УДК 616.12-008.331.1-053.81-07:616.153.915-074 DOI: 10.20969/VSKM.2023.16(3).54-61

ОЦЕНКА ПРЕДИКТОРНОЙ ВОЗМОЖНОСТИ ИСКУССТВЕННО ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ В ПРОГНОЗИРОВАНИИ КАРДИОВАСКУЛЯРНЫХ СОБЫТИЙ

НЕВЗОРОВА ВЕРА АФАНАСЬЕВНА, *ORCID ID: 0000-0002-0117-0349, SCOPUS Author ID: 6603425593; докт. мед. наук, профессор, директор Института терапии и инструментальной диагностики ФГБОУ ВО «Тихоокеанский государственный медицинский университет» Минздрава России, 690002, Россия, Владивосток, проспект Острякова, 2, e-mail: nevzorova@inbox.ru*

ПЛЕХОВА НАТАЛЬЯ ГЕННАДЬЕВНА, *ORCID ID: 0000-0002-8701-7213; SCOPUS Author ID: 6603245380; докт. биол. наук, зав. Центральной научно-исследовательской лабораторией ФГБОУ ВО «Тихоокеанский государственный медицинский университет» Минздрава России, 690002, Россия, Владивосток, проспект Острякова, 2, e-mail: pl_nat@hotmail.com*

ЧЕРНЕНКО ИВАН НИКОЛАЕВИЧ, *ORCID ID: 0000-0001-5261-810X, младший научный сотрудник Центральной научно-исследовательской лаборатории ФГБОУ ВО «Тихоокеанский государственный медицинский университет» Минздрава России, 690002, Россия, Владивосток, проспект Острякова, 2, e-mail: chernencrj2010@mail.ru*

ПРИСЕКО ЛЮДМИЛА ГРИГОРЬЕВНА, *ORCID ID: 0000-0002-3946-2064; очный аспирант специальности 3.1.18 «Внутренние болезни», преподаватель Института терапии и инструментальной диагностики ФГБОУ ВО «Тихоокеанский государственный медицинский университет» Минздрава России, 690002, Россия, Владивосток, проспект Острякова, 2, e-mail: ludmilka.95.95@yandex.ru*

МЕНОВЩИКОВА АННА КОНСТАНТИНОВНА, *ORCID ID: 0000-0002-2375-6973; студентка 5 курса специальности 31.05.01 «Лечебное дело» ФГБОУ ВО «Тихоокеанский государственный медицинский университет» Минздрава России, 690002, Россия, Владивосток, проспект Острякова, 2, e-mail: menovshhikovaan@mail.ru*

Реферат. ***Введение.*** Прогнозирование риска развития кардиоваскулярных заболеваний и неблагоприятных исходов от них является перспективным направлением в медицине. Использование методов машинного обучения при обработке больших баз данных для выявления новых предикторов и установления более сложных и глубоких взаимодействий между ними создает другие прогностические возможности в области персонализированного подхода к оценке рисков. Принципиальным подходом к улучшению прогнозирования является возможность использования методов искусственного интеллекта путем создания комбинации сложных математических моделей и алгоритмов при наличии определенных условий вычислительной мощности и улучшения качества баз данных. ***Цель.*** Оценить точность и надежность моделей на основе машинного обучения для решения задач прогнозирования рисков кардиоваскулярных заболеваний. ***Материалы и методы исследования.*** Данные одномоментного анонимного опроса добровольцев в Приморском крае были собраны в ходе проведения многоцентрового обсервационного исследования «Эпидемиология сердечно-сосудистых заболеваний в регионах Российской Федерации». В исследовании приняли участие 2 131 человек. Отобраны основные и дополнительные предикторы кардиоваскулярных заболеваний для последующего анализа и включения в модели прогнозирования. Модели были созданы и оценены с использованием статистических и нейросетевых библиотек языка программирования Python. Качество моделей определялось с помощью расчета показателя общей площади под кривой операционной характеристики приемника. *Результаты и их обсуждение.* По сравнению с установленным алгоритмом прогнозирования риска с помощью расчета площади под кривой рабочей характеристики приемника машинно-обучающиеся алгоритмы улучшили прогноз: случайный лес +1,7%, логистическая регрессия +3,2%, нейронные сети +3,6%. Алгоритм с наивысшими показателями (нейронные сети) предсказал 320 случаев (чувствительность 70,2%) и 1217 не случаев (специфичность 74,7%), правильно предсказав на 248 (+7,6%) больше пациентов, у которых развилось кардиоваскулярное заболевание, по сравнению с установленной шкалой расчета абсолютного кардиоваскулярного риска Systematic COronary Risk Evaluation. ***Заключение.*** Использование нейронных сетей с многослойным перцептроном, как одного из методов послойного построения алгоритма машинного обучения, является наиболее оптимальным вариантом для создания прогностической модели и в настоящее время дает наиболее достоверный результат.

Ключевые слова: кардииоваскулярный риск, кардиоваскулярные заболевания, Python, нейронная сеть, Random Forest.

Для ссылки: Невзорова В.А., Плехова Н.Г., Черненко И.Н., и др. Оценка предикторной возможности искусственно-интеллектуальных методов в прогнозировании кардиоваскулярных событий // Вестник современной клинической медицины. – 2023. – Т.16, вып.3. – С.54-61. ***DOI:*** 10.20969/VSKM.2023.16(3).54-61.

ASSESSMENT OF THE PREDICTIVE CAPABILITY OF ARTIFICIAL INTELLIGENCE METHODS IN THE PREDICTION OF CARDIOVASCULAR EVENTS

NEVZOROVA VERA A., *ORCID ID: 0000-0002-0117-0349, SCOPUS Author ID:6603425593; D. Med. Sci., professor, director of the Institute of therapy and instrumental diagnostics of Pacific State Medical University, Russia, 690002, Vladivostok, Ostryakov ave., 2, e-mail: nevzorova@inbox.ru (mailto:nevzorova@inbox.ru)*

PLEKHOVA NATALIA G., *ORCID ID: 0000-0002-8701-7213; SCOPUS Author ID: 6603245380; D. Bio. Sci., the Head of Central research laboratory of Pacific State Medical University, Russia, 690002, Vladivostok, Ostryakov ave., 2, tel. 8(423)242-97-78, e-mail: pl_nat@hotmail.com (mailto:pl_nat@hotmail.com)*

CHERNENKO IVAN N., *ORCID ID: 0000-0001-5261-810X; D. Bio. Sci., the Head of Central research laboratory of Pacific State Medical University, Russia, 690002, Vladivostok, Ostryakov ave., 2, tel. 8(423)242-97-78, e-mail: chernencrj2010@mail.ru (mailto:chernencrj2010@mail.ru)*

PRISEKO LUDMILA G., *ORCID ID: 0000-0002-3946-2064, postgraduate student (scientific specialty 3.1.18 «Internal Medicine»), teacher at the Institute of Therapy and Instrumental Diagnostics of the Pacific State Medical University; Address: 2, Ostryakova ave., Vladivostok, Russian Federation 690002; Phone: 89147237764; e-mail: ludmilka.95.95@yandex.ru (mailto: udmilka.95.95@yandex.ru)*
MENOVSHCHIKOVA ANNA K., *ORCID ID: 0000-0002-2375-6973; student of the 5th year of education of the Pacific State Medical University; Address: 2, Ostryakova ave., Vladivostok, Russian Federation 690002; Phone: 89243336797; e-mail: menovshhikovaan@mail.ru (mailto: menovshhikovaan@mail.ru)*

Abstract. ***Introduction.*** Predicting the risk of cardiovascular diseases and adverse outcomes are a promising direction in medicine. Using machine learning methods to process large databases to identify new predictors and establish more complex and deeper interactions between them creates other predictive capabilities in a personalized approach to risk assessment. The new approach to improving forecasting is the ability to use artificial intelligence techniques by creating a combination of sophisticated mathematical models and algorithms under certain conditions of computational power and improved database quality. ***Aim.*** To estimate the accuracy and reliability of models based on machine learning for solving problems of cardiovascular disease risk prediction. ***Materials and Methods.*** Data from a one-stage anonymous survey of volunteers in Primorsky Krai were collected in a multicenter observational study «Epidemiology of Cardiovascular Diseases in the Regions of the Russian Federation». A total of 2 131 participants took part in the study. The main and additional predictors of cardiovascular diseases were selected for further analysis and inclusion in prediction models. Models were created and evaluated using statistical and neural network libraries of the Python programming language. The quality of the models was determined by calculating the total area under the receiver operating characteristic curve. ***Results and discussion.*** Compared to the established risk prediction algorithm by calculating the area under the receiver operating characteristic curve, machine-learning algorithms improved the prediction: random forest +1.7%, logistic regression +3.2%, neural networks +3.6%. The algorithm with the highest performance (neural networks) predicted 320 cases (sensitivity 70.2%) and 1217 non-cases (specificity 74.7%), correctly predicting 248 (+7.6%) more patients who developed cardiovascular disease compared to the established absolute cardiovascular risk calculation scale Systematic Coronary Risk Evaluation. ***Conclusion.*** The use of neural networks with multilayer perceptron, as one of the methods of layer-by-layer construction of machine learning algorithm, is the best option for creating a prognostic model and currently provides the most reliable results.
Key words: cardiovascular risk, cardiovascular disease, Python, neural network, Random Forest.
For reference: Nevzorova VA, Plekhova NG, Chernenko IN, et al. Assessment of the predictive capability of artificial intelligence methods in the prediction of cardiovascular events. The Bulletin of Contemporary Clinical Medicine. 2023; 16(3): 54-61. ***DOI:*** 10.20969/VSKM.2023.16(3).54-61.

Введение. Кардиоваскулярные заболевания (КВЗ) остаются в числе ведущих причин смертности населения. Наиболее доказанным подходом к увеличению продолжительности жизни и снижению смертности от КВЗ является разработка и внедрение научно обоснованных профилактических мероприятий с учетом традиционных и вновь разрабатываемых предикторов [1,2,3,4]. С другой стороны, мощное развитие современных технологий лабораторной диагностики, в том числе молекулярной, предлагает достаточно большое количество параметров для оценки состояния организма, интерпретация которых вызывает затруднения у врачей. С этой точки зрения методы машинного обучения (МО) могут быть полезны как средство определения прогностически значимых предикторов среди миллионов точек фенотипических данных. Большинство из этих переменных можно получить автоматически, в то время как клинические переменные требуют ручного сбора, что отнимает много времени и чревато ошибками [5]. В настоящее время прогнозирование событий является активно развивающимся направлением в эпидемиологии КВЗ, примером чего является Фрамингемское исследование и другие проспективные исследования [6]. Подобные эпидемиологические исследования в области создания прогностических моделей содержат сотни или тысячи переменных и позволяют охарактеризовать субклинические процессы заболевания. Практическое применение таких моделей обосновано выявлением ключевых факторов риска (ФР) развития КВЗ с целью своевременного применения профилактических мер (например, отказ от курения, терапия статинами, контроль артериального давления) [7]. Используемые для этого прогностические шкалы основаны на данных, полученных в результате крупных популяционных исследований, с учетом набора жестко обозначенных ФР. В США для оценки риска КВЗ в первичной медицинской помощи чаще всего используется шкала Framingham Risk Score [8]. В Европе и России широкое применение нашла шкала Systematic COronary Risk Evaluation (SCORE), дополненная в 2021 году (SCORE-2 и SCORE2-Older Persons) [9]. Валидированные шкалы риска имеют ряд ограничений, обусловленных применяемым математическим аппаратом с использованием множественной регрессии, которая предполагает наличие или полное отсутствие связи между ограниченным числом ФР и исходами КВЗ. Попытки повысить прогностическую эффективность прогностических шкал заключаются в выделении экспертными сообществами дополнительных ФР, в частности сахарного диабета, ожирения, хронических неинфекционных воспалительных заболеваний, депрессии и так далее, которые могут повышать кардиоваскулярный риск (КР) без его точного значения [9,10]. С этой же целью, ориентируясь на данные национальной статистики смертности от КВЗ, используются разновидности шкалы SCORE для стран с различными фенотипами КР, разрабатываются национальные шкалы риска, сопоставимые с базой данных (БД) Framingham Heart Study [3,11]. Другим принципиальным подходом к улучшению прогнозирования рисков КВЗ является возможность использования методов искусственного интеллекта путем создания комбинации сложных

математических моделей и алгоритмов при наличии определенных условий вычислительной мощности и улучшения качества БД [12].

Использование методов МО при обработке больших БД для выявления новых предикторов и установления более сложных и глубоких взаимодействий между ними создает другие прогностические возможности в области персонализированного подхода к оценке КР. Однако в настоящее время эта проблема освещена лишь в нескольких исследованиях [13,14,15]. При изучении возможностей использования методов МО внимание исследователей привлекает разрешение и надежность использования того или иного метода для проведения аналитических расчетов и построения прогностических моделей [4].

Цель исследования. Оценить точность и надежность моделей на основе машинного обучения для решения задач прогнозирования рисков кардиоваскулярных заболеваний.

Материалы и методы. Данные одномоментного анонимного опроса добровольцев в Приморском крае были собраны в ходе проведения многоцентрового обсервационного исследования «Эпидемиология сердечно-сосудистых заболеваний в регионах Российской Федерации, (ЭССЕ-РФ)». В исследовании приняли участие 2 131 человек на основании полученного письменного информированного добровольного согласия. Этическое и исследовательское одобрение было получено от Этического комитета Тихоокеанского государственного медицинского университета (соглашение № 46 от 23.11.2014).

Данные добровольцев в возрасте от 24 до 65 лет на момент первого обследования включали восемь основных исходных переменных (пол, возраст, статус курения, систолическое артериальное давление (САД), общую фракцию холестерина (ОФХ), холестерин-липопротеидов высокой плотности (ЛПВП) и диабет), используемых в разработанной American College of Cardiology и American Heart Association (ACC/AHA) 10-летней модели прогнозирования риска [16]. Дополнительно были введены параметры рост, масса тела, индекс массы тела (ИМТ). Последний показатель рассчитывался общепринятым способом путем вычисления отношения массы тела в килограммах к квадрату роста в метрах. Окружность талии (ОТ) измеряли на половине расстояния между нижним краем нижнего ребра и подвздошным гребнем тазобедренной кости на горизонтальном уровне. Значения окружности талии определялись в соответствии с кардио-метаболическим риском.

Для биохимического исследования образцы крови набирались в пробирки натощак от пациентов и в тот же день центрифугировались для отделения сыворотки, которая хранилась в замороженном виде (-80 градусов Цельсия) для последующего анализа. В качестве биохимических маркеров представляли интерес и анализировались следующие: гликемия, мочевая кислота; из липидного профиля ОФХ, ЛПВП и липопротеиды низкой (ЛПНП) плотности, триглицериды (ТГ); С-реактивный белок (СРБ) и натрийуретический пептид (PROBNP). Все переменные определялись колориметрическим методом с использованием автоматического биохимического анализатора Mindray BS-200 (Shenzhen Mindray Bio-Medical Electronics, Китай) и реагентов Alpha Diagnostics (San Antonio, США). Атерогенный индекс плазмы (Atherogenic index of plasma, AIP) и атерогенный коэффициент (Atherogenic coefficient, AC) рассчитывались как log (triglycerides (TG)/ high-density lipoproteins (HDL)) и non-HDL/HDL, соответственно.

Исследование ЭССЕ-РФ инициировано с 1 января 2014 года. Далее в течение 5 лет проводилось динамическое наблюдение за исследуемой когортой жителей Приморского края, промежуточные результаты представлены в публикации [17]. В наблюдаемую группу не вошли лица, имевшие в анамнезе наследственные дислипидемии, а также респонденты с фоновой фармакотерапией гиполипидемическими препаратами.

В дополнение к восьми основным переменным риска развития КВЗ, которые указаны выше и рекомендованные ACC/AHA 2013 [16], алгоритмы МО включают показатели, которые потенциально могут быть связаны с КВЗ. Эти показатели были выбраны на основе литературных данных, которые указывают на их потенциал в качестве предикторов риска развития КВЗ [14-20]. Для общего подхода к работе с недостающими значениями в МО использовалась медианная интерполяция [21]. Была также выдвинута гипотеза, что отсутствующие значения некоторых клинических переменных (например, ИМТ и результаты лабораторных исследований) могут свидетельствовать о том, что некоторые пациенты считают их менее значимыми, учитывая неполную регистрацию нормальных значений ИМТ в медицинских картах первичной медико-санитарной помощи. В общей сложности 26 переменных были проанализированы в моделях МО до базового уровня (за исключением фиктивных переменных для отсутствующих значений) (Таблица 1).

Чтобы сравнить алгоритмы МО риска, исследуемая популяция была разделена в наборе данных на «обучающую» когорту, в которой были получены алгоритмы риска КР, и «валидационную» когорту, в которой алгоритмы были применены и протестированы. Сформированы 2 выборки путем использования функции рандомизации: обучающая (80% участников) и тестовая (20% участников) для оценки эффективности прогноза, в которую вошли данные пациентов с установленным диагнозом КВЗ. Гиперпараметры моделей определялись с помощью сеточного поиска. Для создания алгоритмов МО использовался язык высокого уровня Python 3.9.2 (Python Software Foundation License) с привлечением библиотеки Scilit-learn, представляющей собой набор инструментов для анализа данных. Использовались четыре широко распространенных класса алгоритмов МО: логистическая регрессия [22], случайный лес [23,24] и нейронные сети (НС) [25,26]. Замена отсутствующих данных оценочным значением - импутация для всех моделей осущест-

Таблица 1

Переменные, включенные в алгоритмы МО

Table 1

Variables included in machine learning algorithms

n	Переменные	Характеристика
1	Пол	мужчина 0, женщина 1
2	Возраст	лет
3	Семейный анамнез КВЗ < 60 лет	нет 0, есть 1
4	Курение	нет 0, есть 1
5	Этническая принадлежность	европейцы 0, корейцы/восточные азиаты 1
6	Диабет	нет 0, есть 1
7	Артериальная гипертония (АГ)	нет 0, есть 1
8	Гипоальфахолестеринемия	нет 0, есть 1
9	САД	Мм рт. ст.
10	ИМТ	кг/м²
11	ОТ	мм
12	Частота сердечных сокращений	ударов в 1 минуту
13	Глюкоза	ммоль/л
14	ОФХ	ммоль/л
15	Аполипопротеин А (ApoA-I)	мкмоль/л
16	Аполипопротеин В	мкмоль/л
17	ТГ	ммоль/л
18	ЛПНП	ммоль/л
19	ЛПВП	ммоль/л
20	Креатинин сыворотки	мкмоль/л
21	СРБ	мкмоль/л
22	Инсулин	мкЕД/мл
23	Мочевая кислота	мкмоль/л
24	Фибриноген	г/л
25	Креатинин	мкмоль/л
26	PROBNP	нг/мл

влялась с помощью алгоритма missForest. В качестве базовой модели (БМ) для сравнения выбрана модель шкалы SCORE.

Была представлена описательная характеристика показателей исследуемой популяции, включая количество (%) и среднее значение (SD) для категориальных и непрерывных переменных, соответственно. Точность алгоритмов прогнозирования МО, которые разрабатывались на основе обучающей выборки, оценивалась с помощью 10-кратной стратифицированной кросс-валидации с использованием меры общей площади под кривой операционной характеристики приемника (area under the curve - receiver operating characteristic curve (AUC-ROC)). Кроме того, используя пороговые значения, соответствующие 10-летнему риску КВЗ >7,5%, как рекомендовано ACC/AHA [2], для сравнения наблюдаемых и ожидаемых прогнозов был использован анализ бинарной классификации случаев и неслучаев при валидации когорт. Этот процесс обеспечил чувствительность, специфичность, положительную предсказательную ценность и отрицательную предсказательную ценность. Статистический анализ, оценивающий эффективность алгоритма, проводился с использованием STATA13MP4 (разработчик StataCorp LLC, США).

Протокол исследования был одобрен независимым Междисциплинарным комитетом по этике Федерального государственного бюджетного образовательного учреждения высшего образования «Тихоокеанский государственный медицинский университет» Министерства здравоохранения Российской Федерации (протокол № 5 от 17.01.2022 г.).

Результаты и их обсуждение. Всего в БД на начальном уровне (1 января 2014 года) было 2 500 добровольцев, которые соответствовали критериям отбора. После исключения 369 пациентов с ошибками кодирования (то есть нечисловые записи для САД, ОФХ) и чрезвычайно отдаленных наблюдений (> 5 SD от среднего значения), когорта анализа состояла из 2131 добровольца. Затем когорта была случайным образом разделена на выборку из 1 440 добровольцев (80%) для создания алгоритма МО, а оставшая-

Таблица 2

Заболевания, согласно МКБ-10, выявленные среди исследуемых в зависимости от наличия АГ

Table 2

ICD-10 diseases detected among subjects depending on the presence of AH

Заболевание, согласно МКБ-10	Частота случаев среди лиц без АГ	Частота случаев среди лиц с АГ
Стенокардия	-	51,06%
Нарушения ритма (фибрилляция и трепетание предсердий)	14,44%	11,06%
Инфаркт миокарда в прошлом	9,09%	5,53%
Неуточненный инсульт с АГ	-	6,81%

ся выборка составила 360 добровольцев для его тестирования.

На момент инициации исследования возраст участников составил в среднем 45,75 (11,7) лет, доля лиц мужского пола - 41%. В течение наблюдения (5-95-й процентили: 3,44,7 лет) КВЗ выявлены среди 422 исследуемых в возрасте 60,2±5,6 лет у мужчин и 61,1±4,8 лет у женщин. У лиц без артериальной гипертензии (АГ) (n=1398) КВЗ имели место у 13,38% (n=187) обследуемых, тогда как среди лиц с АГ (n=733) - у 32,06% (n=235). Согласно кодам «Международной статистической классификации болезней» (МКБ-10), в исследуемой популяции зафиксировано наличие следующих нозологий, которые указаны в таблице 2.

При этом абсолютный риск фатальных событий в группе пациентов с гипертонией составил 0,037, что значительно превышало показатель для лиц без гипертонии (0,017, p<0,05) при относительном значении равном 2,146.

Переменные, составляющие входные данные для моделей МО, были обучены на когорте из 1 440 пациентов с 244 случаями КВЗ (35,1%), развившимися в течение 5-летнего периода наблюдения. В качестве первого шага для оценки значимости переменной в обучающем наборе данных мы использовали обучение сгенерированного случайного леса. Для каждого элемента в процессе построения модели обучающей выборки для МО рассчитывалась ошибка на невыбранных выборках (out-of-bag error), значение которой затем усреднялось, применяясь ко всему случайному лесу. Второй шаг заключался в оценке важности отображения параметра после обучения, для чего их значения смешивались и, опять же, вычислялась out-of-bag error. Важность параметра оценивалась путем усреднения по всем деревьям разницы в показателях ошибки out-of-bag до и после смешивания значений. При этом нормальность величины таких ошибок рассчитывалась по стандартному отклонению. Использование приведенного выше метода перестановок позволяет оценить снижение значимости каждой переменной после пермутации этой переменной для каждого дерева решений.

НС – представляет собой модель, построенную по образу и подобию биологических сетей, где каждый нейрон выполняет математические задачи. С помощью библиотеки МО, написанной на языке программирования Python, анализировались данные. В нашей работе мы использовали модель многослойного перцептрона - простейшего вида НС. В такой модели входной сигнал распространяется от слоя к слою в прямом направлении. Сам многослойный перцептрон состоял из трех основных элементов: входного слоя, включающего множество нейронов (20 переменных), нескольких слоев скрытых вычислительных нейронов и одного выходного слоя (наличие установленного КВЗ). Обучение и оптимизация НС проводились в соответствии с алгоритмом адаптивной оценки момента Adam, за 1000 эпох, объем единовременно поступающих данных составлял 20 единиц. Наибольший вклад в реализацию модели вносят следующие переменные: возраст, ОТ, ИМТ, глюкоза, частота сердечных сокращений. Такие параметры, как ЛПНП, аполипопротеин В, ОФХ и другие, имели наименьшее значение в реализации модели.

Для оценки уровня точности методов построения прогностических моделей (деревьев решений, множественной линейной регрессии и НС) исполь-

Таблица 3

Качество прогнозирования рассматриваемых моделей

Table 3

Prediction quality of the models under consideration

Модель	AUC-ROC	Изменение в процентах
SCORE	0.724	БМ
Нейронная сеть	0.789	8.98
Линейная регрессия	0.779	7.60
Random Forest	0.759	4.83

зовался AUC-ROC, рекомендованный для оценки качества моделей на несбалансированных данных. ROC — это линия от (0,0) до (1,1) в координатах отношения доли истинно положительных показателей к доле ложно положительных. Чем выше AUC-ROC, тем лучше классификатор. Модель риска SCORE служила основой для сравнения (AUC 0,728, 95% доверительный интервал (ДИ) 0,723±0,735). Точность прогнозирования в зависимости от дискриминации (с-статистика AUC) представлена в таблице 3 для всех моделей. Достаточно высокие значения ROC-индекса AUC - 0,789 с отношением к БМ SCORE 8,98 свидетельствуют о наибольшей пригодности НС для прогнозирования выбранных диагнозов КВЗ.

БМ SCORE правильно предсказала 1316 случаев из 2100 общих случаев, в результате чего чувствительность составила 62,7%. Алгоритм случайного леса привел к увеличению сети на 191 случай КР по сравнению с БМ, что привело к увеличению чувствительности до 65,3%. НС работали лучше всего для повышения точности прогнозирования, что показало увеличение КР до 1564 случаев (чувствительность 74,5%) правильно предсказанных случаев, соответственно.

Модель Random forest заключается в использовании ансамбля элементов, каждый из которых в единичном исполнении дает очень низкое качество классификации, но при учете их большого количества результат становится достоверным [23,24]. Таким образом, использование ансамбля деревьев решений на основе обучающей выборки позволяет, с учетом источников случайности, уменьшить значение дисперсии в модели, так как по отдельности каждое дерево решений демонстрирует высокий уровень дисперсии и ошибок [25,26]. Объединение различных деревьев решений делает эту модель одной из наиболее точных и часто используемых для решения задач классификации и регрессии. Для определения относительной прогностической значимости переменных мы использовали подход «после факта», чтобы ранжировать их вклад в конечный результат, выдаваемый прогностической моделью. Преимуществом множественной линейной регрессии, по сравнению с простой, является использование в модели нескольких входных переменных, что позволяет увеличить долю объясненной дисперсии выходной переменной, а вместо прямой используется гиперплоскость [27]. Эта модель чаще всего используется в медицине из-за простоты использования и хорошей точности. При добавлении в модель каждой новой переменной коэффициент детерминации увеличивается, и, чтобы исключить мультиколлинеарность, выбранная нами модель с использованием минимального набора независимых переменных объясняет наибольшую долю дисперсии зависимой переменной. Для этого мы использовали метод наименьших квадратов с минимизацией суммы квадратов разницы между зависимой переменной и ее значениями, предсказанными линейной функцией. Таким образом, используя метод наименьших квадратов, из первоначального набора переменных (26 показателей) были удалены показатели с наименьшей значимостью и оставлены показатели с наибольшей значимостью для прогноза (15 показателей).

На основе последних параметров была построена регрессионная модель прогноза для расчета риска развития КВЗ. При построении НС мы использовали модель многослойного перцептрона, в которой входной сигнал распространяется от слоя к слою в прямом направлении. Именно такая структура позволяет НС хорошо справляться с решением не только линейных, но и нелинейных задач [28]. В серии предварительных вычислительных экспериментов с использованием построения прогностических моделей (деревья решений, множественная линейная регрессия и НС) наилучший результат показала модель многослойного перцептрона, которая представляет собой модель послойного построения алгоритма МО. Полученные показатели с помощью AUC-ROC свидетельствовали о высоком уровне прогнозирования КВЗ с помощью методов МО. В то же время модель деревьев решений оказалась предпочтительной для применения решения задачи определения значимости ФР в развитии заболевания. Преимуществом разработанной модели является одномоментное рассмотрение целого набора данных: лабораторных маркеров, совокупности признаков, таких как возраст, ОТ, ИМТ. Кроме того, представленный аналитический алгоритм способен обрабатывать большой объем данных, полученных в результате многоцентрового эпидемиологического исследования «ЭССЕ-РФ» в Приморском крае. Фокусировка на конкретных КВЗ в сочетании с данными первичного обследования условно здоровых лиц, которым еще не поставлен диагноз, позволит повысить точность прогнозирования риска развития заболевания. Для повышения качества моделей необходима дополнительная разработка различных методов предварительной обработки медицинских данных.

По сравнению с установленным алгоритмом прогнозирования риска (AUC 0,728, 95% ДИ 0,723±0,735), машинно-обучающиеся алгоритмы улучшили прогноз: случайный лес +1,7% (AUC 0,745, 95% ДИ 0,739±0,750), логистическая регрессия +3,2% (AUC 0,760, 95% ДИ 0,755±0,766), НС +3,6% (AUC 0,764, 95% ДИ 0,759±0,769). Алгоритм с наивысшими показателями (НС) предсказал 320 случаев (чувствительность 70,2%) и 1217 неслучаев (специфичность 74,7%), правильно предсказав на 248 (+7,6%) больше пациентов, у которых развилось кардиоваскулярное заболевание, по сравнению с установленным алгоритмом SCORE.

Использование в исследовании трех алгоритмов МО показало различия в степени важности ФР в зависимости от метода моделирования. Модель на основе деревьев решений показала превосходство в важности предикторов некоторых факторов, а нейросетевая модель превосходила по эффективности случайные леса. НС имеют категориальные переменные с возможностью анализировать состояние здоровья, связанное с КР, группируя

пациентов с похожими характеристиками в каждой группе. Это может помочь в дальнейшем изучении различных прогностических ФР и разработке в будущем новых подходов к прогнозированию риска и алгоритма развития КВЗ. Наконец, важность отсутствующих значений или отсутствия ответов часто не оценивается с помощью традиционных инструментов для прогнозирования риска КВЗ [16,28]. В то время как наше исследование демонстрирует, что отсутствующие значения, в частности, рутинные биометрические переменные, такие как ИМТ, являются независимыми предикторами КР.

Заключение. Использование НС с многослойным перцептроном, как одного из методов послойного построения алгоритма МО, является наиболее оптимальным для создания прогностической модели и в настоящее время дает наиболее надежный результат. Значительное повышение надежности моделей на основе МО связано с тем, что в них используется большее количество предикторов, чем при расчете по шкале SCORE. Использование набора персональных данных, полученных в результате опроса, клинических и лабораторных исследований, для МО при прогнозировании заболеваний позволяет целенаправленно создавать индивидуальные терапевтические рекомендации.

Прозрачность исследования. *Исследование не имело спонсорской поддержки. Авторы несут полную ответственность за предоставление окончательной версии рукописи в печать.*

Декларация о финансовых и других взаимоотношениях. *Работа является частью государственного задания Минздрава РФ 222040500008-5 «Технологии искусственного интеллекта в фенотипировании тканевого и системного ремоделирования и прогнозировании исходов на этапах развития хронических неинфекционных заболеваний у лиц различных этнических групп».*

ЛИТЕРАТУРА / REFERENCES

1. Conroy R, Pyörälä K, Fitzgerald Ae, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European heart journal. 2003; 24 (11): 987–1003. DOI:10.1016/S0195-668X(03)00114-3
2. Greenland P, Alpert JS, Beller GA, et al. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association task force on practice guidelines developed in collaboration with the American Society of Echocardiography, American Society of Nuclear Cardiology, Society of Atherosclerosis Imaging and Prevention, Society for Cardiovascular Angiography and Interventions, Society of Cardiovascular Computed Tomography, and Society for Cardiovascular Magnetic Resonance. Journal of the American College of Cardiology. 2010; 56 (25): 50–103. DOI: 10.1016/j.jacc.2010.09.001
3. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study British Medical Journal. 2007. 335 (7611): 136. DOI:10.1136/bmj.39261.471806.55
4. Sjostrom L, Lindroos AK, Peltonen M, et al. Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. New England Journal of Medicine. 2004; 351 (26): 2683–2693. DOI:10.1056/NEJMoa035622
5. Rios R, Miller RJH, Hu LH, et al. Determining a minimum set of variables for machine learning cardiovascular event prediction: results from REFINE SPECT registry. Cardiovascular Research. Advance online publication. 2022. 118 (9): 2152–2164. DOI:10.1093/cvr/cvab236
6. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. Circulation. 2010; 121 (15): 1768–1777. DOI: 10.1161/CIRCULATIONAHA.109.849166
7. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European Heart Journal. 2017. 38 (23): 1805–1814. DOI:10.1093/eurheartj/ehw302
8. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008; 117 (6): 743–753. DOI:10.1161/CIRCULATIONAHA.107.699579
9. Mortensen MB, Falk E. Limitations of the SCORE-guided European guidelines on cardiovascular disease prevention. European Heart Journal. 2017; 38 (29): 2259–2263. DOI:10.1093/eurheartj/ehw568
10. Каблуков Д.А., Крукович Е.В., Плехова Н.Г. и др. Персонифицированный подход к оценке и коррекции факторов риска неинфекционной заболеваемости. Тихоокеанский медицинский журнал. – 2019. – No3. – С. 52–56. [Kablukov DA, Krukovich EV, Plekhova NG, et al. Personificirovannyj podhod k ocenke i korrekcii faktorov riska neinfekcionnoj zabolevaemosti [Personified approach to assessment and correction of risk factors of non-communicable diseases]. Tihookeanskij medicinskij zhurnal [Pacific Medical Journal]. 2019; 3: 52–56. (In Russ.)]. DOI:10.17238/PmJ1609-1175.2019.3.52-56
11. Woodward M, Brindle M, Tunstall-Pedoe H, et al. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). Heart. 2007; 93 (2): 172–176. DOI:10.1136/hrt.2006.108167
12. Piepoli MF, Hoes AW, Agewall S, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). European Heart Journal. 2016; 37 (29): 2315–2381. DOI:10.17863/CAM.17
13. 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
14. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circulation Research. 2017; 121 (9): 1092–1101. DOI:10.1161/circresaha.117.311312
15. 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. DOI:10.1371/journal.pone.0174944
16. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014; 129(25 Suppl 2): 49–73. DOI: 10.1161/01.cir.0000437741.48606.98
17. Plekhova NG, Nevzorova VA, Brodskay TA, et al. Association of Cardiovascular Events and Blood Pressure and Serum Lipoprotein Indicators Based on Functional Data Analysis as a Personalized Approach to the Diagnosis. Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing. 2020; 1295: 278–293. DOI: 10.1007/978-3-030-63319-6_24
18. Emerging Risk Factors Collaboration, Kaptoge S, Di Angelantonio E, et al. C-Reactive Protein, Fibrinogen, and Cardiovascular Disease Prediction. New England Journal of Medicine. 2012; 367 (14): 1310–1320. DOI:10.1056/ NEJMoa1107477
19. Osborn DP, Hardoon S, Omar RZ, et al. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. JAMA Psychiatry. 2015; 72 (2): 143–151. DOI:10.1001/jamapsychiatry.2014.2133
20. Wannamethee SG, Shaper AG, Perry IJ. Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke. Stroke; a journal of cerebral circulation. 1997; 28 (3): 557–563. DOI: doi: 10.1161/01.str.28.3.557
21. Batista GE, Monard MC. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence. 2003; 17 (5-6): 519–533. DOI:10.1080/713827181
22. Hosmer DW, Lemeshow JrS, Sturdivant RX. Applied Logistic Regression. Wiley; 3rd edition; 2013; 528 p.
23. Breiman L. Random Forests. Machine Learning. 2001; 45(1): 5–32. DOI:10.1023/A:1010933404324
24. Rigatti SJ. Random Forest. Journal of insurance medicine. 2017; 47 (1): 31–39. DOI:10.17849/insm-47-01-31-39.1
25. Hagan MT, Demuth HB, Beale MH, De Jesus O. Neural Network Design; 2nd edition. Martin Hagan; 2014; 800 p.
26. Uddin S, Khan A, Hossain, ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making. 2019; 19 (1): 281. DOI:10.1186/s12911-019-1004-8
27. Zahid FM, Heumann C. Multiple imputation with sequential penalized regression. Statistical methods in medical research. 2019; 28 (5): 1311–1327. DOI:10.1177/0962280218755574
28. Quesada JA, Lopez-Pineda A, Gil-Guillén VF, et al. Machine learning to predict cardiovascular risk. International journal of clinical practice. 2019; 73 (10): e13389. DOI:10.1111/ijcp.13389