Кардиоваскулярная терапия и профилактика. 2022;21(1):2908.
doi:10.15829/1728-8800-2022-2908
ISSN 1728-8800 (Print)
ISSN 2619-0125 (Online)

Методы машинного обучения в прогнозировании рисков 5-летней смертности (по данным исследования ЭССЕ-РФ в Приморском крае)

Невзорова В. А.[1], Бродская Т. А.[1], Шахгельдян К. И.[2,3], Гельцер Б. И.[2], Костерин В. В.[3], Присеко Л. Г.[1]

[1]ФГБОУ ВО Тихоокеанский государственный медицинский университет Минздрава России, Институт терапии и инструментальной диагностики. Владивосток; [2]ФАОУ ВО Дальневосточный федеральный университет, Школа биомедицины. Владивосток; [3]ФГБОУ ВО Владивостокский государственный университет экономики и сервиса, Институт информационных технологий. Владивосток, Россия

Цель. Разработка и сравнительная оценка точности моделей прогнозирования риска смерти в течение 5 лет по данным исследования ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации) в Приморском крае.

Материал и методы. В исследование включено 2131 человек (1257 женщин и 874 мужчины) в возрасте 23-67 лет с медианой 47 лет, 95% доверительный интервал [46; 48]. Протокол исследования включал: измерение артериального давления (АД), частоты сердечных сокращений (ЧСС), окружности талии, бедер и их соотношение (СТБ). Биохимические показатели крови: общий холестерин (ХС), ХС липопротеинов низкой и высокой плотности, триглицериды, аполипопротеины AI и B, липопротеин(а), N-концевой фрагмент мозгового натрийуретического пропептида (NT-proNBP), D-димер, фибриноген, С-реактивный белок (СРБ), глюкоза, креатинин, мочевая кислота. Конечной точкой исследования был факт смерти от всех причин в течение 5 лет проспективного наблюдения (2013-2018гг). Группу умерших за этот период составили 42 (2%) человека, продолживших исследование — 2089 (98%). Для обработки и анализа данных применяли тесты χ^2, Фишера и Манна-Уитни, однофакторную логистическую регрессию (ЛР). Для построения прогностических моделей использовали методы машинного обучения (МО): многофакторную ЛР, вейбул-регрессию и стохастический градиентный бустинг.

Результаты. Разработанные на основе методов МО прогностические модели с использованием в их структуре показателей возраста, пола, факта курения, систолического АД (САД) и уровня общего ХС имели более высокие метрики качества, чем шкала SCORE (Systematic Coronary Risk Evaluation). Включение в состав предикторов показателей СРБ, глюкозы, NT-proNBP и ЧСС повышало точность всех моделей с максимальным подъемом метрик качества в модели многофакторной ЛР. Тестирование предиктивного потенциала других факторов (СТБ, показатели липидного спектра, фибриноген, D-димер и др.) не улучшало качество прогнозирования. Анализ степени влияния отдельных предикторов на показатель смертности указывал на превалирующий вклад 5 факторов: возраста, уровней общего ХС, NT-proNBP, СРБ и глюкозы. Менее заметное влияние ассоциировалось с уровнем ЧСС, САД и курения, а вклад гендерной принадлежности был минимальный.

Заключение. Применение современных методов МО повышает надежность прогностических моделей и обеспечивает более высокую эффективность риск-стратификации обследованных, особенно среди лиц с низким и умеренным риском смерти от болезней системы кровообращения.

Ключевые слова: методы машинного обучения, прогнозирование, факторы риска, смертность, ЭССЕ-РФ.

Отношения и деятельность. Работа выполнена при поддержке грантов РФФИ 19-29-01077 и 18-29-03131.

Поступила 13/05-2021
Рецензия получена 30/06-2021
Принята к публикации 28/07-2021

Для цитирования: Невзорова В. А., Бродская Т. А., Шахгельдян К. И., Гельцер Б. И., Костерин В. В., Присеко Л. Г. Методы машинного обучения в прогнозировании рисков 5-летней смертности (по данным исследования ЭССЕ-РФ в Приморском крае). *Кардиоваскулярная терапия и профилактика.* 2022;21(1):2908. doi:10.15829/1728-8800-2022-2908

Machine learning for predicting 5-year mortality risks: data from the ESSE-RF study in Primorsky Krai

Nevzorova V. A.[1], Brodskaya T. A.[1], Shakhgeldyan K. I.[2,3], Geltser B. I.[2], Kosterin V. V.[3], Priseko L. G.[1]

[1]Pacific State Medical University, Institute of Therapy and Instrumental Diagnostics. Vladivostok; [2]Far Eastern Federal University, School of Biomedicine. Vladivostok; [3]Vladivostok State University of Economics and Service, Institute of Information Technologies. Vladivostok, Russia

Aim. To develop and perform comparative assessment of the accuracy of models for predicting 5-year mortality risks according to the Epidemiology of Cardiovascular Diseases and their Risk Factors in Regions of Russian Federation (ESSE-RF) study in Primorsky Krai.

*Автор, ответственный за переписку (Corresponding author):
e-mail: brodskaya@mail.ru
Тел.: +7 (914) 651-71-00

[Невзорова В. А. — д.м.н., профессор, директор института терапии и инструментальной диагностики, ORCID: 0000-0002-0117-0349, Бродская Т. А.* — д.м.н., доцент, профессор института терапии и инструментальной диагностики, ORCID: 0000-0002-9836-6339, Шахгельдян К. И. — д.т.н., зав. лабораторией "Анализ больших данных в здравоохранении и биомедицины" Школы биомедицины, директор института информационных технологий, ORCID: 0000-0002-4539-685X, Гельцер Б. И. — д.м.н., профессор, член-корр. РАН, директор департамента клинической медицины Школы биомедицины, ORCID: 0000-0002-9250-557X, Костерин В. В. — стажёр-исследователь Лаборатория цифрового моделирования и анализа данных физики и биомедицины, ORCID: 0000-0003-3747-7438, Присеко Л. Г. — клинический ординатор Института терапии и инструментальной диагностики, ORCID: 0000-0002-3946-2064].

Material and methods. The study included 2131 people (1257 women and 874 men) aged 23-67 years with a median of 47 years (95% confidence interval [46; 48]). The study protocol included measurement of blood pressure (BP), heart rate (HR), waist circumference, hip circumference, and waist-to-hip ratio (WHR). The following blood biochemical parameters: total cholesterol (TC), low and high density lipoprotein cholesterol, triglycerides, apolipoproteins AI and B, lipoprotein(a), N-terminal pro-brain natriuretic peptide (NT-proNBP), D-dimer, fibrinogen, C-reactive protein (CRP), glucose, creatinine, uric acid. The study endpoint was 5-year all-cause death (2013-2018). The group of deceased patients during this period consisted of 42 (2%) people, while those continued the study — 2089 (98%). The χ^2, Fisher and Mann-Whitney tests, univariate logistic regression (LR) were used for data processing and analysis. To build predictive models, we used following machine learning (ML) methods: multivariate LR, Weibull regression, and stochastic gradient boosting.

Results. The prognostic models developed on the ML basis, using parameters of age, sex, smoking, systolic blood pressure (SBP) and TC level in their structure, had higher quality metrics than Systematic COronary Risk Evaluation (SCORE) system. The inclusion of CRP, glucose, NT-proNBP, and heart rate into the predictors increased the accuracy of all models with the maximum rise in quality metrics in the multivariate LR model. Predictive potential of other factors (WHR, lipid profile, fibrinogen, D-dimer, etc.) was low and did not improve the prediction quality. An analysis of the influence degree of individual predictors on the mortality rate indicated the prevailing contribution of five factors as follows: age, levels of TC, NT-proNBP, CRP, and glucose. A less noticeable effect was associated with the level of HR, SBP and smoking, while the contribution of sex was minimal.

Conclusion. The use of modern ML methods increases the accuracy of predictive models and provides a higher efficiency of risk stratification, especially among individuals with a low and moderate death risk from cardiovascular diseases.

Keywords: machine learning methods, prediction, risk factors, mortality, ESSE-RF.

Relationships and Activities. This work was supported by Russian Foundation for Basic Research grants 19-29-01077 and 18-29-03131.

Nevzorova V. A. ORCID: 0000-0002-0117-0349, Brodskaya T. A.* ORCID: 0000-0002-9836-6339, Shakhgeldyan K. I. ORCID: 0000-0002-4539-685X, Geltser B. I. ORCID: 0000-0002-9250-557X, Kosterin V. V. ORCID: 0000-0003-3747-7438, Priseko L. G. ORCID: 0000-0002-3946-2064.

*Corresponding author: brodskaya@mail.ru

Received: 13/05-2021
Revision Received: 30/06-2021
Accepted: 28/07-2021

For citation: Nevzorova V. A., Brodskaya T. A., Shakhgeldyan K. I., Geltser B. I., Kosterin V. V., Priseko L. G. Machine learning for predicting 5-year mortality risks: data from the ESSE-RF study in Primorsky Krai. *Cardiovascular Therapy and Prevention*. 2022;21(1):2908. (In Russ.) doi:10.15829/1728-8800-2022-2908

АГ — артериальная гипертензия, апо — аполипопротеин, БСК — болезни системы кровообращения, ВР — вейбулл-регрессия, ДИ — доверительный интервал, ИА — индекс атерогенности, ИК — индекс курения, ИМТ — индекс массы тела, ЛВП — липопротеины высокой плотности, ЛНП — липопротеины низкой плотности, Лп(а) — липопротеин(а), ЛР — логистическая регрессия, Ме — медиана, МК — мочевая кислота, МЛР — многофакторная ЛР, МО — машинное обучение, ОБ — окружность бедер, ОТ — окружность талии, ОШ — отношение шансов, ПАД — пульсовое артериальное давление, САД — систолическое артериальное давление, СГБ — стохастический градиентный бустинг, СРБ — С-реактивный белок, ССР — сердечно-сосудистый риск, СТБ — соотношение ОТ/ОБ, ТГ — триглицериды, у.е. — условные единицы, Фг — фибриноген, ХС — холестерин, ЧСС — частота сердечных сокращений, ЭССЕ-РФ — Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации, ACC — точность анализа, AUC — площадь под ROC-кривой, NT-proBNP — N-терминальный фрагмент мозгового натрийуретического пропептида, SCORE — Systematic Coronary Risk Evaluation (шкала оценки риска смерти), Sen — чувствительность, Spec — специфичность.

Введение

Структура смертности населения с сохранением лидерства болезней системы кровообращения (БСК) остается неизменной в большинстве стран современного мира даже в период пандемии новой коронавирусной инфекции COVID-19 (COronaVIrus Disease 2019) [1]. Доказано, что увеличение продолжительности социально и экономически активного возраста возможно только на основе широкого внедрения в систему здравоохранения профилактических стратегий популяционного и высокого сердечно-сосудистого риска (ССР), реализация которых позволит добиться снижения смертности от БСК [2, 3]. Для формирования эффективной популяционной стратегии разработаны и продолжают совершенствоваться прогнозные шкалы оценки вероятности сердечно-сосудистых событий, основанные на проведении крупномасштабных исследований, примером которых являются Framingham Risk Score, SCORE (Systematic Coronary Risk Evaluation) и др. [2, 4]. Вместе с тем, по мнению экспертного сообщества, традиционные шкалы не всегда точны в отношении прогноза в одной из самых сложных для профилактических мероприятий категории лиц с умеренным или низким риском смерти от БСК. В связи с этим в последние годы осуществляется поиск новых подходов, позволяющих персонифицировать риски неблагоприятных событий за счет модификации классических прогностических шкал. Так, на основе Framingham Heart Study была создана шкала ASSIGN (Assessing Cardiovascular Risk to Scottish Intercollegiate Guidelines Network/SIGN to Assign Preventative Treatment), учитывающая национальные особенности вклада отдельных факторов ССР в развитие БСК [5]. Иным примером расширения возможностей используемых шкал является введение поправочных коэффициентов для оценки традиционных факторов риска и их совокупного влияния на вероятность развития индексных сердечно-сосудистых событий [3]. Однако ни один из таких подходов до настоящего времени не привел к принципиальным изменениям, позволяющим индивидуализировать ССР [4]. В качестве базовой модели верификации степени 10-летнего суммарного ССР экспертами признается европейская система

SCORE, разработанная с использованием данных российских исследований [2, 3]. Вместе с тем, прогностическая точность этой системы составляет около 65%, что требует ее совершенствования, в т.ч. на основе проспективных исследований в российской популяции [6]. Примером такого подхода служит исследование ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации) с проспективным этапом наблюдения, продолжающимся уже в течение 8 лет в 12 регионах РФ, включая Приморский край. В рутинной практике обработка медицинских данных для оценки популяционного риска осуществляется с помощью методов математической статистики. Вместе с тем, в последние годы в превентивной и клинической кардиологии для интеллектуального анализа больших данных все шире используются современные технологии машинного обучения (МО), с помощью которых разрабатываются модели, обеспечивающие более высокую точность предсказания сердечно-сосудистых событий [7-9].

Цель исследования состояла в разработке и сравнительном анализе точности моделей прогнозирования риска смерти в течение 5 лет по результатам исследования ЭССЕ-РФ в Приморском крае.

Материал и методы

Исследование выполнено на основе данных многоцентрового эпидемиологического исследования ЭССЕ-РФ, проведенного на территории Приморского края (2013-2015гг). Подробный протокол исследования ЭССЕ-РФ был представлен ранее [10]. Путем систематической стратифицированной многоступенчатой случайной выборки по методу Киша были отобраны и обследованы 2131 человек (1257 женщин и 874 мужчины), жителей Приморского края, включая городское (n=1655) и сельское (n=462) население, в возрасте 23-67 лет с медианой (Me) 47 лет и 95% доверительным интервалом (ДИ): 46-48.

Критериями невключения были: острые и хронические заболевания в фазе обострения, ожирение, симптоматическая артериальная гипертензия (АГ). Дизайн исследования одобрен междисциплинарным этическим комитетом ФГБОУ ВО ТГМУ Минздрава России. Все пациенты дали письменное информированное согласие на участие в исследовании. Протокол включал 61 характеристику, полученную по результатам антропометрических, общеклинических, лабораторных и инструментальных обследований. Выполнено измерение роста, массы тела, окружности талии (ОТ) и бедер (ОБ), расчет соотношения ОТ/ОБ (СТБ) и индекса массы тела (ИМТ). В исследование включена информация о гендерной принадлежности, статусе курения, наличии АГ, семейном анамнезе. Использовался стандартный вопросник, разработанный на основе адаптированных международных методик, который включает 12 модулей. Всем пациентам проводилось взятие венозной крови натощак и измерение артериального давления. Биохимические исследования выполнены на анализаторе AbbotArchitect c8000 (США) с использованием диагностических наборов этой же фирмы. Определяли уровни общего холестерина (ХС), ХС липопротеинов низкой плотности (ЛНП), ХС липопротеинов высокой плотности (ЛВП), триглицеридов (ТГ), аполипопротеинов (апо) В и А, липопротеина(а) (Лп(а)), N-концевого фрагмента мозгового натрийуретического пропептида (NT-proNBP), глюкозы, инсулина, мочевой кислоты (МК), фибриногена (Фг), D-димера, С-реактивного белка (СРБ), креатинина. Рассчитывали индекс атерогенности (ИА) по формуле: ИА=(общий ХС–ХС ЛВП)/ХС ЛВП. В группу активных курильщиков включали лиц, которые ежедневно выкуривают хотя бы одну сигарету. К бросившим курить относились лица, курившие в прошлом и не употребляющие табачные изделия на протяжении ⩾ 6 мес. Индекс курения (ИК) в пачка-лет рассчитывали по формуле: количество выкуренных сигарет/сут., умноженное на общий стаж курения (в годах), деленное на 20. Согласно дизайну исследования ЭССЕ-РФ, в отличие от шкалы SCORE, конечной точкой явился факт смерти от всех причин в течение 5 лет проспективного наблюдения (2013-2020гг). Анализ смертности среди обследованных проводился по данным Единой информационной системы здравоохранения Приморского края. Группу умерших за этот период составили 42 (2%) человека, из которых мужчин было 23, женщин — 19, а группу продолжающих исследование — 2089 (98%) человек (851 мужчин и 1238 женщин).

Входные признаки (потенциальные предикторы) были представлены в форме непрерывных или категориальных переменных. Для обработки и анализа данных использовали методы статистического анализа: Me и 95% ДИ, тесты хи-квадрат (χ^2), Фишера, Манна-Уитни, корреляционный анализ (по Пирсону и Спирмену), однофакторную логистическую регрессию (ЛР). Методы МО были представлены многофакторной ЛР (МЛР), вейбулл-регрессией (ВР), стохастическим градиентным бустингом (СГБ). Статистическая достоверность признаков и проверка гипотез подтверждалась значениями $p<0,05$. Разработка МЛР выполнялась с использованием в их структуре только одной из коррелируемых переменных для исключения проблем мультиколлинеарности. Качество моделей оценивали по 4 метрикам: площадь под ROC-кривой (AUC), чувствительность (Sen), специфичность (Spec) и точность (ACC). Модели были разработаны на обучающей выборке (3/4 пациентов) и верифицированы на тестовой (1/4). Процедура кросс-валидации выполнялась с усреднением метрик качества не <100 раз по случайно выбранным данным.

Дизайн исследования включал 4 этапа. На первом из них применяли статистический анализ, с помощью которого проводили межгрупповые сравнения исследуемых параметров. Для непрерывных переменных использовали тест Манна-Уитни, а для категориальных — χ^2 и точный тест Фишера с расчетом отношения шансов (ОШ). На 2-ом этапе по нормализованным признакам с помощью однофакторной ЛР относительно конечной точки исследования определяли весовые коэффициенты потенциальных предикторов. На 3-ем этапе были разработаны прогностические модели на основе МЛР, ВР и СГБ. При увеличении метрик качества моделей считали, что включенный в их структуру фактор может рассматриваться в качестве предиктора риска общей 5-летней смерти. На 4-ом этапе определяли относительный вклад

Таблица 1

Сравнительный анализ клинико-демографических
и функционально-метаболических показателей в группах сравнения (Me, 95% ДИ)

Показатели	Первая группа n=2089	Вторая группа n=42	ОШ, 95% ДИ	p
Мужской пол, абс. (%)	851 (40,7%)	23 (54,8%)	1,76 (0,95-3,29)	0,091
Возраст, лет	47 [46; 48]	56,5 [53; 61]	–	<0,0001
Бросившие курить, n (%)	496 (23,7%)	8 (19%)	1,17 (0,46-2,7)	0,92
Курящие, n (%)	448 (21,4%)	18 (42,9%)	2,87 (1,44-5,77)	0,003
ИК, пачка-лет	0 [0; 0] 5,9 [5,5; 6,3]	5 [0; 10] 8,4 [5,7; 11,2]	–	0,02
ОТ, см	89 [88; 89]	95 [91; 99]	–	0,0015
ОБ, см	103 [102; 103]	104 [102; 110]	–	0,083
СТБ, у.е.	0,86 [0,85; 0,86]	0,91 [0,89; 0,94]	–	0,0005
ИМТ, кг/м^2	27 [26,7; 27,3]	28,3 [26,5; 30,9]	–	0,02
САД, мм рт.ст.	131,5 [131; 132,5]	140 [135; 144]	–	0,007
ДАД, мм рт.ст.	79,5 [78,5; 80]	81,75 [72,5; 85,5]		0,71
ЧСС, уд./мин	74 [74; 75]	79 [74; 82]	–	0,039
Общий ХС, ммоль/л	5,53 [5,47; 5,60]	5,45 [4,68; 6,30]	–	0,85
ХС ЛНП, ммоль/л	3,5 [3,45; 3,55]	3,68 [2,91; 4,07]	–	0,86
ХС ЛВП, ммоль/л	1,39 [1,37; 1,41]	1,31 [1,19; 1,43]	–	0,065
Лп(а), ммоль/л	12,7 [11,2; 14,2]	13,6 [8,1; 18,4]	–	0,44
апо AI, г/л	1,83 [1,81; 1,85]	1,81 [1,72; 1,87]	–	0,43
апоВ, г/л	0,82 [0,81; 0,83]	0,85 [0,75; 0,95]	–	0,57
ИА, у.е.	2,9 [2,84; 2,96]	3,03 [2,66; 4,00]	–	0,26
ТГ, ммоль/л	1,14 [1,1; 1,18]	1,05 [0,9; 1,55]	–	0,65
Глюкоза, ммоль/л	5,15 [5,11; 5,18]	5,47 [5,09; 5,99]	–	0,016
Креатинин, мкмоль/л	67,7 [67,1; 68,3]	68,15 [62,2; 72,8]	–	0,77
Клиренс креатинина, мкмоль/л	117,4 [116; 118,8]	115,2 [106; 124,3]		0,62
МК, ммоль/л	320 [310; 330]	325 [290; 380]	–	0,136
D-димер, нг/мл	153 [151; 155]	167,5 [146; 200]	–	0,079
СРБ, мг/л	1,3 [1,23; 1,4]	2,22 [1,41; 3,54]	–	0,0046
Фг, г/л	3,74 [3,69; 3,83]	4,56 [3,69; 4,96]	–	0,048
NT-proNBP, пг/мл	13,8 [13,1; 14,5]	19,1 [14,3; 30,9]	–	0,0028

Примечание: ОШ рассчитывался только для факторов в категориальной форме. Первая группа — продолжающие исследование, вторая группа — умершие от всех причин (пояснения в тексте). апо — аполипопротеин, ДАД — диастолическое артериальное давление, ДИ — доверительный интервал, ИА — индекс атерогенности, ИК — индекс курения, ИМТ — индекс массы тела, ЛВП — липопротеины высокой плотности, ЛНП — липопротеины низкой плотности, Лп(а) — липопротеин(а), МК — мочевая кислота, ОБ — окружность бедер, ОТ — окружность талии, ОШ — отношение шансов, САД — систолическое артериальное давление, СРБ — С-реактивный белок, СТБ — соотношение ОТ/ОБ, ТГ — триглицериды, у.е. — условные единицы, Фг — фибриноген, ХС — холестерин, ЧСС — частота сердечных сокращений, NT-proBNP — N-терминальный фрагмент мозгового натрийуретического пропептида.

предикторов на реализацию конечной точки по данным СГБ-анализа лучшей прогностической модели. Обработка и анализ данных выполнялись на языке R в среде R-studio и на языке Python с помощью пакетов xgsboost и tensorflow.

Результаты

На первом этапе исследования анализировали возможные различия клинико-демографических и функционально-метаболических показателей в группах сравнения (таблица 1).

Результаты сопоставления указывали на достоверные межгрупповые различия 12 факторов, к которым относили возраст обследованных, активный статус курения, ИК, систолическое артериальное давление (САД), частоту сердечных сокращений (ЧСС), уровни глюкозы, СРБ, Фг, NT-proNBP, ОТ, СТБ, ИМТ. Наиболее высоким уровнем достоверности отличались признаки возраста, СТБ и ОТ. Расчет ОШ демонстрировал увеличение риска общей смерти на горизонте 5 лет в 2,9 раза при активном курении табака. При этом факт курения в прошлом и принадлежность к мужскому полу существенно не влияли на вероятность неблагоприятного исхода в анализируемый период. Таким образом, предварительный анализ данных позволил выделить факторы, которые можно рассматривать в качестве потенциальных реклассифицирующих критериев прогноза.

На следующем этапе исследования с помощью однофакторной ЛР определяли весовые коэффициенты, уточняющие прогностический потенциал

Таблица 2

Оценка весовых коэффициентов потенциальных предикторов риска 5-летней смертности

Показатели	Коэффициент	p
Мужской пол	0,57	0,071
Возраст	3,45	<0,0001
Бросившие курить	0,14	0,74
Курящие	1,06	0,0024
ИК, пачка-лет	1,58	0,068
ОТ	3,13	0,0067
ОБ	2,03	0,15
СТБ	3,51	0,0059
ИМТ	2,61	0,101
САД	3	0,004
ДАД	0,70	0,501
ЧСС	2,56	0,0454
Общий ХС	-0,52	0,669
Лп(а)	-1,42	0,162
ХС ЛНП	-0,53	0,658
ХС ЛВП	-2,23	0,067702
ТГ	1,33	0,483
апо AI	-0,8449	0,529
апоВ	0,7648	0,617
ИА	1,8073	0,159
Глюкоза	4,2	0,000003
Креатинин	2,9	0,133
МК	1,67	0,0817
D-димер	2,10	0,252
СРБ	4,56	0,000226
Фг	1,95	0,0475
NT-proNBP	3,813	0,0426

Примечание: ДАД — диастолическое артериальное давление, ИА — индекс атерогенности, ИК — индекс курения, ИМТ — индекс массы тела, ЛВП — липопротеины высокой плотности, ЛНП — липопротеины низкой плотности, Лп(а) — липопротеин(а), МК — мочевая кислота, ОБ — окружность бедер, ОТ — окружность талии, САД — систолическое артериальное давление, СРБ — С-реактивный белок, СТБ — соотношение ОТ/ОБ, ТГ — триглицериды, ХС — холестерин, ЧСС — частота сердечных сокращений, Фг — фибриноген, NT-proBNP — N-терминальный фрагмент мозгового натрийуретического пропептида.

Таблица 3

Оценка точности авторских моделей для прогнозирования риска 5-летней смертности

Прогностические модели и их предикторы	Sen	Spec	AUC	ACC
Предикторы SCORE				
SCORE (1)	0,691	0,702	0,736	0,702
МЛР (2)	0,72	0,716	0,757	0,716
ВР (3)	0,72	0,718	0,758	0,718
СГБ (4)	0,725	0,739	0,762	0,738
Предикторы SCORE + СРБ				
МЛР (5)	0,725	0,727	0,767	0,727
ВР (6)	0,725	0,73	0,768	0,73
СГБ (7)	0,703	0,703	0,777	0,703
Предикторы SCORE + СРБ + глюкоза				
МЛР (8)	0,745	0,7134	0,78	0,714
ВР (9)	0,725	0,748	0,781	0,748
СГБ (10)	0,715	0,73	0,759	0,73
Предикторы SCORE + СРБ + глюкоза + NT-proNBP				
МЛР (11)	0,735	0,731	0,786	0,731
ВР (12)	0,71	0,717	0,784	0,717
СГБ (13)	0,71	0,7	0,766	0,7
Предикторы SCORE + СРБ + глюкоза + NT-proNBP + ЧСС				
МЛР (14)	0,735	0,752	0,786	0,752
ВР (15)	0,735	0,73	0,784	0,73
СГБ (16)	0,71	0,717	0,767	0,717

Примечание: ВР — вейбулл-регрессия, МЛР — многофакторная ЛР, СГБ — стохастический градиентный бустинг, СРБ — С-реактивный белок, ЧСС — частота сердечных сокращений, ACC — точность анализа, AUC — площадь под ROC-кривой, NT-proBNP — N-терминальный фрагмент мозгового натрийуретического пропептида, SCORE — Systematic Coronary Risk Evaluation (шкала оценки риска смерти), Sen — чувствительность, Spec — специфичность.

анализируемых факторов. Данный подход расширяет возможности для отбора предикторов за счет более детальной оценки их взаимосвязей с результирующей переменной (таблица 2). В результате исследования были выделены 10 факторов, весовые коэффициенты которых демонстрировали статистически значимый уровень влияния на конечную точку наблюдения. К ним относились возраст, факт курения, САД, уровни глюкозы, СРБ, Фг, NT-proNBP, ОТ, СТБ, ЧСС. Таким образом, на втором этапе отбора предикторов их количество сократилось на 2 показателя (ИК и ИМТ). Анализ абсолютных значений весовых коэффициентов показал, что доминирующее прямое влияние на вероятность смерти в течение 5 лет оказывали показатели, характеризующие в непрерывной форме содержание в крови СРБ (4,56) и глюкозы (4,2), а также возраст обследованных (4,16). Менее заметную, но достоверную связь с конечной точкой имели параметры NT-proNBP (3,8), СТБ (3,5), ОТ (3,1), САД (3), ЧСС (2,5), Фг (1,9). Влияние на риск смерти факта курения (1,1) было статистически значимым, но менее существенным по сравнению с другими факторами.

На следующем этапе исследования были разработаны и валидированы многофакторные модели оценки риска смерти в течение 5 лет на основе МЛР, ВР и СГБ. В качестве базовых предикторов в указанных моделях использовали 5 факторов шкалы SCORE (возраст, гендерная принадлежность, САД, курение и уровень общего ХС), которые пошагово дополнялись отдельными показателями обследованных из программы ЭССЕ-РФ (таблица 3). Метрики качества разработанных моделей на каждом этапе сопоставлялись с прогностической точностью “классической” шкалы

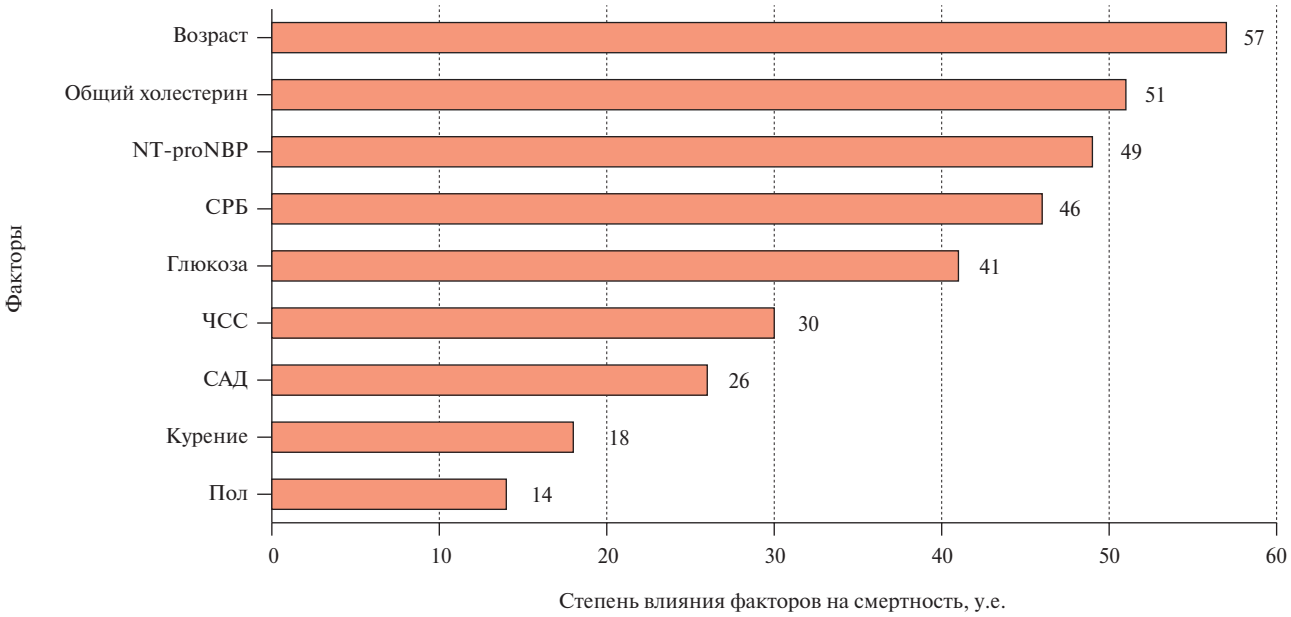

Рис. 1 Относительный вклад отдельных предикторов в реализацию конечной точки.

SCORE. Предварительный расчет риска смерти с использованием этого алгоритма показал, что среднее значение данного показателя в группе умерших составляло 6,2%, Me — 3,7% с интерквартильным размахом [Q25; Q75]: [0,1; 8,3], а в когорте живых аналогичные индикаторы были на уровне 2,1%, Me — 0,7% [0,1; 2,5] ($p<0,0001$). Эти данные свидетельствуют о высокой эффективности шкалы SCORE для оценки популяционного риска смерти на 5-летнем горизонте наблюдения. Вместе с тем, прогностическая система на основе стандартной модели SCORE демонстрировала недостаточную точность стратификации персонифицированных рисков смерти за счет более низкого показателя Sen — 0,69, характеризующего способность идентифицировать долю истинно положительных результатов. При этом авторские модели с использованием только предикторов SCORE отличались более высокими критериями качества. Так, модель (3) на основе ВР обладала лучшими метрикам AUC и Sen (0,76 и 0,745, соответственно), а модель СГБ (4) — более высоким уровнем Spec и точности (ACC) (0,734). Включение в состав предикторов SCORE СРБ повышало показатель ACC, что в большей степени относилось к модели ВР (6), метрика AUC которой составила 0,768. Расширение спектра предикторов за счет включения глюкозы, NT-proNBP и ЧСС демонстрировало наилучшую предсказательную способность у модели (14), разработанной методом МЛР. В этом случае параметр AUC достигал 0,786, ACC и Spec — 0,752, что по классификации С-статистики соответствует хорошей прогностической точности. Последующее пошаговое включение в структуру различных моделей других факторов (показателей СТБ, липидного спектра, Фг, D-димера и др.) демонстрировало либо отсутствие существенной динамики со стороны индикаторов прогностической точности, либо снижение их уровня.

На четвертом этапе исследования методом СГБ определяли вклад отдельных предикторов в реализацию конечной точки исследования (рисунок 1). Анализ проводили по данным модели МЛР (14), обладающей лучшими прогностическими характеристиками. Было установлено, что доминирующее воздействие на результирующую переменную оказывают 5 факторов: возраст обследованных (57 у.е.), уровни общего ХС (51 у.е.), NT-proNBP (49 у.е.), СРБ (46 у.е.) и глюкозы (41 у.е.). Менее заметное влияние на риск смерти в течение 5 лет оказывали показатели ЧСС (30 у.е.), САД (26 у.е.) и факт курения (18 у.е.), а вклад гендерной принадлежности был минимальным (14 у.е.).

Обсуждение

Прогностические исследования в клинической медицине относятся к одним из перспективных и быстро развивающихся направлений, представленных в Национальном проекте “Здравоохранение” 2018г на перспективный период до 2030г (разработан Минздравом России во исполнение Указа Президента Российской Федерации от 7 мая 2018г № 204). Их реализация трудоемка, базируется, как правило, на результатах анализа больших данных, полученных в крупных проспективных исследованиях, одним из наиболее значимых из которых является ЭССЕ-РФ, а также разработке прогностических алгоритмов и программного обеспечения для формирования персонализированных рекомендаций и управления

рисками. В последние годы для решения этих задач все чаще используются методы МО. В отличие от традиционных статистических методов, которые обеспечивают исследование взаимосвязей между ограниченным числом переменных, технологии МО предоставляют возможности для обработки больших и разнородных данных. Кроме того, алгоритмы МО основаны на меньшем количестве допущений и имеют более высокую прогностическую точность [7, 8, 11].

Вероятность развития БСК тесно ассоциирована с факторами ССР, использующимися в многочисленных инструментах прогнозирования, к которым относится шкала SCORE. Последняя была разработана для европейской популяции и широко применяется в РФ [2, 4]. По мнению ряда авторов, недостатками шкалы SCORE являются ограниченность входящих в ее структуру факторов ССР и валидация на популяции в возрастном диапазоне 45-70 лет, что сужает возможности для трансляции этой прогностической системы на более молодые когорты населения [2, 12]. В настоящем исследовании возраст обследованных из регионального сегмента программы ЭССЕ-РФ варьировал от 23 до 67 лет, а спектр факторов риска, включенных в авторские модели, был расширен за счет многоступенчатого алгоритма отбора предикторов с использованием методов математической статистики и расчета весовых коэффициентов. Применение данного подхода позволило уже на предварительном этапе оценить прогностический потенциал анализируемых показателей. Сопоставление эффективности моделей в стратификации риска смерти от всех причин, наиболее значимый вклад в которую носят сердечно-сосудистые события на 5-летнем горизонте наблюдения в сравнении с сердечно-сосудистой смертностью, оцененной с помощью шкалы SCORE, указывало на недостаточную прогностическую точность последней в этом же временном интервале (5 лет). С одной стороны, полученные различия можно объяснить разными конечными точками исследований — общая и сердечно-сосудистая смерть, различными интервалами наблюдения — 5 и 10 лет. С другой стороны, это может быть связано с различными математическими аппаратами, используемыми для анализа результатов и различным набором переменных, используемых для оценки персонифицированного риска участников. Вместе с тем, даже при таких различиях в конечных точках и интервалах наблюдения модель SCORE демонстрировала хорошую эффективность в расчете популяционного риска: 6,2% в группе умерших и 2,1% в когорте продолжающих участие в исследовании.

В настоящем исследовании варианты повышения прогностической эффективности моделей апробировались путем пошагового включения в их структуру новых факторов ССР. При этом максимальный уровень прогностической точности фиксировался на этапе комбинации предикторов SCORE с такими показателями, как СРБ, глюкоза, NT-proNBP и ЧСС. Последовательное включение в модели других факторов, характеризующих функционально-метаболический статус обследованных (показатели липидного спектра, Фг, D-димер, СТБ и др.), существенно не влияло (модели МЛР) или снижало (модели ВР и СГБ) уровень индикаторов качества. Ограничение точности моделей в этих случаях может объясняться проявлениями мультиколлинеарности, влияние которой на качество прогноза возрастает по мере увеличения количества анализируемых факторов [13]. В некоторых исследованиях показано, что попытка модернизации шкалы SCORE за счет включения в структуру этого алгоритма показателя ХС ЛВП не только не улучшало, но даже ухудшало стратификацию пациентов по степени риска [14]. Прогностическое значение повышенного уровня СРБ в оценке риска неблагоприятных исходов определяется доказанной взаимосвязью системного воспаления с заболеваниями атеросклеротического генеза, в т.ч. с фатальными аритмиями, инфарктом миокарда, мозговым инсультом и другими БСК, занимающими ведущее место в структуре смертности населения [15]. В этой работе уровень СРБ среди умерших был в 1,7 выше, чем в когорте живых, что подтверждает прогностическую роль данного фактора как индикатора смертности. Большинство эпидемиологических исследований свидетельствуют о прямой положительной взаимосвязи между высоким риском смерти и уровнем глюкозы в сыворотке крови [2, 3]. Высокий предиктивный потенциал этого показателя, установленный в настоящей работе, подтверждает его прогностическое значение для оценки вероятности фатальных событий. В ранее опубликованных работах, в т.ч. по данным многоцентрового исследования ЭССЕ-РФ, была установлена достоверная взаимосвязь между содержанием NT-proNBP и частотой развития фибрилляции предсердий, инфаркта миокарда, АГ, гипертрофии левого желудочка [16]. Показано, что данный показатель является предиктором неблагоприятного прогноза ишемической болезни сердца [16]. В настоящем исследовании уровень NT-proNBP в группе умерших был на 38% выше, чем в когорте продолжающих участие в исследовании, что также подтверждает его высокий предиктивный потенциал в отношении риска смерти в течение 5 лет. Предсказательная ценность ЧСС в отношении смертности на разных горизонтах наблюдения доказана во многих исследованиях [17]. В нашей работе предиктивный потенциал ЧСС был менее заметным, чем NT-proNBP, СРБ и глюкозы и был сопоставим с САД.

Ограничения исследования могут быть связаны с недостаточным количеством анализируемых факторов и необходимостью расширения спектра используемых методов МО, включая искусственные нейронные сети.

Заключение

В настоящей работе по данным регионального сегмента исследования ЭССЕ-РФ в Приморском крае был апробирован многоступенчатый алгоритм отбора предикторов и разработаны многофакторные прогностические модели риска общей смерти на горизонте 5 лет. Модели МЛР, ВР и СГБ с использованием в их структуре показателей возраста, гендерной принадлежности, факта курения, уровней САД и общего ХС имели более высокие метрики качества, чем примененная в сравнительном аспекте в нетрадиционном временном интервале классическая шкала SCORE. Включение в состав предикторов таких показателей, как СРБ, глюкоза, NT-proNBP и ЧСС, существенно повышало точность модели МЛР. Тестирование предиктивного потенциала других факторов (СТБ, показателей липидного обмена, уровней Фг, D-димера и др.) не улучшало качество прогнозирования. Анализ степени влияния отдельных факторов риска, включенных в исследование ЭССЕ-РФ, на показатель смерти от всех причин указывал на превалирующий вклад 5 факторов: возраста, уровней общего ХС, NT-proNBP, СРБ и глюкозы. Менее заметное влияние ассоциировалось с уровнем ЧСС, САД и курением, а вклад гендерной принадлежности был минимальным. Таким образом, применение современных методов МО повышает надежность прогностических моделей, использование которых в проспективных исследованиях обеспечивает более высокую эффективность риск-стратификации обследованных и позволяет оценить вклад отдельных предикторов в реализацию выбранной конечной точки, что особенно актуально для оценки персонифицированного риска неблагоприятных событий, в т.ч. у лиц более молодого возраста.

Отношения и деятельность. Работа выполнена при поддержке грантов РФФИ 19-29-01077 и 18-29-03131.

Литература/References

1. The European Society for Cardiology. ESC Guidance for the Diagnosis and Management of CV Disease during the COVID-19 Pandemic. https://www.escardio.org/Education/COVID-19-and-Cardiology. (Last update: 10 June 2020).
2. Boytsov SA, Pogosova NV, Bubnova MG, et al. Cardiovascular prevention 2017. National guidelines. Russian Journal of Cardiology. 2018;(6):7-122. (In Russ.) Бойцов С. А., Погосова Н. В., Бубнова М. Г. и др. Кардиоваскулярная профилактика 2017. Российские национальные рекомендации. Российский кардиологический журнал. 2018;23(6):7-122. doi:10.15829/1560-4071-2018-6-7-122.
3. Arnett DK, Blumenthal RS, Albert MA, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: Executive Summary. A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2019;74(10):1376-414. doi:10.1016/j.jacc.2019.03.009.
4. Brodskaya TA, Nevzorova VA, Repina NI, Bogdanov DYu. Cardiovascular Therapy and Prevention. 2017;16(4):93-9. (In Russ.) Бродская Т. А., Невзорова В. А., Репина Н. И., Богданов Д. Ю. Вопросы оценки сердечно-сосудистого риска в зависимости от этнической принадлежности и поражения органов-мишеней. Кардиоваскулярная терапия и профилактика. 2017;16(4):93-9. doi:10.15829/1728-8800-2017-4-93-99.
5. Berger JS, Jordan CO, Lloyd-Jones D, Blumenthal R. Blumenthal screening for cardiovascular risk in asymptomatic patients. Rational Pharmacotherapy in Cardiology. 2010;6(3):381-90. doi:10.20996/1819-6446-2010-6-3-381-390.
6. Boytsov SA, Drapkina OM. Modern content and improvement of high cardiovascular risk strategy in reducing mortality from cardiovascular diseases. Terapevticheskii Arkhiv. 2021;93(1):4-6. (In Russ.) Бойцов С. А., Драпкина О. М. Современное содержание и совершенствование стратегии высокого сердечно-сосудистого риска в снижении смертности от сердечно-сосудистых заболеваний. Терапевтический архив. 2021;93(1):4-6. doi:10.26442/00403660.2021.01.200543.
7. Gusev AV, Gavrilov DV, Korsakov IN, et al. Prospects for the use of machine learning methods for predicting cardiovascular disease. Physician and information technology. 2019;3:41-7. (In Russ.) Гусев А. В., Гаврилов Д. В., Корсаков И. Н. и др. Перспективы использования методов машинного обучения для предсказания сердечно-сосудистых заболеваний. Врач и информационные технологии. 2019;3:41-7.
8. Shakhgeldyan C, Geltser B, Kriger A, et al. Feature Selection Strategy for Intrahospital Mortality Prediction after Coronary Artery Bypass Graft Surgery on an Unbalanced Sample. Proceeding of 4th International Conference on Computer Science and Application Engineering. CSAE. 2020;108:1-7. doi:10.1145/3424978.3425090.
9. Geltser BI, Tsivanyuk MM, Shakhgeldyan KI, Rublev VYu. Machine learning as a tool for diagnostic and prognostic research in coronary artery disease. Russian Journal of Cardiology. 2020;25(12):3999. (In Russ.) Гельцер Б. И., Циванюк М. М., Шахгельдян К. И., Рублев В. Ю. Методы машинного обучения как инструмент диагностических и прогностических исследований при ишемической болезни сердца. Российский кардиологический журнал. 2020;25(12):3999. doi:10.15829/1560-4071-2020-3999.
10. Scientific and Organizing Committee of the Russian Federation essay. The epidemiology of cardiovascular disease in different regions of Russia (ESSE-RF). The rationale for and design of the study. The Russian Journal of Preventive Medicine. 2013;6:25-34. (In Russ.) Научно-организационный комитет проекта ЭССЕ-РФ. Эпидемиология сердечно-сосудистых заболеваний и в различных регионах России (ЭССЕ-РФ). Обоснование и дизайн исследований. Профилактическая медицина. 2013;6:25-34.
11. Plekhova NG, Nevzorova VA, Rodionova LV, et al. Scale of Binary Variables for Predicting Cardiovascular Risk Scale for Predicting

Cardiovascular Risk, 2018. 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). 2018:1-4. doi:10.1109/RPC.2018.8482216.
12. Williams B, Mancia G, Spiering W, et al. ESC Scientific Document Group. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J. 2018;1;39(33):3021-104. doi:10.1093/eurheartj/ehy339.
13. Geltser BI, Shahgeldyan KJ, Rublev VY, et al. Machine Learning Methods for Prediction of Hospital Mortality in Patients with Coronary Heart Disease after Coronary Artery Bypass Grafting. Kardiologiia. 2020;60(10):38-46. (In Russ.) Гельцер Б. И., Шахгельдян К. И., Рублев В. Ю. и др. Методы машинного обучения в прогнозировании летальных исходов в стационаре у больных ишемической болезнью сердца после коронарного шунтирования. Кардиология. 2020;60(10):38-46. doi:10.18087/cardio.2020.10.n1170.
14. Belyalov FI. Application of prediction scores in clinical medicine. Russian Journal of Cardiology. 2016;(12):23-7. (In Russ.) Белялов Ф. И. Использование шкал прогноза в клинической медицине. Российский кардиологический журнал. 2016;(12):23-7. doi:10.15829/1560-4071-2016-12-23-27.
15. Zhukova VA, Shalnova SA, Metelskaya VA. C-reactive protein: modern state of the problem. Cardiovascular Therapy and Prevention. 2011;10(1):90-5. (In Russ.) Жукова В. А., Шальнова С. А., Метельская В. А. С-реактивный белок: современное состояние проблемы. Кардиоваскулярная терапия и профилактика. 2011;10(1):90-5.
16. Shalnova SA, Imaeva AE, Deev AD, et al. Elevated Level of the Natriuretic Peptide Among Adult Population in Regions Participating in the ESSE-RF Study and its Association with Cardiovascular Diseases and Risk Factors. Kardiologiia. 2017;57(12):43-52. (In Russ.) Шальнова С. А., Имаева А. Э., Деев А. Д. и др. Повышенный уровень натрийуретического пептида в популяции взрослого населения регионов — участников ЭССЕ-РФ и его ассоциации с сердечно-сосудистыми заболеваниями и факторами риска. Кардиология. 2017;57(12):43-52. doi:10.18087/cardio.2017.12.10065.
17. Chen X, Barywani SB, Hansson P, et al. Impact of changes in heart rate with age on all-cause death and cardiovascular events in 50-year-old men from the general population. Open Heart. 2019;6:e000856. doi:10.1136/openhrt-2018-000856.