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

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

Методы машинного обучения в прогнозировании исходов и рисков сердечно-сосудистых заболеваний у пациентов с артериальной гипертензией (по материалам ЭССЕ-РФ в Приморском крае)

Невзорова В. А.[1], Плехова Н. Г.[1], Присеко Л. Г.[1], Черненко И. Н.[1], Богданов Д. Ю.[2], Мокшина М. В.[1], Кулакова Н. В.[1]

Цель. Оценить возможность применения технологий искусственного интеллекта в прогнозировании исходов и рисков сердечно-сосудистых заболеваний (ССЗ) у пациентов с артериальной гипертонией (АГ).

Материал и методы. Создана компьютерная программа для извлечения в полуавтоматическом режиме информации из анкет респондентов, проанализированы библиотеки с предобработкой данных. Проведен анализ основных и дополнительных показателей факторов риска развития ССЗ (35 параметров) у 2131 человек при выполнении регионального этапа "ЭССЕ-РФ, 2014-2019гг". Для создания модели прогнозирования применен высокоуровневый язык Python 2.7 с использованием объектно-ориентированного программирования и включением обработки исключений с поддержкой многопоточных вычислений. С помощью функции рандомизирования сформированы обучающая (488 человек) и тестовая (245 человек) выборки, в которые вошли данные пациентов с установленным диагнозом АГ.

Результаты. Распространенность АГ среди обследуемых составила 34,39%. К значимым признакам для прогнозирования развития ССЗ отнесены антропометрические параметры, наличие курения, данные биохимического анализа крови (общий холестерин, АроА, АроВ, глюкоза, Д-димер, С-реактивный белок). В результате 5-летнего наблюдения ССЗ установлены у 235 человек (32,06%) с АГ и у 187 человек (13,38%) без АГ; показатели смертности составили 1,27% у лиц с АГ и 1,12% без АГ. Абсолютный риск фатального исхода среди лиц с АГ (0,037) был значимо выше ($p<0,05$), чем у пациентов без АГ (0,017). Для построения нейросети (НС) применяли базовую модель Sequential из библиотеки Keras. При машинном обучении в качестве входных данных использовались 26 значимых для развития ССЗ переменных и выходными были определены 9 нейронов, которые соответствовали количеству установленных сердечно-сосудистых событий. Созданная НС обладала предсказующей способностью до 97,9%, что превышало таковую на 34,9% шкалы SCORE.

Заключение. Полученные данные указывают на важность фенотипирования факторов риска с использованием антропометрических маркеров и параметров биохимии крови, при определении их значимости в списках 20 топ-предикторов для прогнозирования ССЗ. Основанный на языке Python метод машинного обучения обеспечивает прогнозирование ССЗ согласно стандартным оценкам риска.

Ключевые слова: факторы риска сердечно-сосудистых заболеваний, артериальная гипертензия, искусственный интеллект.

Отношения и деятельность: работа была поддержана грантом РФФИ 19-29-01077.

[1]ФГБОУ ВО Тихоокеанский государственный медицинский университет Минздрава России, Владивосток; [2]КГБУЗ Владивостокская клиническая больница № 1, Владивосток, Россия.

Невзорова В. А.* — д.м.н., профессор, директор Института терапии и инструментальной диагностики, главный внештатный специалист по терапии Дальневосточного федерального округа, ORCID: 0000-0002-0117-0349, Плехова Н. Г. — д.б.н., профессор, зав. Центральной научно-исследовательской лабораторией, ORCID: 0000-0002-8701-7213, Присеко Л. Г. — ординатор Института терапии и инструментальной диагностики, ORCID: 0000-0002-3946-2064, Черненко И. Н. — м.н.с., Центральная научно-исследовательской лаборатория, ORCID: 0000-0001-5261-810X, Богданов Д. Ю. — врач-кардиолог, ORCID: 0000-0002-8388-5566, Мокшина М. В. — доцент Института терапии и инструментальной диагностики, ORCID: 0000-0003-3663-1560, Кулакова Н. В. — доцент Института терапии и инструментальной диагностики, ORCID: 0000-0001-6473-5653.

*Автор, ответственный за переписку (Corresponding author):
nevzorova@inbox.ru

АроА — аполипопротеин А, АроВ — аполипопротеин В, АГ — артериальная гипертензия, АД — артериальное давление, ИМТ — индекс массы тела, ЛП(а) — липопротеин "а" малое, ЛВП — липопротеиды высокой плотности, ЛНП — липопротеиды низкой плотности, МКБ-10 — международная статистическая классификация болезней, МО — машинное обучение, НС — нейросеть, ОТ — окружность талии, ОХС — общий холестерин, ПД — пульсовое давление, САД — систолическое АД, СРБ — С-реактивный белок, ССЗ — сердечно-сосудистые заболевания, ТГ — триглицериды, ФР — фактор(-ы) риска, ЧСС — частота сердечных сокращений.

Рукопись получена 13.02.2020
Рецензия получена 21.02.2020
Принята к публикации 12.03.2020

Для цитирования: Невзорова В. А., Плехова Н. Г., Присеко Л. Г., Черненко И. Н., Богданов Д. Ю., Мокшина М. В., Кулакова Н. В. Методы машинного обучения в прогнозировании исходов и рисков сердечно-сосудистых заболеваний у пациентов с артериальной гипертензией (по материалам ЭССЕ-РФ в Приморском крае). *Российский кардиологический журнал*. 2020;25(3):3751. doi:10.15829/1560-4071-2020-3-3751

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

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

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

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

Results. The prevalence of HTN among subjects was 34,39%. There were following significant factors for predicting CVD: anthropometric parameters, smoking, bio-

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

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

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

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

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

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

Received: 13.02.2020 **Revision Received:** 21.02.2020 **Accepted:** 12.03.2020

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

Наиболее часто для прогнозирования риска развития сердечно-сосудистых заболеваний (ССЗ) разрабатываются модели с использованием многомерных регрессионных методов анализа, что объединяет информацию об ограниченном числе точно установленных факторов риска (ФР). Такой алгоритм предполагает, что все учитываемые факторы связаны с исходами ССЗ линейным образом при ограничении или отсутствии взаимодействия между ними. По причине такого ограничительного подхода к моделированию и предикторам подобные алгоритмы, в частности, шкалы Framingham, SCORE и DECODE демонстрируют недостаточную прогностическую эффективность при ограниченном наборе признаков [1]. В различных предметных областях, в т.ч. и в медицине, наиболее качественный результат при построении прогностической модели показывает метод интеллектуального анализа данных, а именно, создание глубоких нейросетей (НС). На данный момент появилось достаточное количество готовых к использованию библиотек, на основании которых после незначительной модификации возможно применение НС для решения практических задач. Подобные методы, основанные на машинном обучении (МО), повышают эффективность прогнозирования рисков за счет использования объемных хранилищ данных при независимой идентификации новых предикторов риска и сложных взаимодействий между ними. Известно небольшое количество исследований, где были изучены потенциальные преимущества использования подходов МО для прогнозирования риска ССЗ. Продемонстрировано, что, по сравнению с приведенными выше шкалами оценки, МО значительно повышает точность прогнозирования риска ССЗ, увеличивая количество пациентов, которые могли бы получить пользу в большей степени от профилактического лечения до проявления клинически значимых признаков [2-4].

В настоящей работе приведена потенциальная ценность использования подходов МО для построения модели прогнозирования риска ССЗ с учетом показателей артериального давления (АД). Проведен проспективный анализ данных, полученных при одномоментном обследовании 2800 жителей Приморского края без наличия ССЗ на исходном уровне. Данное обследование было проведено с 2014 по 2019гг при выполнении регионального этапа Российского многоцентрового эпидемиологического исследования "Эпидемиология сердечно-сосудистых заболеваний в регионах РФ" (ЭССЕ-РФ). Для разработки модели прогнозирования риска использовали современный автоматизированный высокоуровневый язык Python, его библиотеку Keras с открытым программным кодом и набором функций и надстроек. Обучение и оптимизация НС проводились по алгоритму Adam. Осуществлялась оценка прогностической значимости НС в общей популяции здоровых лиц, включая клинически значимую подгруппу пациентов с артериальной гипертонией (АГ).

Цель исследования — оценить возможность применения технологий искусственного интеллекта в прогнозировании исходов и рисков ССЗ у пациентов с АГ.

Материал и методы

В период выполнения регионального этапа "ЭССЕ-РФ, 2014-2019гг" проведено одномоментное обследование жителей Приморского края [5]. Исследование проведено по стандартам надлежащей клинической практики (Good Clinical Practice) и принципам Хельсинской Декларации. Использована программа кардиологического скрининга, принятая при популяционных исследованиях, включающая комплекс показателей оценки донозологического состояния организма. Для формирования репрезентативной выборочной совокупности применялся сплош-

ной метод путем индивидуального приглашения на обследование. Критерии включения: подписание информированного согласия, возраст (от 24 до 65 лет), полное заполнение разработанной анкеты, наличие информации по ФР развития ССЗ. Критерии невключения: отказ от участия в исследовании, наличие онкологических заболеваний. Всего включено в исследование 2800 человек, завершили программу обследования к 2019г 2131 человек (76,1%). Путем систематического отбора с использованием алгоритма корректировки данных проведено формирование выборок в компьютерной программе извлечения в полуавтоматическом режиме информации из анкет респондентов (рис. 1).

Анализировали частоту встречаемости основных ФР: избыточной массы тела с вычислением индекса массы тела (ИМТ) по формуле Кетле (масса тела, кг/рост, м^2), окружность талии (ОТ), уровни АД и пульсового давления (ПД); частота сердечных сокращений (ЧСС), факт курения, гиподинамии; 10-летнего риска развития ССЗ по шкале SCORE (у лиц ⩾40 лет и ⩽65 лет) на основании пола, возраста, систолического АД (САД), общего холестерина (ОХС) и статуса курения. Уровни АД оценивали в соответствии с рекомендациями [6], согласно которым показатели 140/90 мм рт.ст. и выше относили к АГ. Наличие наследственной отягощенности, факта курения и злоупотребления алкоголем уточняли при сборе анамнеза. Проводили определение показателей липидного профиля (ОХС, триглицериды (ТГ), липопротеиды низкой (ЛНП) и высокой плотности (ЛВП), липопротеин “а” малое (ЛП(а)), аполипопротеин А (АроА), аполипопротеин В (АроВ)); уровня глюкозы, креатинина, мочевой кислоты, Д-димера, С-реактивного белка (СРБ).

Рис. 1. Блок-схема, описывающая общие рамки исследования.

Для нейросетевого анализа данных применялся высокоуровневый язык Python 2.7 (лицензия Python Software Foundation License) на основании объектно-ориентированного программирования с включением механизма обработки исключений и поддержки многопоточных вычислений. Для инициации МО после анализа библиотек Python (TensorFlor, Keras) использовался Keras с открытым программным кодом и набором функций и надстроек. Обучение и оптимизация НС проводились по алгоритму Adam (адаптивный момент оценки, adaptive moment estimation) с вычислением адаптивной скорости обучения для каждого параметра. Аналогично импульсу, Adam сохраняет экспоненциально убывающее среднее значение прошлых квадратов градиентов AdaDelta и прошлых градиентов M (t).

Статистический анализ данных проводился с помощью программного обеспечения Stata 11.2 и R 3.2.1 (StataCorp LP, США). Непрерывные переменные представлены медианами значений с межквартильными интервалами (МКИ), сравнение проводилось с помощью параметрического критерия Стьюдента. Для сравнения дискретных переменных использовался критерий χ^2 или критерий Фишера. Кумулятивные вероятности развития ССЗ оценивались по методу Каплана-Мейера и сравнивались с помощью логарифмического рангового критерия. Для оценки влияния различных переменных на риск развития ССЗ использовались одно- и многофакторные регрессионные модели пропорциональных рисков Кокса. Представлены отношения рисков и их 95% доверительные интервалы с соответствующими значениями p. Статистически значимым считалось значение $p<0,05$. Эффективность алгоритмов прогнозирования МО, разработанных на основе обучающей когорты, оценивалась с использованием коэффициента валидации.

Исследование поддержано грантом РФФИ 19-29-01077.

Результаты и обсуждение

Характеристика исследуемой популяции. Для 2131 участника определена полная информация с использо-

Таблица 1

Клинико-лабораторная характеристика исследуемых лиц

Показатели (M±m)	Группа здоровых лиц (n=1398)	Группа лиц с АГ (n=733)
Средний возраст, лет	42,68±11,45	51,56±9,82*
Рост, см	168,82±0,25	168,02±0,36
Вес, кг	75,62±0,44	85,44±0,63*
Индекс массы тела, кг/м^2	26,47±0,14	30,35±0,22*
Окружность талии, см	85,97±0,39	96,71±0,53*
Среднее САД, мм рт.ст.	123,87±0,27	156,48±0,58*
Среднее ДАД, мм рт.ст.	75,39±0,22	89,19±0,39*
Среднее ПД, мм рт.ст.	48,49±0,22	67,29±0,50*
Среднее ЧСС, уд./мин	74,91±0,32	77,75±0,68*
Общий холестерин, ммоль/л	5,49±0,03	5,87±0,05*
ЛНП, ммоль/л	3,49±0,03	3,76±0,04*
ЛВП, ммоль/л	1,45±0,01	1,4±0,01*
Триглицериды, ммоль/л	1,24±0,02	1,67±0,04*
ЛП(а), мг/дл	20,19±0,65	20,62±0,92
АроА, г/л	1,76±0,01	1,81±0,02*
АроВ, г/л	0,82±0,01	0,89±0,01*
Глюкоза, ммоль/л	5,23±0,03	5,86±0,08*
Креатинин, мкмоль/л	69,12±0,44	71,55±0,77*
Мочевая кислота, мкмоль/л	315,87± 2,71	356,38±4,01*
Д-димер, мкг/л	212,30±7,16	186,05±4,93*
С-реактивный белок, мг/л	2,63±0,16	3,78±0,25*

Примечание: различия значимы при * — p<0,05.

Сокращения: ПД — пульсовое давление, ЛП(а) — липопротеин класса а, АроА — аполипопротеин А, АроВ — аполипопротеин В, ДАД — диастолическое артериальное давление, ЛВП — липропротеины высокой плотности, ЛНП — липропротеины низкой плотности, САД — систолическое артериальное давление, ЧСС — частота сердечных сокращений.

ванием рандомизированного алгоритма корректировки данных в компьютерной программе (табл. 1). Средний возраст участников в начале исследования составил 45,75 (11,7) лет, лица мужского пола — 874 (41%). В течение периода наблюдения 5 лет (5-95-й процентиль: 3,4-4,7 года), выявлено 422 случая ССЗ в возрастном диапазоне 60,2±5,6 года для мужчин и 61,1±4,8 год для женщин. В группе лиц без АГ (n=1398) наличие установленных ССЗ отмечено у 13,38% (n=187) исследуемых, тогда как среди лиц с АГ (n=733) у 32,06% (n=235). По частоте встречаемости, согласно кодам “Международной статистической классификации болезней” (МКБ-10), стенокардия обнаружена у 51,06% лиц с АГ, нарушение ритма (фибрилляция и трепетание предсердий) у 14,44% лиц без АГ и у 11,06% с АГ, перенесенный в прошлом инфаркт миокарда у 9,09% лиц без АГ и у 5,53% лиц с АГ, инсульт неуточненный отмечен у 6,81% с АГ. Абсолютный риск фатальных событий в группе пациентов с АГ составил 0,037, что значимо превышало показатель для лиц без АГ (0,017, p<0,05) при относительном значении 2,146.

Предикторы риска развития ССЗ. Выявленные статистические различия средних величин показателей изучаемых ФР между группами лиц без АГ (группа сравнения) и с наличием АГ (основная группа) представлены в таблице 1.

У всех обследованных установлено наличие избыточной массы тела. Среди лиц с АГ среднее значение ИМТ было выше, по сравнению с лицами без АГ (p=0,00001). ОТ мужчин не превысила рекомендованного значения с наибольшим показателем 98,5±0,67 см в основной группе. У женщин основной группы ОТ составила 95,13±0,78 см vs 82,89±0,49 см в группе сравнения (p<0,0001).

Уровень ПД превышал порог среди лиц с АГ, при этом максимальное значение (68,88±0,71 мм рт.ст.) отмечено у женщин. Среднее значение ЧСС в группах находилось в пределах допустимых цифр.

Средний уровень ОХС превышал нормальное значение у всех исследуемых. Максимальное среднее значение фракции ЛНП (3,88±0,05 ммоль/л) отмечено у женщин с АГ. Значимое различие по уровню ЛВП выявлено между женщинами с АГ и без АГ (p=0,007). Средний уровень ТГ превысил норму только у мужчин с АГ (1,77±0,08 ммоль/л).

Гликемия натощак >5,6 ммоль/л считается ФР возникновения сахарного диабета и ССЗ. Между группами исследуемых выявлены значимые различия (p<0,001). Превышение порогового значения отмечено у всех лиц основной группы.

Средний уровень креатинина не превышал допустимых значений в 100% случаев. Однако по данному

Таблица 2

Показатели, включенные в алгоритм машинного обучения (данные пациентов с АГ)

Факторы риска (M±m)	ССЗ (n=293)	Без ССЗ (n=440)	Значение P
Женщины, %	67,8	42,65	-
Средний возраст, лет	52,67±0,85	52,16±0,50	0,61
Курение, %	33,9	39,53	-
Рост, см	165,87±0,94	167,48±0,52	0,02*
Вес, кг	85,12±1,57	85,25±0,91	0,94
Индекс массы тела, кг/м^2	31,03±0,55	30,42±0,30	0,33
Окружность талии, см	97,33±1,35	97,47±0,79	0,92
Окружность бедер, см	107,43±0,95	107,40±0,56	0,97
Среднее САД, мм рт.ст.	156,29±1,59	157,55±0,85	0,48
Среднее ДАД, мм рт.ст.	87,52±0,91	89,64±0,58	0,05
Среднее ПД, мм рт.ст.	67,91±0,73	69,08±1,43	0,46
Среднее ЧСС, усл. ед.	76,83±1,27	77,26±0,62	0,76
Глюкоза, ммоль/л	5,77±0,13	5,96±0,13	0,31
Общий холестерин, ммоль/л	5,92±0,12	6,01±0,07	0,51
ЛВП, ммоль/л	1,40±0,03	1,40±0,02	1
ЛНП, ммоль/л	3,83±0,10	3,86±0,06	0,79
Триглицериды, ммоль/л	1,66±0,09	1,68±0,06	0,85
ЛП(а), мг/дл	20,09±0,4	20,22±0,2	0,77
АпоА, г/л	1,84±0,04	1,85±0,02	0,82
АпоВ, г/л	0,89±0,02	0,92±0,01	0,18
С-реактивный белок, мг/л	3,34±0,61	3,78±0,34	0,52
Креатинин, мкмоль/л	68,93±0,95	72,04±1,30	0,05
Мочевая кислота, мкмоль/л	353,56±8,98	354,29±5,56	0,94
Д-димер, мкг/л	178,99±9,82	185,92±6,16	0,55

Примечание: различия значимы при * — p<0,05.

Сокращения: ПД — пульсовое давление, ЛП(а) — липопротеин класса а, АроА — аполипопротеин А, АроВ — аполипопротеин В, ДАД — диастолическое артериальное давление, ЛВП — липропротеины высокой плотности, ЛНП — липропротеины низкой плотности, САД — систолическое артериальное давление, ЧСС — частота сердечных сокращений.

ФР исследуемые группы статистически значимо (p=0,006) отличались между собой. Наибольшее среднее значение ФР составило 318,80±4,96 мкмоль/л среди женщин с АГ.

ЛП(а) представляет собой атерогенную фракцию липидов и имеет прогностическое значение для развития атеросклероза и ССЗ, в частности, ишемической болезни сердца. Допустимые значения показателя находятся в пределах 5-18 мг/дл. В основной группе и группе сравнения данный ФР составил 20,62±0,93 и 20,19±0,65мг/дл, соответственно, без статистически значимости различий (p=0,704). Интересно, что среди мужчин без АГ среднее значение ЛП(а) было несколько выше (20,70±1,09 мг/дл) по сравнению с мужчинами основной группы (18,16±1,28 мг/дл) без статистически значимой достоверности.

Предполагается, что уровни АроА и АроВ могут быть решающими в определении риска возникновения атеросклероза, особенно, когда другие показатели липидного спектра не превышают норму и/или нет клинических проявлений сосудистого поражения [7]. Статистически значимые различия между значениями в основной группе и группе сравнения АроА (p=0,025) и АроВ (p=0,00001).

При сравнении средних значений Д-димера выявлено его более высокое содержание у лиц без АГ, в отличие от исследуемых с АГ (табл. 1), с значимым различием (p=0,0026). Также значимое (p=0,0001) различие обнаружено между женщинами группы сравнения (236,51±10,56 мкг/л) и основной группы (190,51±5,72 мкг/л).

При сравнении средних значений между группами уровень СРБ был выше у лиц с АГ по сравнению с лицами без АГ независимо от пола. Различия между исследуемыми группами оказались статистически значимы (p=0,0001).

Таким образом, в результате проведенной статистической обработки полученных данных к значимым признакам для прогнозирования развития ССЗ отнесены антропометрические параметры (рост, вес, ИМТ, ОТ) и показатели биохимического анализа крови (уровень ОХС, показатели гликемии натощак, содержание фракций АроА и АроВ, Д-димера и СРБ).

Модель МО для прогнозирования исходов ССЗ у пациентов с АГ. Для создания НС используются различные языки программирования, где поддерживаются базовые математические операторы и многомерные массивы. К ним относятся такие интерпретируемые Си языки, как Python, с использованием

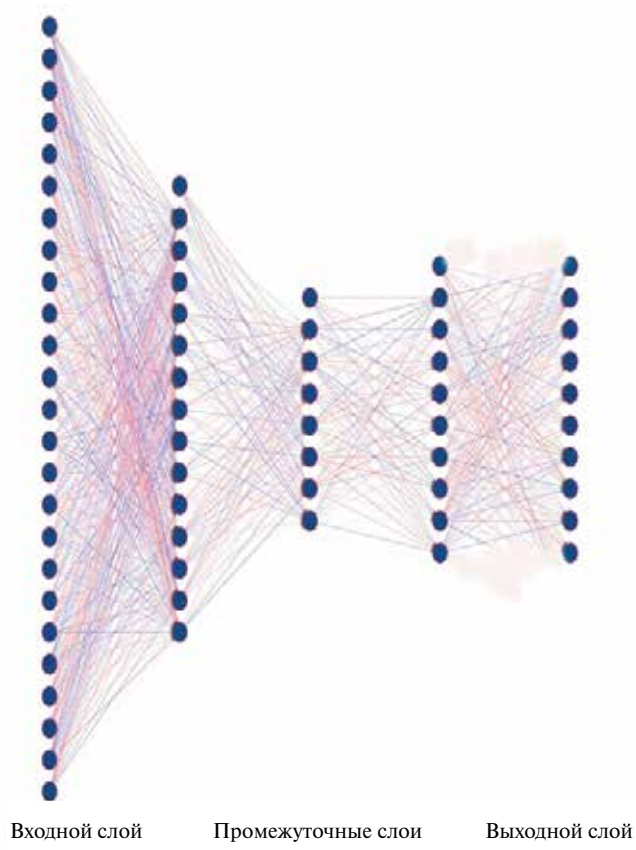

Рис. 2. Модель нейронной сети.

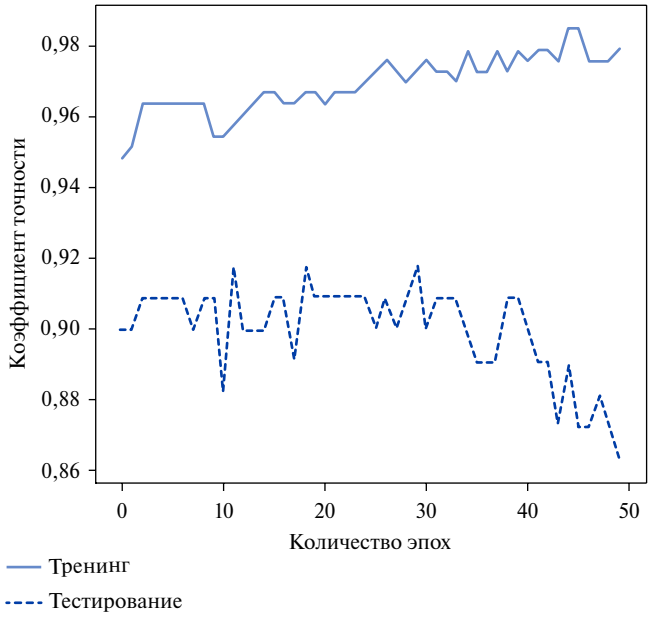

Рис. 3. Изменение точности нейронной сети в процессе обучения и тестирования (фрагмент тестирования 50 эпох).

которого мы провели МО. Для построения НС использовали базовую модель Sequential из библиотеки Keras, которая представлена набором слоев разной плотности с возможностью их комбинирования для постройки многослойного персептрона по Румельхарту. Из общего массива данных с помощью функции рандомизирования X_train, X_test, y_train, y_test=train_test_split (X,Y,test_size=0,40, random_state=42) сформировано 2 выборки: обучающая (488 человек) и тестовая (245 человек, рис. 1), в которые вошли данные пациентов с установленным диагнозом АГ. Из всех обследуемых с АГ (n=733) количество курящих составило 144 человек, курили и бросили — 170, некурящие — 419. В качестве входных данных использовались 26 наиболее важных переменных, что составило входной слой прогностической модели (табл. 2, рис. 2). Скрытые слои были определены эмпирическим путем: первый слой включал 15 нейронов (позиции, где происходит умножение матрицы весовых коэффициентов и матрицы входных данных предыдущих нейронов); второй содержал результат минимизации ошибки — 8 нейронов и вводился третий с целью уточнения прогноза, который охватывал 10 нейронов. Выходной слой состоял из 9 нейронов, каждый из которых соответствовал количеству событий, соответствующих диагнозу согласно МКБ-10 (табл. 3).

Обучение и оптимизация НС проводились по алгоритму адаптивной оценки момента Adam, который вычисляет адаптивные скорости обучения для каждого параметра. В дополнение к хранению экспоненциально убывающего среднего значения прошлых квадратов градиентов, таких как AdaDelta, Adam аналогично импульсу сохраняет экспоненциально убывающее среднее значение прошлых градиентов M (t). Алгоритм Adam отличается от других адаптивных методов быстрой скоростью обучения модели и эффективностью. Изменение точности НС в процессе обучения и тестирования представлено на рисунке 3.

Объем выборки для МО составил 66,6% от всех обследуемых с АГ. Обучение и оптимизация НС проводилась в 1000 эпохах, объем подаваемых единовременно данных составил 32 единицы. В результате тестирования с применением алгоритма адаптивной оценки момента Adam точность НС достигла 97,9%, а величина потерь находилась в пределах 10^{-7}-10^{-8} (рис. 3). При проведении тестирования точность сети снизилась до 95,5% (рис. 3).

Классификационный анализ. Чтобы оценить клиническую значимость наших результатов, мы сравнили нашу модель с традиционной оценкой SCORE при прогнозировании риска ССЗ (порог для начала терапии АГ в соответствии с рекомендациями). В этой рабочей точке базовая модель SCORE правильно предсказала 145 ССЗ из 465 случаев, чувствительность составила 61,7%, коэффициент прогноза составил 1,5%. Наша модель автопрогнозирования с использованием метода МО правильно предсказала 230 ССЗ из 733 обследуемых, что привело к чувствительности 97,9%. Полученная разница в показателях соответ-

Таблица 3

Стратификация обследованных лиц от 24 до 65 лет с АГ без наличия ССЗ в начале исследования в зависимости от наличия первого сердечно-сосудистого события после 5-летнего периода наблюдения

№ п/п	№ кода по МКБ-10	Наименование заболевания	Количество, человек	Удельный вес
1	I20.8	Другие формы стенокардии	120	51,06%
2	I48	Фибрилляция и трепетание предсердий	26	11,06%
3	I25. 2	Перенесенный в прошлом инфаркт миокарда	13	5,53%
4	I64.0	Инсульт неуточнённый	16	6,81%
5	I70.2	Атеросклероз артерий конечностей	26	11,06%
6	I20.1	Стенокардия с подтвержденным спазмом	14	5,96%
7	I69.3	Последствия инфаркта мозга	7	2,98%
8	I69.4	Последствия инсульта неуточненные	6	2,55%
9	I20.0 + I21.9	Острый коронарный синдром (нестабильная стенокардия и острый инфаркт миокарда)	7	2,98%

Сокращение: МКБ-10 — международная статистическая классификация болезней.

ствует 36,2% увеличения точности предсказания возникновения ССЗ в случае использования методов МО.

Заключение

Результаты проведенного исследования показывают, что методы МО могут эффективно использоваться для значимого прогнозирования рисков развития ССЗ при фенотипических эпидемиологических исследованиях. Метод, основанный на языке Python, обеспечивает прогнозирование ССЗ по стандартным оценкам риска. Применение функции рандомизирования для отбора переменных с последующими методами регрессии Кокса позволяет улучшить прогнозирование результатов без проблем переоснащения и несовпадения при учете нелинейностей. Результаты также указывают на важность расширенного фенотипирования обследованных лиц с использованием антропометрических маркеров, параметров биохимии крови, при определении их значимости в списках 20 топ-предикторов для прогнозирования ССЗ.

В известных исследованиях MESA показано, что такие показатели как возраст, воспаление и сосудистые заболевания доминируют в прогнозе смерти. Также указывается, что нарушение метаболизма глюкозы и артериальная гипертония приводят к прогнозированию инсульта, а маркеры субклинического атеросклероза занимают центральное место в прогнозировании общих ССЗ — будь они ограничены ишемической болезнью сердца или вовлекают системное сосудистое русло [8]. Примененный нами метод МО уникален тем, что демонстрирует закономерности изменения составляющих предикторов, различающихся для конкретных исходов заболевания. Достаточно высокие показатели точности прогнозирования (от 86 до 98%) указывают на приемлемость использования метода МО при расчете риска развития ССЗ. Преимуществом проведенного исследования является рассмотрение совокупности антропометрических данных, результатов лабораторных анализов и других важных предикторов развития ССЗ. Таким образом, МО в сочетании с расширенным фенотипированием повышает точность прогнозирования сердечно-сосудистых событий в популяции обследуемых с наличием такого ФР их развития, как АГ. Разработанные подходы позволяют подойти к более точному пониманию маркеров субклинических заболеваний без априорных предположений о причинности их возникновения.

Отношения и деятельность: исследование поддержано грантом РФФИ 19-29-01077.

Литература/References

1. Siontis GC, Tzoulaki I, Siontis KC, et al. Comparisons of established risk prediction models for cardiovascular disease: systematic review. BMJ. 2012;344:e3318. doi:10.1136/bmj.e3318.
2. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PLoS One. 2017;12(4):e0174944. Published 2017 Apr 4. doi:10.1371/journal.pone.0174944.
3. Ahmad T, Lund LH, Rao P, et al. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. Journal of the American Heart Association. 2018;7(8):e008081. doi:10.1161/JAHA.117.008081.
4. Plekhova NG, Nevzorova VA, Rodionova LV, et al. Scale of Binary Variables for Predicting Cardiovascular Risk Scale for predicting cardiovascular risk. Proceedings of the 2018 3rd Russian-Pacific Conf. on computer technology and applications (RPC). 2018. doi:10.1109/RPC.2018.8482216.
5. The Scientific and Organizing Committee of the project of the ESSE-RF. Epidemiology of cardiovascular diseases in various regions of Russia (ESSE-RF). Rationale and design of the study. Prophylactic medicine. 2013;6:25-34. (In Russ.) Научно-организационный комитет проекта ЭССЕ-РФ. Эпидемиология сердечно-сосудистых заболеваний в различных регионах России (ЭССЕ-РФ). Обоснование и дизайн исследования. Профилактическая медицина. 2013;6:25-34.
6. Mancia G, Fagard R, Narkiewicz K, et al. Recommendations for the treatment of arterial hypertension. ESH/ESC 2013. Russian Journal of Cardiology. 2014;(1):7-94. (In Russ.) Mancia G, Fagard R, Narkiewicz K, et al. Рекомендации по лечению артериальной гипертонии. ESH/ESC 2013. Российский кардиологический журнал. 2014;(1):7-94. doi:10.15829/1560-4071-2014-1-7-94.
7. Plekhova NG, Nevzorova VA, Rodionova LV, et al. Indicators of lipoprotein metabolism in young patients with arterial hypertension. Bulletin of modern clinical medicine. 2019;4:44-51. (In Russ.) Плехова Н.Г., Невзорова В. А., Родионова Л. В. и др. Показатели липопротеинового метаболизма у пациентов молодого возраста с артериальной гипертонией. Вестник современной клинической медицины. 2019;4:44-51. doi:10.20 969/VSKM.2019.12(4).44-51.
8. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res. 2017;121(9):1092-101. doi:10.1161/CIRCRESAHA.117.311312.