротической бляшкой в сонных артериях были старше, среди них было больше мужчин, они имели более высокие ИМТ, окружность талии, уровни САД, гликемии натощак и холестерина (ХС) липопротеинов низкой плотности (ЛНП), индекс массы миокарда левого желудочка, а также более низкую расчетную скорость клубочковой фильтрации (таблица 1).

Результаты МЛРА данных с пошаговым включением предикторов окончательной выборки представлены в таблице 2. В ходе выполнения анализа среди параметров липидного спектра крови в качестве возможного предиктора регрессионного уравнения не учитывались значения общего ХС для устранения эффекта мультиколлинеарности.

Предикторами вероятности выявления СКА в полученном регрессионном уравнении являются возраст, ИМТ, уровни САД и ХС ЛНП, при этом последний предиктор характеризуется максимальным стандартизированным коэффициентом регрессии — 0,382 (p<0,001). Общий процент верных классификаций для полученной данным методом математической модели составил 73,2%, процент верных отрицательных предсказаний — 80,1%, процент верных положительных предсказаний — 63,4%. При проведении ROC-анализа площадь под ROC-кривой составила 0,76 — 95% доверительный интервал (ДИ): 0,71-0,80 (p=0,006). При отрезном значении ХС ЛНП 3,8 ммоль/л чувствительность и специфичность математической модели составили 70,4 и 82,6%, соответственно.

На втором этапе анализа данных при использовании алгоритмов МО окончательная выборка была разделена на тренировочную и тестовую части в соотношении 80:20. На тренировочной части подбирались гиперпараметры и проводилось обучение модели, на тестовой части проводился замер и сравнение метрик обученных моделей, при полу-

Таблица 2

Результаты МЛРА данных окончательной выборки
в отношении прогнозирования вероятности наличия СКА

Предиктор	Нестандартизированный коэффициент	Стандартизированный коэффициент	p
Возраст	0,192	0,063	<0,01
ИМТ	0,013	0,012	<0,01
САД	0,174	0,071	<0,01
ХС ЛНП	6,251	0,382	<0,001
Константа	-51,243	3,298	<0,001

Примечание: ИМТ — индекс массы тела, ЛНП — липопротеины низкой плотности, САД — систолическое артериальное давление, ХС — холестерин.

Таблица 3

Метрики для различных алгоритмов МО при создании математической прогностической модели оценки вероятности наличия СКА с указанием их значения и 95% ДИ

Метрика	Алгоритм МО		
	Random Forest	AdaBoostClassifier	KNeighborsClassifier
Точность	0,95 (0,90-0,99	0,90 (0,84-0,96)	0,78 (0,70-0,87)
Положительная предсказательная способность	0,97 (0,93-0,99)	0,89 (0,83-0,95)	0,75 (0,66-0,84)
Чувствительность	0,92 (0,86-0,98)	0,87 (0,80-0,94)	0,77 (0,68-0,86)
Специфичность	0,98 (0,95-0,99)	0,92 (0,86-0,98)	0,80 (0,72-0,88)
Площадь под ROC-кривой	0,97 (0,93-0,99)	0,94 (0,89-0,99)	0,88 (0,81-0,95)

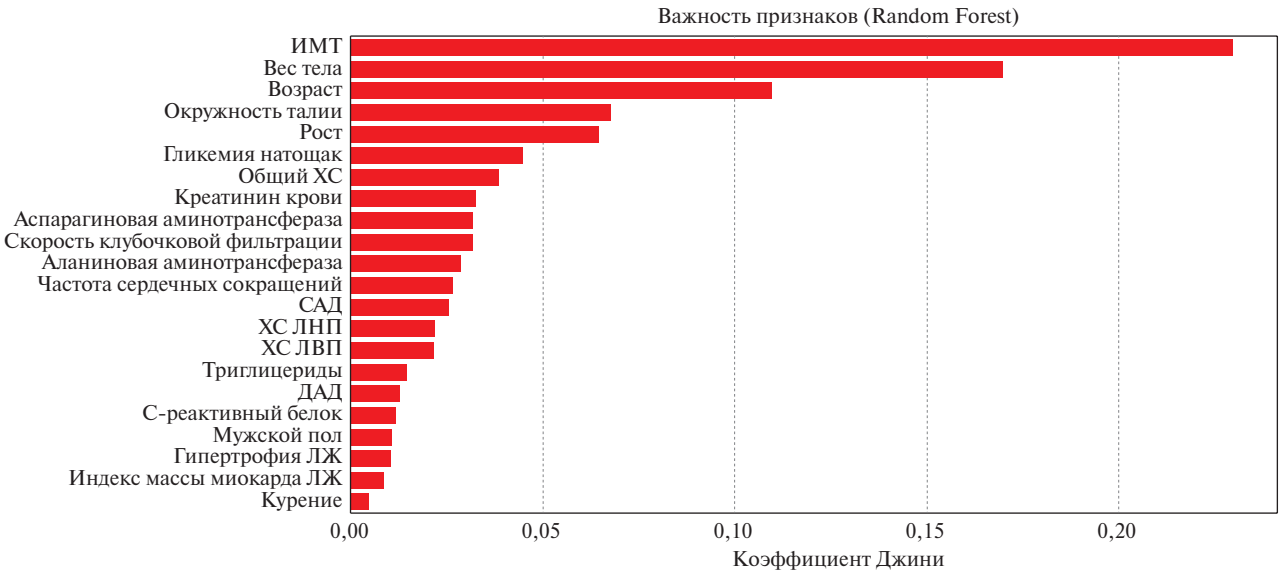

Рис. 1 Значимость клинических и лабораторно-инструментальных параметров при обучении математической прогностической модели алгоритмом Random Forest.

Примечание: ДАД — диастолическое артериальное давление, ЛВП — липопротеины высокой плотности, ЛНП — липопротеины низкой плотности, ЛЖ — левый желудочек, САД — систолическое артериальное давление.

чении лучшего результата информация об обученной модели сохранялась.

Гиперпараметры моделей подбирались путем перебора, по заранее определенным разработчиком значениям[2], с учетом метрик, полученных на тестовой части выборки. В результате для каждой из математических моделей были зафиксированы следующие параметры: Random Forest (число деревьев в ансамбле: 100, критерий разделения вершины дерева: коэффициент джини), AdaBoostClassifier (число деревьев в ансамбле: 50, коэффициент скорости обучения: 1.0), KNeighborsClassifier (число ближайших соседей: 7).

Предсказания моделей интерпретировались следующим образом: алгоритм Random Forest — для каждого дерева оценивалось относительное число

[2] Scikit-learn Machine Learning in Python. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. (28 April 2022).

целевых объектов (вероятность) в листе, далее было взято среднее значение по всем деревьям; алгоритм AdaBoostClassifier — среднее значение по взвешенным предсказаниям каждого классификатора; алгоритм KNeighborsClassifier — метка наиболее представленного класса среди соседних объектов.

Характеристики (метрики) математических моделей в зависимости от используемого алгоритма МО, полученные на тестовой выборке, приведены в таблице 3. Полученные модели оценки вероятности наличия СКА характеризуются предсказательной способностью от 75 до 97%, чувствительностью от 77 до 92%, специфичностью от 80 до 98%, площадью под ROC-кривой от 0,88 до 0,97. С учетом различий величины прогностической точности математических моделей для окончательной работы был выбран алгоритм Random Forest.

При проведении данного алгоритма МО была определена значимость клинических и лабораторно-инструментальных параметров в отношении вероятности прогнозируемого моделью целевого события (рисунок 1). Наиболее важными предикторами, согласно данному алгоритму, являлись возраст, рост и вес тела, ИМТ и окружность талии, уровни гликемии натощак и общего ХС крови.

Обсуждение

Согласно существующим клиническим рекомендациям, выявление атеросклеротических бляшек в сонной артерии у асимптомного в отношении ССЗ пациента с исходно низким или умеренным ССР по шкалам-рискометрам позволяет провести реклассификацию величины риска и, следовательно, пересмотреть целевые уровни показателей липидного спектра с назначением в большинстве случаев гиполипидемической терапии [2, 3]. Кроме того, визуализация СКА может способствовать формированию более высокой приверженности пациента к выполнению рекомендаций по соблюдению здорового образа жизни и приему лекарственных препаратов [23].

В связи с этим внедрение в клиническую практику диагностических алгоритмов, основанных на математических моделях, позволяющих при направлении пациента на УЗИ с высокой вероятностью прогнозировать получение положительного результата в отношении детекции атеросклеротической бляшки сонной артерии, представляется актуальной задачей.

Выбор в качестве исследуемой группы для разработки математических прогностических моделей нормотензивных пациентов с избыточным весом и ожирением, без сахарного диабета и/или ССЗ обусловлен целью включения в анализ пациентов с исходно “невысоким” ССР, у которых оптимизация системы риск-стратификации наиболее целесообразна, а также решения проблемного вопроса прогнозирования риска, связанного с феноменом гетерогенности фенотипов ожирения [10, 11].

Ранее созданные модели прогнозирования вероятности выявления СКА методом МЛРА, предикторами в которых выступали параметры, оцениваемые при выполнении УЗИ, магнитно-резонансной и компьютерной томографии, суточного мониторирования показателей артериальной жесткости, анализов лабораторных маркеров воспаления, фиброза и нейрогуморальной активности висцеральной жировой ткани, характеризовались достаточно высоким уровнем прогностической способности [14, 15]. Вместе с тем, они не могут быть имплементированы в реальную клиническую практику, оставаясь, в первую очередь, предметом интереса научных исследований.

Вместе с тем, применяя данный метод математической обработки информации о пациентах, включающей комплекс широкодоступных в клинической практике клинических и лабораторно-инструментальных параметров, мы не можем получить регрессионное уравнение оценки вероятности целевого события с высокой предсказательной способностью. В настоящем исследовании общий процент верных классификаций данной математической модели составил всего 73,2% при проценте верных положительных предсказаний 63,4%.

Напротив, разработанная математическая прогностическая модель оценки вероятности выявления СКА с помощью алгоритма МО Random Forest отличается высокой положительной предсказательной способностью (97%), чувствительностью (92%) и специфичностью (98%) [20]. Данная прогностическая модель существенно превосходит математическую модель, полученную в ходе МЛРА, при сравнении площади под ROC-кривой: 0,97 (95% ДИ: 0,93-0,99) vs 0,76 (95% ДИ: 0,71-0,80).

Кроме того, данная модель способна функционировать и в случае меньшего, чем использовался при ее создании и валидации набора предикторов, с незначительным снижением своей прогностической способности [16, 21].

Безусловно, перед внедрением в практическое здравоохранение прогностической математической модели, полученной методом МО, необходимо проведение ее внешней валидации на независимой от исходной базе данных; кроме того, целесообразна ее оптимизация в отношении расширения критериев включения при анализе аналогичных изучаемых выборок.

Заключение

Проведенный сравнительный анализ математических прогностических моделей оценки вероятности выявления СКА, созданных в результате обработки данных с помощью МЛРА и методов МО, выявил существенное превосходство последней

в отношении предсказательной способности целевого события при изучении комплекса широкодоступных клинических и лабораторно-инструментальных параметров.

Интеграция математической модели, созданной с помощью методов МО, в диагностический алгоритм принятия решения о направлении пациента на УЗИ БЦА в рамках проведения риск-стратификации пациенту с "невысоким" ССР по шкалам-рискометрам, позволит значительно увеличить ее точность, оптимизируя при этом расходы на оказание медицинской помощи.

Отношения и деятельность. Исследование выполнено на уникальной научной установке "Многокомпонентный программно-аппаратный комплекс для автоматизированного сбора, хранения, разметки научно-исследовательских и клинических биомедицинских данных, их унификации и анализа на базе центра обработки данных с использованием технологий искусственного интеллекта" (регистрационный номер 2075518), при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2021-665.

Литература/References:

1. Boytsov SA, Drapkina OM, Shlyakhto EV, et al. Epidemiology of Cardiovascular Diseases and their Risk Factors in Regions of Russian Federation (ESSE-RF) study. Ten years later. Cardiovascular Therapy and Prevention. 2021;20(5):3007. (In Russ.) Бойцов С. А., Драпкина О. М., Шляхто Е. В. и др. Исследование ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации). Десять лет спустя. Кардиоваскулярная терапия и профилактика. 2021;20(5):3007. doi:10.15829/1728-8800-2021-3007.
2. Cardiovascular prevention 2017. National guidelines. Russian Journal of Cardiology. 2018;(6):7-122. (In Russ.) Кардиоваскулярная профилактика 2017. Российские национальные рекомендации. Российский кардиологический журнал. 2018;(6):7-122. doi:10.15829/1560-4071-2018-6-7-122.
3. Kukharchuk VV, Ezhov MV, Sergienko IV, et al. Diagnostics and correction of lipid metabolism disorders in order to prevent and treat of atherosclerosis. Russian recommendations, VII revision. The Journal of Atherosclerosis and Dyslipidemias. 2020;1(38):7-40. (In Russ.) Кухарчук В. В., Ежов М. В., Сергиенко И. В. и др. Диагностика и коррекция нарушений липидного обмена с целью профилактики и лечения атеросклероза. Российские рекомендации, VII пересмотр. Атеросклероз и дислипидемии. 2020;1(38):7-40. doi:10.34687/2219-8202.JAD.2020.01.0002.
4. Rossello X, Dorresteijn J, Janssen A, et al. Risk prediction tools in cardiovascular disease prevention: A report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP). Eur J Prev Cardiol. 2019;26(14):1534-44. doi:10.1177/2047487319846715.
5. Smirnova MD, Svirida ON, Fofanova TV, et al. Algorithm for predicting cardiovascular events in low/moderate risk patients using traditional and new factors: data from 10-year follow-up study. Cardiovascular Therapy and Prevention. 2021;20(6):2799. (In Russ.) Смирнова М. Д., Свирида О. Н., Фофанова Т. В. и др. Алгоритм прогнозирования сердечно-сосудистых осложнений у больных низкого/умеренного риска с использованием "классических" и "новых" факторов (по данным десятилетнего наблюдения). Кардиоваскулярная терапия и профилактика. 2021;20(6):2799. doi:10.15829/1728-8800-2021-2799.
6. Wong Y, Tse H. Circulating Biomarkers for Cardiovascular Disease Risk Prediction in Patients with Cardiovascular Disease. Front Cardiovasc Med. 2021;8:713191. doi:10.3389/fcvm.2021.713191.
7. Baber U, Mehran R, Sartori S, et al. Prevalence, impact, and predictive value of detecting subclinical coronary and carotid atherosclerosis in asymptomatic adults: the BioImage study. J Am Coll Cardiol. 2015;65(11):1065-74. doi:10.1016/j.jacc.2015.01.017.
8. Nezu T, Hosomi N. Usefulness of carotid ultrasonography for risk stratification of cerebral and cardiovascular disease. J Atheroscler Thromb. 2020;27(10):1023-35. doi:10.5551/jat.RV17044.
9. Li H, Xu X, Luo B, Zhang Y. The Predictive Value of Carotid Ultrasonography With Cardiovascular Risk Factors–A "SPIDER" Promoting Atherosclerosis. Front Cardiovasc Med. 2021;8:706490. doi:10.3389/fcvm.2021.706490.
10. Drapkina OM, Eliashevich SO, Shepel RN. Obesity as a risk factor for chronic noncommunicable diseases. Russian Journal of Cardiology. 2016;(6):73-9. (In Russ.) Драпкина О. М., Елиашевич С. О., Шепель Р. Н. Ожирение как фактор риска хронических неинфекционных заболеваний. Российский кардиологический журнал. 2016;(6):73-9. doi:10.15829/1560-4071-2016-6-73-79.
11. Druzhilov MA, Kuznetsova TY. Heterogeneity of obesity phenotypes in relation to cardiovascular risk. Cardiovascular Therapy and Prevention. 2019;18(1):161-7. (In Russ.) Дружилов М. А., Кузнецова Т. Ю. Гетерогенность фенотипов ожирения в отношении сердечно-сосудистого риска. Кардиоваскулярная терапия и профилактика. 2019;18(1):161-7. doi:10.15829/1728-8800-2019-1-162-168.
12. Chumakova GA, Kuznetsova TY, Druzhilov MA, et al. Visceral adiposity as a global factor of cardiovascular risk. Russian Journal of Cardiology. 2018;(5):7-14. (In Russ.) Чумакова Г. А., Кузнецова Т. Ю., Дружилов М. А. и др. Висцеральное ожирение как глобальный фактор сердечно-сосудистого риска. Российский кардиологический журнал. 2018;(5):7-14. doi:10.15829/1560-4071-2018-5-7-14.
13. Kuznetsova TY, Chumakova GA, Druzhilov MA, et al. Clinical application of quantitative echocardiographic assessment of epicardial fat tissue in obesity. Russian Journal of Cardiology. 2017;(4):81-7. (In Russ.) Кузнецова Т. Ю., Чумакова Г. А., Дружилов М. А. и др. Роль количественной эхокардиографической оценки эпикардиальной жировой ткани у пациентов с ожирением в клинической практике. Российский кардиологический журнал. 2017;(4):81-7. doi:10.15829/1560-4071-2017-4-81-87.
14. Shenkova NN, Veselovskaya NG, Chumakova GA, et al. Risk prediction for subclinical atherosclerotic lesion of brachiocephalic arteries in obese women. Russian Journal of Cardiology. 2017;(4):54-60. (In Russ.) Шенкова Н. Н., Веселовская Н. Г., Чумакова Г. А. и др. Прогнозирование

риска субклинического атеросклероза брахиоцефальных артерий у женщин с ожирением. Российский кардиологический журнал. 2017;(4):54-60. doi:10.15829/1560-4071-2017-4-54-60.

15. Druzhilova OY, Druzhilov MA, Otmakhov VV, et al. Role of assessment of arterial wall stiffness in predicting carotid artery atherosclerosis in patients with abdominal obesity. Terapevticheskii Arkhiv. 2016;88(4):24-8. (In Russ.) Дружилова О. Ю., Дружилов М. А., Отмахов В. В. и др. Роль оценки жесткости артериальной стенки при прогнозировании атеросклероза сонной артерии у пациентов с абдоминальным ожирением. Терапевтический архив. 2016;4(88):24-8. doi:10.17116/terarkh201688424-28.
16. Gusev AV, Gavrilov DV, Novitsky RE, et al. Improvement of cardiovascular risk assessment using machine learning methods. Russian Journal of Cardiology. 2021;26(12):4618. (In Russ.) Гусев А. В., Гаврилов Д. В., Новицкий Р. Э. и др. Совершенствование возможностей оценки сердечно-сосудистого риска при помощи методов машинного обучения. Российский кардиологический журнал. 2021;26(12):4618. doi:10.15829/1560-4071-2021-4618.
17. Narain R, Saxena S, Goyal A. Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. Patient Prefer Adherence. 2016;10:1259-70. doi:10.2147/PPA.S108203.
18. Dimopoulos A, Nikolaidou M, Caballero F, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol. 2018;18(1):179. doi:10.1186/s12874-018-0644-1.
19. Quesada J, Lopez-Pineda A, Gil-Guillén V, et al. Machine learning to predict cardiovascular risk. Int J Clin Pract. 2019;73(10):e13389. doi:10.1111/ijcp.13389.
20. Gavrilov DV, Kuznetsova TYu, Druzhilov MA, et al. Predicting the subclinical carotid atherosclerosis in overweight and obese patients using a machine learning model. Russian Journal of Cardiology. 2022;27(4):4871. (In Russ.) Гаврилов Д. В., Кузнецова Т. Ю., Дружилов М. А. и др. Прогнозирование наличия субклинического каротидного атеросклероза у пациентов с избыточным весом и ожирением при помощи модели машинного обучения. Российский кардиологический журнал. 2022;27(4):4871. doi:10.15829/1560-4071-2022-4871.
21. Gavrilov DV, Serova LM, Korsakov IN, et al. Cardiovascular diseases prediction by integrated risk factors assessment by means of machine learning. Vrach. 2020;31(5):41-6. (In Russ.) Гаврилов Д. В., Серова Л. М., Корсаков И. Н. и др. Предсказание сердечно-сосудистых событий при помощи комплексной оценки факторов риска с использованием методов машинного обучения. Врач. 2020;31(5):41-6. doi:10.29296/25877305-2020-05-08.
22. Iliyasu R, Etikan I. Comparison of quota sampling and stratified random sampling. Biom Biostat Int J. 2021;10(1):24-7. doi:10.15406/bbij.2021.10.00326.
23. Bengtsson A, Norberg M, Ng N, et al. The beneficial effect over 3 years by pictorial information to patients and their physician about subclinical atherosclerosis and cardiovascular risk: Results from the VIPVIZA randomized clinical trial. Am J Prev Cardiol. 2021;7:100199. doi:10.1016/j.ajpc.2021.100199.