https://doi.org/10.29296/25877305-2020-05-08

Предсказание сердечно-сосудистых событий при помощи комплексной оценки факторов риска с использованием методов машинного обучения

Д.В. Гаврилов[1],
Л.М. Серова[1], кандидат технических наук,
И.Н. Корсаков[1], кандидат физико-математических наук,
А.В. Гусев[1], кандидат технических наук,
Р.Э. Новицкий[1],
Т.Ю. Кузнецова[2], доктор медицинских наук
[1]ООО «К-Скай», Петрозаводск
[2]Петразоводский государственный университет
E-mail: dgavrilov@webiomed.ai

Актуальность: *профилактика сердечно-сосудистых заболеваний (ССЗ) является актуальной проблемой, связанной с лидирующем уровнем смертности от них в мире, различными способами оценки сердечно-сосудистого риска, точностью его определения.*

Цель: *разработать модель при помощи машинного обучения для предсказания сердечно-сосудистого риска и валидировать ее с использованием российских медицинских данных.*

Материал и методы: *набор данных для обучения получен из Фрамингемского исследования, в него входили 4363 пациента без ССЗ, из которых 852 (19,5%) умерли от инфаркта миокарда и инсульта в течение 10 лет с начала наблюдения. Входящие признаки модели: пол, возраст, систолическое АД, холестерин, курение, индекс массы тела, частота сердечных сокращений. Исходный набор данных был разделен на 2 части: учебный набор данных «train» (80% записей) и набор данных проверки «validate» (оставшиеся 20%). Дополнительно было проведено тестирование модели на внешнем наборе данных «test», который включал 411 деперсонифицированных данных пациентов российской популяции.*

Результаты: *итогом работы стала модель WML.CVD.Score, построенная методом последовательной нейронной сети с одним входным, двумя скрытыми и одним выходным слоем. Результаты точности на учебном наборе данных: Accuracy – 81,15%, AUC – 0,80. Эти же показатели на проверочном наборе данных «validate» составили: Accuracy – 81,1%, AUC 0,76. Результаты тестирования на наборе данных «test»: Accuracy – 79,07, AUC – 0,86. На российских тестовых данных AUC для шкалы SCORE составила 0,81 против 0,86 для разработанной модели, что показало обоснованность применения машинного обучения с целью повышения прогностической модели.*

Заключение: *разработанная модель продемонстрировала высокую точность предсказания сердечно-сосудистых событий как при внутренней, так и при внешней валидации.*

Ключевые слова: кардиология, сердечно-сосудистые заболевания, факторы риска, моделирование риска, машинное обучение.

Для цитирования: 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): 41–46. https://doi.org/10.29296/25877305-2020-05-08

Сердечно-сосудистые заболевания (ССЗ) занимают лидирующую позицию среди причин смерти во всем мире, обусловливая на протяжении ряда последних лет более 17,5 млн смертей в год [1].

В Российской Федерации (РФ) показатели смертности от ССЗ занимают 1-е место, обусловливая более 840 тыс. случаев (2018) и значительно доминируют над другими причинами [2]. В последние годы в РФ показатель смертности от ССЗ снизился, что объясняется во многом эффективной реорганизацией медицинской помощи больным с острым коронарным синдромом, кроме того, совершенствуется система диспансерного наблюдения пациентов с ССЗ амбулаторно, более широко внедряется система первичной профилактики. Но необходимо дальнейшее снижение показателя смертности до значений, обозначенных в национальном проекте «Здравоохранение» [3]. Для достижения поставленных целей необходима глубокая трансформация существующих моделей оказания кардиологической помощи с созданием эффективной системы управления сердечно-сосудистыми рисками (ССР) [4].

Одной из возможностей управления ССР является разработка систем прогнозирования развития и течения ССЗ. Сегодня практикующий терапевт и кардиолог применяют в своей работе различные принципы прогнозирования (оценка риска ССЗ по шкалам, стратификация риска при конкретной патологии), но в реальной практике часто это представляет определенные трудности [5].

В настоящее время пройден значительный путь от понимания отдельных факторов риска (ФР) до их сложного патофизиологического воздействия на развитие ССЗ, выразившегося в форме разработок общепризнанных прогностических шкал. Однако существующие шкалы прогноза ССЗ имеют ряд значимых ограничений, влияющих на предсказательную способность, с возможностью как усугубления ССР, так и его недооценки [6]. Недостатки существующих прогностических шкал может нивелировать новый подход к прогнозированию ССР – это применение методов машинного обучения, которые часто объединяют в собирательное понятие «искусственный интеллект» [7–9]. Эти методы обладают рядом доказанных преимуществ: выявляют неочевидные и скрытые закономерности между ФР и исходами, обладают быстротой анализа, не сопоставимой с традиционными методиками, отсутствием необходимости длительного проспективного анализа медицинских данных. Построение прогностических шкал при помощи машинного обучения потенциально способно улучшить точность прогноза развития ССЗ [10].

Целью данного исследования является разработка модели определения ССР для лиц без ССЗ.

Задачами исследования явились построение и валидация на внешних медицинских данных модели прогноза индивидуальной вероятности смерти от ИБС и инсульта для лиц без ССЗ, разработанной при помощи методов машинного обучения, а также сравнение полученной модели с имеющейся Европейской шкалой для расчета риска смерти от сердечно-сосудистого заболевания в ближайшие 10 лет (SCORE).

МАТЕРИАЛ И МЕТОДЫ

В качестве исходного набора данных для обучения модели были использованы данные Фрамингемского исследования (Framingam Heart Study, США) [11], состоящего из 4363 пациентов без ССЗ на момент обследования, из которых 852 (19,5% от когорты) умерли от инфаркта миокарда и инсульта в течение 10 лет с момента начала наблюдения.

Для формирования модели были взяты признаки, используемые в самой распространенной прогностической шкале в Европе SCORE, наиболее популярной в РФ. Помимо этого, в разработку модели дополнительно были включены 2 ФР: индекс массы тела (ИМТ) и частота сердечных сокращений (ЧСС). В настоящее время данные ФР развития ССЗ рассматриваются как значимые в определении ССР, они легковоспроизводимы, что немаловажно для практического применения. Таким образом, клинические признаки пациентов, используемые для оценки ССР в данной модели, – это те признаки, которые определены клиническими исследованиями Европейского общества кардиологов по проекту SCORE, а также другими исследованиями по изучению факторов, влияющих на смертельные исходы от ССЗ [12, 13]: пол, возраст, систолическое АД, общий холестерин, курение/отсутствие курения, ИМТ, ЧСС.

Перечень и характеристики признаков для построения модели и диапазон их значений приведены в табл. 1.

Таблица 1

Характеристика клинических признаков, используемых для построения модели

Table 1

Characterization of the clinical features used to build the model

Наименование признака	Возраст, годы	Пол, [1, 0]	Курение, [1, 0]	ОХ, г/моль	САД, мм рт. ст.	ИМТ, кг/м²	ЧСС в минуту
Условное наименование признака	AGE	SEX	SMOKE	TOTCHOL	SYSBP	BMI	HEART
Среднее (mean)	49,9	0,44	0,49	6,16	132,8	25,8	75,86
Стандартное отклонение (std)	8,65	0,50	0,50	1,16	22,3	4,10	12,02
Минимальное значение (Min)	32	0	0	2,935	83,5	15,54	44
Максимальное значение (Max)	70	1	1	18,08	295	56,8	143

Примечание. ОХ – общий холестерин, САД – систолическое АД.

Возраст пациентов в когорте – от 32 до 70 лет (средний возраст – 49,9 года, стандартное отклонение – 8,65). Минимальное и максимальное значения возраста в наборе данных являются ограничивающими факторами модели. Признаки «пол» (SEX), «курение» (SMOKE) для повышения точности были разложены на два бинарных. Поэтому в модели по факту используются не 7, а 9 признаков. В итоге когорта располагала полными данными, необходимыми для обучения модели.

Исходная дата набора данных была установлена 1 января 2005 г., что позволило всем пациентам когорты находиться под наблюдением в течение 10 лет, дата конца периода наблюдения была определена как 1 января 2015 г. Лица, имеющие в анамнезе ССЗ, наследственные нарушения липидного обмена, или вне возрастного диапазона были исключены из анализа. Под сердечно-сосудистой смертью понималась смерть от ИБС и инсульта.

Исходный набор данных был разделен на две части. Первая – это учебный набор данных «train» (80% от числа записей исходного набора), предназначенный для обучения модели. Второй – это набор данных для проверки «validate» (оставшиеся 20%), он применялся для внутреннего тестирования с целью оценки алгоритмов согласно апробированным подходам использования машинного обучения при прогнозировании ССР [14].

С учетом разработки модели не на российской популяции для дополнительной проверки точности модели была выполнена внешняя валидация на тестовом наборе данных «test». С этой целью была составлена база деперсонифицированных медицинских данных из 411 пациентов российской популяции, которые модель ранее «не видела». Характеристика выборки: средний возраст – 47,9±5,7 года, мужчин – 242 (58,9%), женщин – 169 (41,1%), курильщики – 155 (37,7%), ОХ – 5,16±0,75 ммоль/л, САД – 128,39±14,8 мм рт. ст., ИМТ – 27,5±4,4 кг/м^2, ЧСС – 57,0±21,7 в минуту, период наблюдения – 10 лет или до появления сердечно-сосудистого события. Смерть от сердечно-сосудистого события в течение 10 лет регистрировалась у 87 (21,1%) человек. Полученные характеристики базы данных признаны соответствующими общим популяционным характеристикам и задаче, которую должна была решать данная модель.

Схема построения, валидации и тестирования модели представлена на рис. 1.

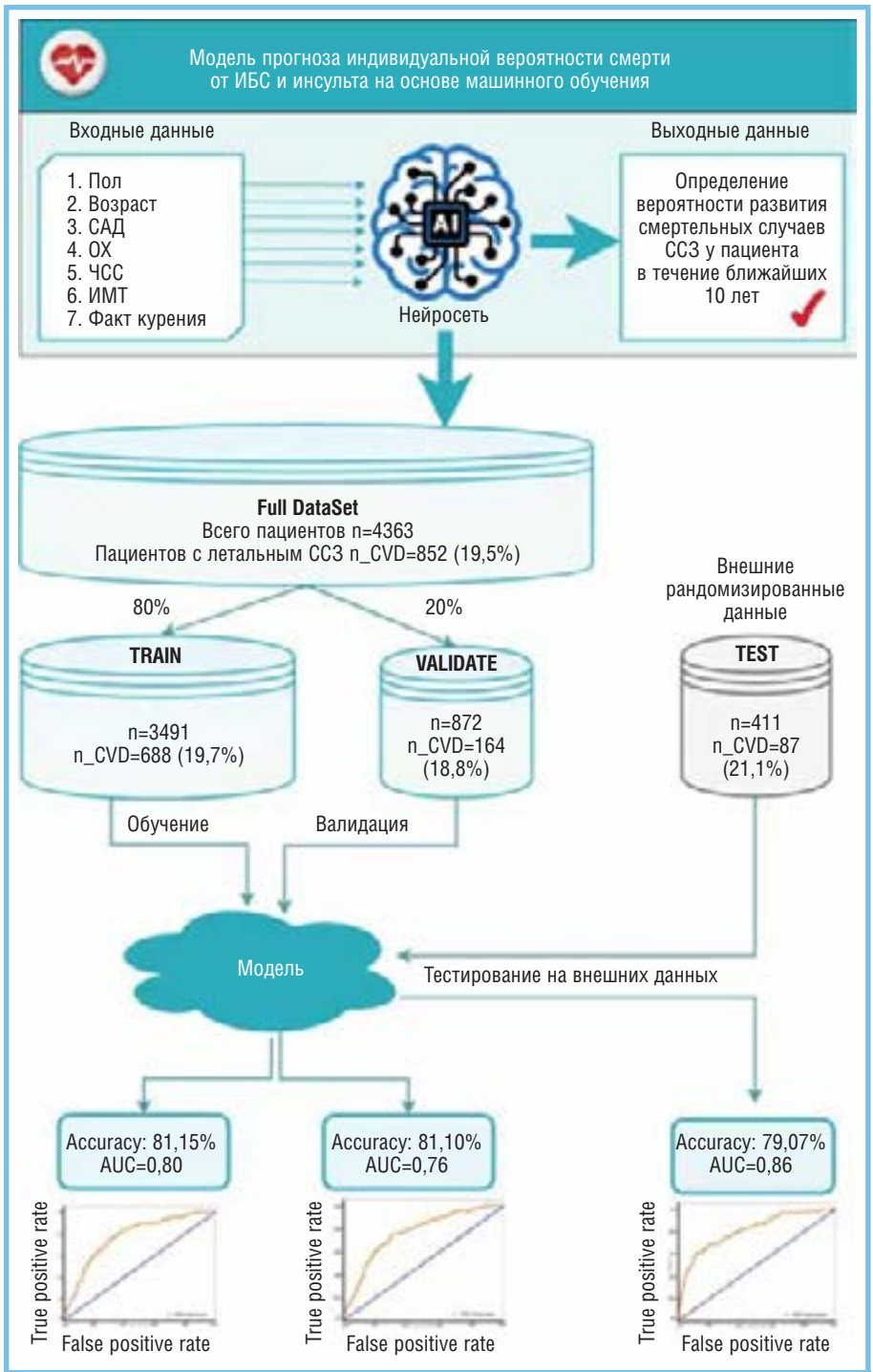

Рис. 1. Концептуальная схема разработки и валидации модели прогнозирования индивидуальной вероятности смерти от ИБС и инсульта WML.CVD.Score

Fig. 1. A conceptual framework for the development and validation of the WML.CVD.Score model for predicting an individual probability of death from coronary heart disease and stroke

Расчет показателей точности ROC-анализа осуществлялся при помощи стороннего программного обеспечения медицинской статистики MedCalc (ПО MedCalc, адрес: *https://www.medcalc.org*, разработчик MedCalc Software Ltd., Бельгия). Параметры точности модели оценивались по методу ROC-анализа [15], основная концепция которого сводится к задаче классификации, чтобы относить ранее неизвестные моделируемые случаи ССЗ с фактическими событиями. В результате классификации получаются четыре результата: истинноположительный (true-positive), ложноположительный (false-positive), истинноотрицательный (true-negative), ложноотрицательный (false-negative). На основании указанной классификации для выбранного порога активации (cut-off) модели рассчитываются показатели: TP – количество истинноположительных результатов, FP – количество ложноположительных результатов, TN – количество истинноотрицательных результатов, FN – количество ложноотрицательных результатов. Для оценки диагностической ценности модели использованы следующие метрики: чувствительность Se=TP/(TP+FN), специфичность Sp=TN/(TN+FP), точность Accuracy=(TP+TN)/(TP+FP+FN+TN), площадь под ROC-кривой (AUC) – площадь, ограниченная ROC-кривой и абсциссой, при этом ROC-кривая – график зависимости Se от (1-Sp).

Таблица 2

Количество данных для обучения и результаты обучения модели

Table 2

Amounts of data for learning and the results of its model

Набор данных	Общее количество записей	Число пациентов с летальным исходом, %	Accuracy, %	AUC
«Train»	3491	688 (19,7)	81,15	0,80
«Validate»	872	164 (18,8)	81,10	0,76
«Test»	411	87 (21,1)	79,07	0,86

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результатом машинного обучения с использованием метода нейронной сети была получена модель WML.CVD.Score для прогноза смерти в течение 10 лет от ССЗ.

Для построения модели использована последовательная нейронная сеть с одним входным, двумя скрытыми и одним выходным слоем. Для предотвращения переобучения использовано исключение («dropout»). На каждом слое применена функция «dense» для полного соединения слоев друг с другом. В скрытых слоях используется функция активации «relu». В качестве оптимизатора алгоритма, который изменяет веса и смещения во время обучения, был использован «rmsprop». В качестве функции потерь («loss») использована бинарная кросс-энтропия, в качестве метрики оценки – AUC (Area Under the Curve).

Результаты точности разработанной модели представлены в табл. 2 и рис. 2 и 3.

При проверке модели на наборе данных «validate» при пороге активации модели CutOff=0,22 получены следующие показатели матрицы ROC-анализа: TP=122, TN=496, FP=212, FN=42, Se=0,75, Sp=0,70, AUC=0,76. На тестовых внешних данных российских пациентов при оптимальном CutOff=0,212: TP=51, TN=304, FP=20, FN=36, Se=0,718, Sp=0,89, AUC=0,86. Полученные результаты демонстрируют высокие значения точности, чувствительности и специфичности модели как на проверочном эталонном наборе американских данных, так и тестовом наборе реальных клинических данных российских пациентов.

Сравнение результатов оценки риска модели WML.CVD.Score с результатами по шкале SCORE

Поскольку созданная модель отвечает на тот же вопрос, что и шкала SCORE – прогноз смерти от инфаркта

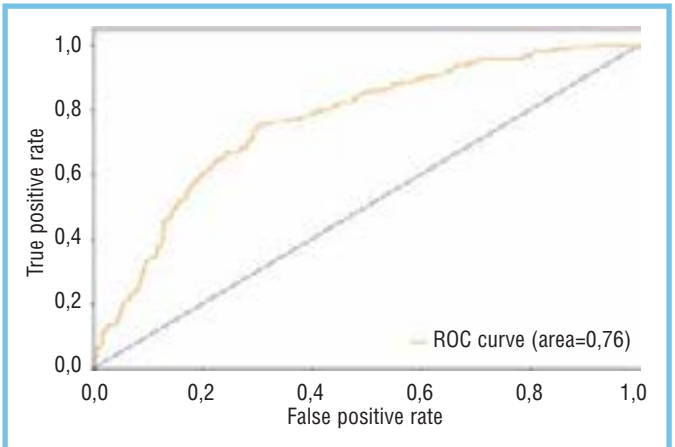

Рис. 2. ROC-кривая модели на наборе данных «validate»
Fig. 2. ROC curve of the model on the «validate» dataset

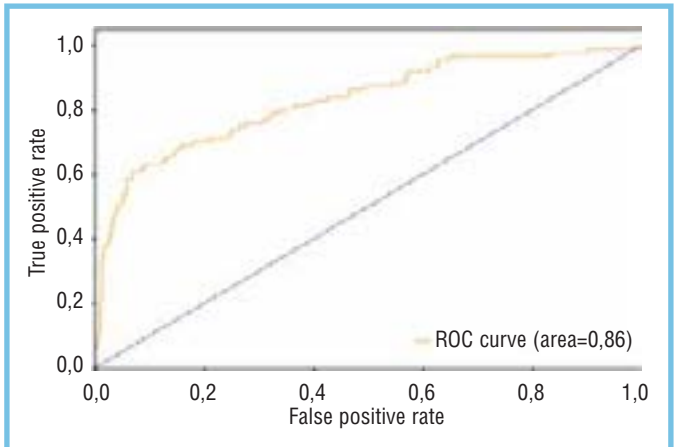

Рис. 3. ROC-кривая модели на наборе данных «test»
Fig. 3. ROC curve of the model on the «test» dataset

миокарда и инсульта, то было проведено сравнение точности результатов оценки рисков, полученных по данным разработанной модели и по данным шкалы SCORE.

Тестовый набор данных «test» из 411 пациентов был подан на вход калькулятора-шкалы SCORE. Из них для 28 пациентов шкала оказалась неприменимой в силу ограничения по возрасту, и тестовый набор был сокращен до 383 пациентов, из которых у 65 (17%) наступил летальный исход. По шкале SCORE были получены результаты оценки риска в баллах от 1 до 13, при этом для значений свыше 5 баллов присваивается умеренный риск смерти от ССЗ.

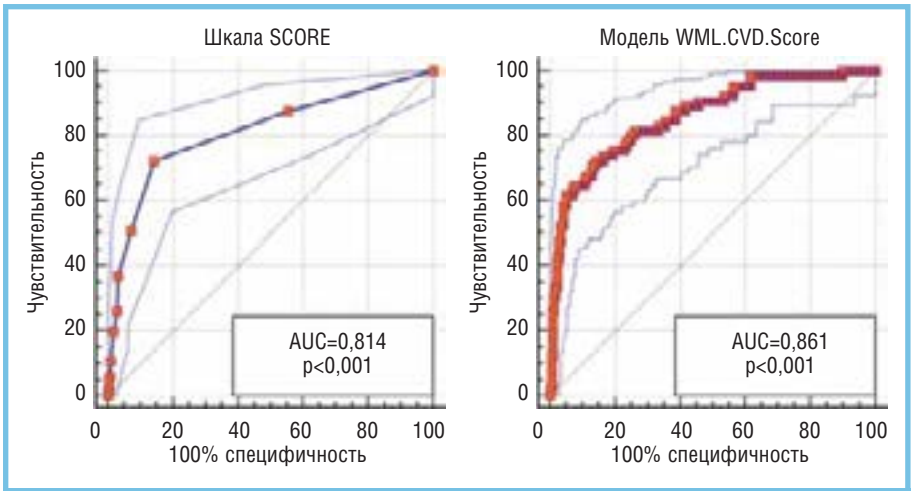

Рис. 4. ROC-кривые для шкалы SCORE и модели WML.CVD.Score на тестовом наборе данных российских пациентов (n=383)
Fig. 4. ROC curves for the SCORE scale and the WML.CVD.Score model on the test dataset on Russian patients (n=383)

Полученные результаты оценки по шкале SCORE были оценены на основе метрик ROC-анализа. Для тестового набора данных при оптимальном CutOff=5 баллов были получены следующие метрики модели: TP=17 (против 50 для модели), TN=309, FP=48 (против модельных 19), FN=9, Se=0,65, Sp=0,86. AUC для шкалы SCORE составил 0,81 против 0,86 для модели на тех же тестовых данных, представленных на рис. 4.

Таким образом, проведенное сравнение показало, что созданная модель и шкала SCORE обладают высоким качеством AUC, при этом показатель у модели WML.CVD.Score лучше на 5%, что для профилактики заболеваемости является существенным улучшением. Модель имеет более высокую чувствительность Se=75% и предсказывает правильнее сердечно-сосудистые события – TP=50 против TP=17 для шкалы SCORE.

Разработанная при помощи машинного обучения модель прогнозирования индивидуальной вероятности смерти от ИБС и инсульта для лиц без ССЗ демонстрирует высокие показатели точности предсказания сердечно-сосудистых событий как на валидационном наборе данных Фрамингемского исследования, так и наборе российских данных пациентов из электронных медицинских карт.

По сравнению со шкалой SCORE разработанная с помощью машинного обучения модель WML.CVD.Score обладает более высокой чувствительностью и правильнее предсказывает положительные сердечно-сосудистые события. Таким образом, алгоритмы машинного обучения демонстрируют лучшую предсказательную способность при расчете ССР.

Встраивание таких моделей в системы поддержки принятия врачебных решений с автоматическим анализом электронных медицинских карт будут способствовать лучшему управлению ССР. Подходы машинного обучения открывают перспективу достижения улучшенной и более индивидуализированной оценки риска ССЗ. Это может помочь движению к персонализированной медицине, лучшей адаптации управления рисками к отдельным пациентам.

* * *

Конфликт интересов: *все авторы заявляют об отсутствии конфликта интересов, требующего раскрытия в данной статье.*

Литература/Reference

1. WHO Global Action Plan for the Prevention and Control of Non-communicable Diseases 2013–2020 (resolution WHA66.10, 27 May 2013) Available at: *http://apps.who.int/gb/ebwha/pdf_files/WHA66/A66_R10-en.pdf?ua=1* [Accessed 27 Mar. 2020].
2. Федеральная служба государственной статистики [Federal State Statistic Service (in Russ.)]. Available at: *https://gks.ru/folder/13721* [Accessed 27 Mar. 2020].
3. Паспорт национального проекта «Здравоохранение» (утв. Президиумом Совета при Президенте РФ по стратегическому развитию и национальным проектам, протокол от 24.12.2018 №6) [Passport of the national project «Healthcare» (in Russ.)] Available at: *http://www.consultant.ru* [Accessed 27 Mar. 2020].
4. Шляхто Е.В., Звартау Н.Э., Виллевальде С.В. и др. Система управления сердечно-сосудистыми рисками: предпосылки к созданию, принципы организации, таргетные группы. *Рос. кардиол. журн.* 2019; 24 (11): 69–82 [Shlyakhto E.V., Zvartau N.E., Villevalde S.V. et al. Cardiovascular risk management system: prerequisites for developing, organization principles, target groups. *Russian Journal of Cardiology*. 2019; 24 (11): 69–82 (in Russ.)]. DOI: 10.15829/1560-4071-2019-11-69-82
5. Белялов Ф.И. Шкалы прогноза сердечно-сосудистых заболеваний. *Архив внутренней медицины*. 2015; 5: 19–21 [Belyalov F.I. Prognostic scores for cardiovascular diseases. *The Russian Archives of Internal Medicine*. 2015; 5: 19–21 (in Russ.)].
6. Бойцов С.А., Шальнова С.А., Деев А.Д. и др. Моделирование риска развития сердечно-сосудистых заболеваний и их осложнений на индивидуальном и групповом уровнях. *Тер. арх.* 2013; 85 (9): 4–10 [Boitsov S.A., Shalnova S.A., Deev A.D. et al. Simulation of a risk for cardiovascular diseases and their events at individual and group levels. *Therapeutic archive*. 2013; 85 (9): 4–10 (in Russ.)].
7. Weng S.F., 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

8. Angraal S., Mortazavi B.J., Gupta A. et al. Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction. *JACC: Heart Failure*. 2020; 8 (1): 12–21. https://doi.org/10.1016/j.jchf.2019.06.013

9. Meyer A., Zverinski D., Pfahringer B. et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir. Med*. 2018; 6 (12): 905–14. https://doi.org/10.1016/S2213-2600(18)30300-X

10. Kuznetsova T., Novitskiy R., Gusev A. et al. Deep and machine learning models to improve risk prediction of cardiovascular disease using data extraction from electronic health records. *Eur. Heart J*. 2019; 40 (Suppl. 1): 1923–4. https://doi.org/10.1093/eurheartj/ehz748.0670

11. Clinical Practice Research Datalink, reference number: CPRD00039761. Available at: *https://www.cprd.com*

12. European guidelines on cardiovascular disease prevention in clinical practice: third joint task force of European and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of eight societies and by invited experts). *Eur. J. Cardiovasc. Prev. Rehabil*. 2003; 10 (4): 1–10. DOI: 10.1097/01.hjr.0000087913.96265.e2

13. Conroy R.M., Pyorala K., Fitzgerald A.P. et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur. Heart J*. 2003; 24: 987–1003. DOI: 10.1016/s0195-668x(03)00114-3

14. Beunzaa J.-J., Puertasa E. et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *J. Biomed. Inform*. 2019; 97: 103257. DOI:10.1016/j.jbi.2019.103257

15. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27 (8): 861–74. https://doi.org/10.1016/j.patrec.2005.10.010

CARDIOVASCULAR DISEASES PREDICTION BY INTEGRATED RISK FACTORS ASSESSMENT BY MEANS OF MACHINE LEARNING

***D. Gavrilov*[1]; *L. Serova*[1]**, *Candidate of Engineering Sciences;* ***I. Korsakov*[1]**, *Candidate of Physico-Mathematical Sciences;* ***A. Gusev*[1]**, *Candidate of Engineering Sciences;* ***R. Novitsky*[1]; *T. Kuznetsova*[2]**, *MD*

[1]*K-SkAI, Petrozavodsk*

[2]*Petrozavodsk State University*

Aim. *To develop a model by machine learning to predict the risk of cardiovascular diseases (CVD) and validate the model using Russian medical data.*

Materials and methods. *The data set was obtained from the Framingham study, consisting of 4,363 patients without CVD, 852 (19.5%) of which died of myocardial infarction and stroke within 10 years of observation. Incoming model features: gender, age, systolic blood pressure, cholesterol, smoking, body mass index, heart rate. The original data set was divided into 2 parts: the training data set (80% of the records) and the validate data set (the remaining 20%). Additionally, the model was evaluated by an external data set included 411 depersonalized patient data from the Russian citizens.*

Results. *The WML.CVD.Score model was created by the serial neural network with one input, two hidden and one output layer. Accuracy results on a training dataset: Accuracy 81.15%, AUC 0.80. The same indicators on the validate data set were: Accuracy 81.1%, AUC 0.76. Test results for the test data set: Accuracy 79.07, AUC 0.86. On the Russian test data, the AUC for the SCORE scale was 0.81 versus 0.86 for the developed model, which showed the validity of the use of machine learning in order to increase the predictive model.*

Conclusion. *The developed model has demonstrated high accuracy to CVD predicting in both internal and external validation. The model can be used in medical practice for patients in Russia.*

Key words: *cardiology, cardiovascular diseases, risk factors, risk modeling, machine learning.*

For citation: *Gavrilov D., Serova L., Korsakov I. et al. Cardiovascular diseases prediction by integrated risk factors assessment by means of machine learning. Vrach. 2020; 31 (5): 41–46. https://doi.org/10.29296/25877305-2020-05-08*

Об авторах / About the Authors: *Gavrilov D.V. ORCID: 0000-0002-8745-857X, Serova L.M. ORCID: 0000-0001-6259-2492, Korsakov I.N. ORCID: 0000-0003-2343-9641, Gysev A.V. ORCID: 0000-0002-7380-8460, Novitsky R.E. ORCID: 0000-0002-2350-977X, Kuznetsova T.Yu. ORCID: 0000-0002-6654-1382*