



Адылова Ф.Т.¹ ✉, Тригулова Р.Х.²

¹ Институт математики имени В.И. Романовского Академии наук Республики Узбекистан, Ташкент, Узбекистан

² Республиканский специализированный научно-практический медицинский центр кардиологии, Ташкент, Узбекистан

Модели машинного обучения в прогнозе сердечно-сосудистых заболеваний на бенчмарке мультиэтнических данных: обзор литературы

Конфликт интересов: не заявлен.

Вклад авторов: сбор и анализ источников литературы, написание текста – Адылова Ф.Т.; концепция работы, редактирование статьи – Тригулова Р.Х.

Подана: 29.01.2025

Принята: 13.05.2025

Контакты: dilafuz_akhmedova-9449@mail.ru

Резюме

Методы машинного обучения, такие как методы случайного выживания, могут быть эффективной статистической методологией для обработки биомедицинских данных масштаба big data, скорости и разнообразия, даже большой размерности. Методы не требуют априорных предположений о причинно-следственной связи и потому способны выявить новые биомаркеры в прогнозировании б сердечно-сосудистых заболеваний. В данном обзоре сделана попытка найти ответ на вопрос об эффективности фенотипирования в прогнозе в целом с акцентом на мультиэтнических данных. Показано, что методы машинного обучения лучше подходят для значимого прогнозирования риска в рамках масштабных эпидемиологических исследований с тщательным фенотипированием, чем обычные модели Кокса или оценки риска. Особо выделяются в этом смысле случайные леса выживаемости как инструмент машинного обучения для прогнозирования сердечно-сосудистых заболеваний и оценки риска в больших популяциях.

Ключевые слова: машинное обучение, мультиэтнические данные, прогнозируемые конечные точки, модели прогнозирования, глубокое фенотипирование

Adilova F.¹✉, Trigulova R.²

¹ V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent, Uzbekistan

² Republican Specialized Scientific and Practical Medical Center of Cardiology, Tashkent, Uzbekistan

Machine Learning Models in Cardiovascular Disease Forecasting on Multiethnic Data Benchmark: A Literature Review

Conflict of interest: nothing to declare.

Authors' contribution: collection and analysis of literary sources, text writing – Adilova F.; study concept, editing – Trigulova R.

Submitted: 29.01.2025

Accepted: 13.05.2025

Contacts: dilafruz_akhmedova-9449@mail.ru

Abstract

Machine learning methods, such as random survival methods, can be an effective statistical methodology for processing biomedical data of big data scale, speed and diversity, even those of high dimensionality. The methods do not require a priori causality assumptions, and therefore may reveal new biomarkers in predicting six cardiovascular diseases. In this review, an attempt is made to answer the question about phenotyping's effectiveness in prognosis in general, with an emphasis on multiethnic data. Machine learning methods have been shown to work better in meaningful risk prediction in large-scale epidemiologic studies with thorough phenotyping than conventional Cox models or risk assessment. Random survival forests stand out in this regard as a machine learning tool for predicting cardiovascular diseases and risk assessment in large populations.

Keywords: machine learning, multiethnic data, predicted endpoints, prediction models, deep phenotyping

■ ВВЕДЕНИЕ

Искусственный интеллект относится к вычислительным системам, которые способны выполнять задачи, требующие человеческого интеллекта и принятия решений [1]. Глубокое обучение является особым типом машинного обучения (МО), который может формулировать сложные представления признаков, связывая простые математические функции в нейронной сети [2]. Глубокое обучение использует нейронные сети со многими слоями нейронов, которые могут выявлять сложные зависимости в исходных данных (рис. 1).

Известно, что прогнозирование событий является краеугольным камнем сердечно-сосудистой эпидемиологии, о чем свидетельствуют Фрамингемское и другие проспективные исследования [3]. Их основной целью было предсказание событий на относительно длительные периоды времени, такие как 10 лет или целая жизнь. Исследования, используемые для построения таких прогностических моделей, часто содержат сотни или тысячи переменных, поэтому методы МО могут быть здесь

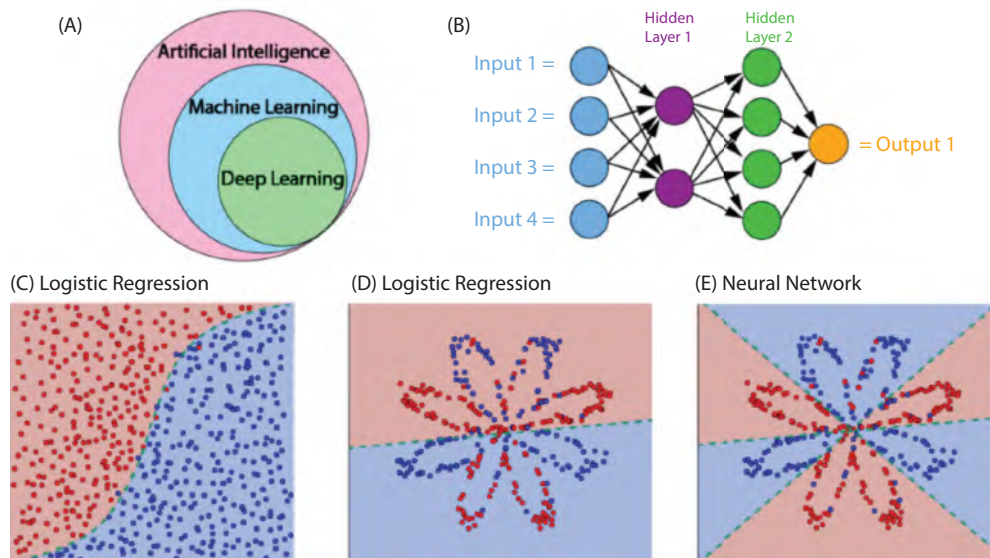


Рис. 1. А – глубокое обучение является подмножеством МО, которое является подмножеством искусственного интеллекта; **В** – архитектура сверточной нейронной сети МО (CNN) с двумя скрытыми слоями. При числе слоев больше 3 архитектура квалифицируется как глубокое обучение; **С–Е** – логистическая регрессия против CNN. Традиционные оценки сердечно-сосудистого риска используют модели логистической регрессии, которые превосходны, когда данные линейно разделимы (**С**), но не так хороши в сложных ситуациях (**Д**), а МО может генерировать более сложные границы принятия решений (**Е**) [1]

Fig. 1. A – deep learning is a subset of machine learning (ML), which is a subset of artificial intelligence; **B** – ML convolutional neural network (CNN) architecture with two hidden layers. When the number of layers is greater than 3, the architecture qualifies as DL; **C–E** – logistic regression vs. CNN. Conventional cardiovascular risk assessments use logistic regression models, which are excellent when data is linearly separable (**C**), but not so good in complicated situations (**D**), while ML can generate more sophisticated decision boundaries (**E**) [1]

полезны как средство выявления наилучших предикторов исходов из миллионов фенотипических данных. Отметим, что кардиология одна из первых стала использовать методы МО [4–6], поскольку они не требуют априорных предположений о причинно-следственной связи и могут применяться для определения роли новых биомаркеров в прогнозировании сердечно-сосудистых заболеваний (ССЗ). МО работает эффективнее в прогнозировании риска в рамках масштабных эпидемиологических исследований с тщательным фенотипированием, т. е. множественной оценкой различных аспектов конкретного патологического процесса. До сих пор эта важная сторона проблемы прогноза была мало изучена. Проводилось исследование, в котором показано, что случайные леса выживаемости могут быть наилучшей стратегией МО для прогнозирования сердечно-сосудистых событий и стратификации риска в значимых по численности популяциях с большими наборами фенотипических данных [7]. Несмотря на то что исходными данными в этом исследовании являются контингенты пациентов с атеросклерозом (и пациенты с сахарным диабетом 2-го типа), оно

продолжает тему, затронутую в наших публикациях [8, 9], в которых удалось выявить различие в предикторах прогноза ССЗ в разных по географии и этническим факторам исследованиях. На сегодня существует мало исследований с использованием методов МО с глубоким фенотипированием для прогнозирования сердечно-сосудистых событий. Поэтому авторы [7] изучили возможность сочетания глубокого фенотипирования с МО для прогнозирования сердечно-сосудистых событий в рамках MESA (многоэтнического исследования атеросклероза). Метод прогнозирования риска, основанный на оценке вероятности случайного выживания, позволил получить совершенно неожиданный взгляд на прогнозирование конкретных исходов, таких как смерть, инсульт, сердечно-сосудистые события, сердечная недостаточность и фибрилляция предсердий, с большей точностью, чем установленные оценки риска. Кроме того, метод объясняет использование важных переменных для прогнозирования конкретных событий и предлагает пути предотвращения исходов ССЗ, основываясь на выявлении механизмов заболевания. В связи с этим целью данного обзора является анализ и обсуждение новых результатов этого исследования.

■ МАТЕРИАЛЫ И МЕТОДЫ

Данные

Исходными данными послужило исследование MESA (Multi-Ethnic Study of Atherosclerosis) – многоэтническое проспективное популяционное когортное исследование атеросклероза, в котором приняли участие 6814 мужчин и женщин, представляющих 4 расовые/этнические группы, в возрасте от 45 до 84 лет без клинически достоверных ССЗ на момент регистрации [10]. В список маркеров MESA включили данные анкет, демографические данные, традиционные факторы риска, антропometriю, прием лекарств, биохимию, магнитно-резонансную томографию сердца и аорты, компьютерную томографию коронарных артерий, ультразвуковое исследование сонных артерий, ЭКГ и лодыжечно-плечевой индекс (ankle-brachial index, ABI).

Прогнозируемые исходы

В качестве конечных точек были использованы: случаи смерти от всех причин, инсульта, всех сердечно-сосудистых заболеваний (ССЗ, CVD), ишемической болезни сердца (ИБС, CHD), фибрилляции предсердий (ФП, AF) и сердечной недостаточности (CH, HF), которые были определены в рамках MESA. Два врача просматривали все медицинские записи для независимой классификации конечных точек и событий. Инсульт определялся как быстрое развитие документально подтвержденного очагового неврологического дефицита (сосудистые причины), длящегося 24 часа или до наступления смерти либо, если менее 24 часов, при клинически значимом поражении головного мозга. Критерии ИБС включали инфаркт миокарда, остановку сердца после реанимации, стенокардию в явном виде, вероятную стенокардию с последующей реваскуляризацией и смерть от ИБС. Сердечно-сосудистые исходы охватывали смерть от ССЗ, инсульт и ИБС. Критерии для определения СН состояли из симптоматической СН, диагностированной врачом и пациентом, получающим медикаментозное лечение по поводу СН, в дополнение к: 1) отеку / застойным явлениям в легких и 2) дилатации желудочка, или плохой функции левого желудочка (ЛЖ), или признакам диастолической дисфункции ЛЖ. Критерии для определения ФП как конечной точки требовали постановки диагноза ФП в стационаре в соответствии с кодами МКБ-9.

Статистический анализ

На рис. 2 показаны процедуры статистического анализа в рамках анализируемого исследования. Часть набора данных (66%) была случайным образом выбрана из общей группы участников для обучения, оставшиеся 33,3% были использованы для тестирования/валидации. Набор обучающих данных был применен для построения 9 моделей с использованием различных подходов и оптимизирован для уменьшения ошибки прогнозирования: Random Forest (все предикторы), Cox Model (все предикторы), LASSO-Cox (все предикторы), AIC-Cox Backward Stepwise Regression (все предикторы), Random Forest (20 RF топ-предикторов), Cox Model (20 RF топ-предикторов), LASSO-Cox (20 RF топ-предикторов), AIC-Cox Backward Stepwise Regression (20 RF топ-предикторов), AIC-Cox Forward Regression (все предикторы). Затем эти модели были опробованы на тестовом наборе данных для проверки эффективности моделей и определения наилучших предикторов.

Тестирование моделей

В ходе анализа протестировали 9 различных моделей. Первая модель использовала алгоритм RSF для всех доступных переменных [11]. RSF – метод дерева ансамблей

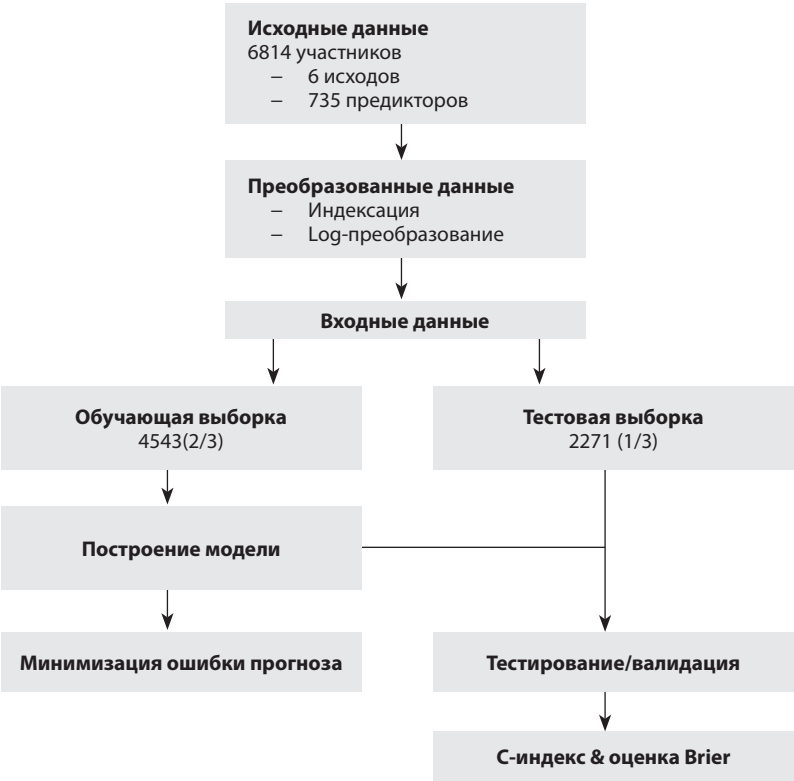


Рис. 2. Блок-схема, описывающая общие рамки исследования. Модели были построены на обучающей выборке, а тестовые данные использовали для вычисления C-индекса и оценки Brier [7]
Fig. 2. Flow chart describing the overall framework of the study. Models were built on the training set, and the test data were used to calculate the C-index and Brier score [7]

для анализа правильно цензурированных данных. Для каждого дерева совокупный уровень опасности случая определяется на основе конечного узла, который его содержит. Совокупная функция опасности (и вероятность выживания) оценивается путем усреднения по всем деревьям.

Информационный критерий Akaike [12] для регрессии Кокса (Akaike Information Criterion – AIC-Cox) с пошаговым исключением в обратном направлении и AIC-Cox с пошаговой регрессией в прямом направлении, а также наименьшая абсолютная усадка и оператор отбора для регрессии Кокса (LASSO-Cox) были протестированы в дополнение к Cox-PHM (Cox Proportional Hazard regression mode). LASSO-Cox минимизирует логарифмическую частичную вероятность при условии, что сумма абсолютных значений параметров ограничена константой, что позволяет эффективно выбирать переменные [13]. Хотя RSF допустимо применять вместо регрессии Кокса для прогнозирования рисков, он может быть более адекватен в качестве эффективного метода выбора переменных. Для этого переменные ранжируют по среднему значению минимальной глубины максимального поддерева (в среднем по 1000 деревьев). Переменные, расположенные выше по дереву, имеют более высокий ранг. Затем 20 переменных, вошедших в топ-20, используют в моделях RSF, AIC-Cox.

Оценка эффективности

Эффективность каждой модели прогнозирования на наборе тестовых данных оценивали по индексу соответствия Харрелла (Harrell concordance index, или C-index) [14], а точность прогнозирования – по шкале Брайера (Brier score, BS) [15]. Высокий C-индекс и низкий BS указывают на более высокую эффективность прогнозирования. Для оценки проблем переобучения были рассчитаны C-индекс и BS для моделей, сгенерированных с использованием подмножеств предикторов (отобранных на основе их возрастающей важности). Для объективности сравнили результаты методов RSF с опубликованными оценками риска [4–6]. Анализ данных проводился на R с использованием доступных библиотек для методов Cox-PHM [16], LASSO-Cox [13], AIC-Cox 5 и RSF [17].

■ РЕЗУЛЬТАТЫ

В исследовании приняли участие 6814 человек. Средний возраст участников составил 62 года, из них 53% женщины, 28% чернокожие, 38% белые, 12% американцы китайского происхождения и 22% испаноязычные. 13% участников страдали сахарным диабетом, 45% были классифицированы как гипертоники в соответствии с критериями Joint National Committee VI, а 50% были курильщиками в настоящем или прошлом. В среднем за 11,2 года (межквартильный интервал 10,6–11,7 года) MESA выявила 831 смерть от всех причин, 710 ССЗ, включая 498 случаев ИБС, среди которых 229 были нефатальными инфарктами миокарда, 200 инсультов, 259 случаев СН и 317 случаев ФП.

Предикторы по исходам

В табл. 1 показаны в качестве примера топ-20 предикторов, использующих RSF для двух исходов, упорядоченных по минимальной глубине максимального поддерева. Эти предикторы были использованы в моделях RSF-20, Cox-20, LASSO-20 и AIC-20.

Таблица 1
Топ-20 переменных, ранжированных по важности методом RSF* [7]
Table 1
Top 20 variables ranked by their importance using the RSF method* [7]

Рей- тинг	ИБС	RVI	Все ССЗ	RVI
1	Оценка уровня кальция в коронарных артериях	0.00	Оценка уровня кальция в коро- нарных артериях	0.00
2	Фактора некроза тканей-α растворимый рецептор	0.28	Фактора некроза тканей-α раство- римый рецептор	0.24
3	Тропонин-Т	0.31	NT-proBNP	0.25
4	NT-proBNP	0.35	Интерлейкин-2 растворимый рецептор	0.28
5	Код Миннесоты 1: лидирующая группа F	0.36	Тропонин-Т	0.35
6	Лодыжечно-плечевой индекс	0.37	Лодыжечно-плечевой индекс	0.40
7	Толщина интима-медиа общей сонной артерии	0.44	Толщина интима-медиа общей сонной артерии	0.41
8	Интерлейкин-2 растворимый	0.48	Пульсовое давление	0.41
9	Пачка – годы курения	0.50	Максимальная площадь восходящей аорты	0.42
10	Толщина интима-медиа внутренней сонной артерии	0.50	Толщина интима-медиа внутрен- ней сонной артерии	0.42
11	Фактор VIII	0.50	Возраст	0.42
12	Конечная систолическая толщина стенки межжелудочковой перегородки	0.52	Амплитуда R в отведении V ₄	0.44
13	Максимальная площадь нисходящей аорты	0.53	Систолическое артериальное давление	0.44
14	Конечная систолическая толщина передне-перегородочной межжелудочковой стенки	0.54	Фактор VIII	0.46
15	Максимальная площадь восходящей аорты	0.54	Растяжимость восходящей аорты	0.47
16	Амплитуда S в отведении aVr	0.55	Соотношение талии/бедер	0.47
17	Конечная диастолическая толщина базаль- ной перегородки	0.55	Код Миннесоты 1: лидирующая группа F	0.48
18	Конечная диастолическая толщина базаль- ной перегородки	0.56	Конечная диастолическая толщи- на базальной перегородки	0.48
19	Пульсовое давление	0.56	Плазмин-антиплазминовый комплекс	0.49
20	Растяжимость нисходящей аорты	0.56	Конечная диастолическая толщи- на базальной перегородки	0.49

Примечания: * относительную важность (RVI) для каждой переменной можно оценить, используя нормализованную минимальную глубину максимального поддерева. Значения RVI варьируются от 0 (наиболее важные) до 1 (наименее важные). NT-proBNP – N-концевой натрийуретический пептид про-B-типа.

Увеличение возраста было наиболее важным предиктором смерти от всех причин. Воспаление и иммунный ответ, измеряемые по повышенным уровням IL-6 (интерлейкина-6), фибриногена, гомоцистеина, TNF-α SR (растворимого рецептора фактора некроза тканей-α) и IL-2 SR, а также нарушения гемостаза, измеряемые по повышенным уровням D-димера, комплекса плазмин – антиплазмин и фактора VIII, были одними из лучших показателей топ-20 маркеров смерти от всех причин, подчеркивающих роль воспаления и тромбоза как распространенных путей развития хронических заболеваний, приводящих к смерти.



Аналогично, сердечный стресс, измеряемый по повышению уровня NT-proBNP (N-Terminal Pro-B-Type Natriuretic Peptide), и повреждение миоцитов из-за повышения уровня сердечного тропонина-Т были одними из главных прогностических маркеров смерти, отражающих роль СН в смертности.

В дополнение к биомаркерам, которые были заметны, низкие и высокие значения ABI, увеличение толщины интимы-медиа сонных артерий, увеличение показателя кальция в коронарных артериях (CAC), размера и растяжимости аорты также показали более низкую вероятность выживания, что указывает на важность атеросклероза и сосудистых аномалий для смертности в общей популяции.

Повышенный уровень глюкозы натощак был наиболее важным фактором риска инсульта, но как высокое давление, так и возраст тоже вошли в топ-20. Показатели атеросклероза, наиболее важным из которых является стеноз сонных артерий, и воспаления также вошли в топ-5 предикторов инсульта.

Как и ожидалось, совокупность показателей, характеризующих атеросклероз (низкий и высокий ABI, увеличение толщины интимы-медиа сонных артерий, снижение растяжимости аорты), была одним из наиболее важных предикторов ИБС, которая представляет собой подгруппу сердечно-сосудистых событий, при этом наиболее важным был CAC, отражающий специфическое влияние коронарного атеросклероза. Увеличение толщины региональной стенки ЛЖ (гипертрофия миокарда), снижение фракции выброса и дилатация аорты, а также биомаркеры нарушения гемостаза, воспаления, стресса и повреждения миокарда были включены в число других основных предикторов ИБС. Важно отметить, что гипертрофия ЛЖ на ЭКГ и значительные нарушения зубца Q и реполяризации были маркерами ИБС и сердечно-сосудистых событий в MESA. Кроме того, среди традиционных факторов риска курение в течение нескольких лет и пульсовое давление были одними из главных предикторов ИБС, в то время как систолическое артериальное давление и возраст были одними из главных предикторов всех ССЗ.

При возникновении СН в качестве конечной точки наиболее важными маркерами были напряжение в камерах сердца (увеличение объема ЛЖ и повышение уровня NT-proBNP) и снижение функции ЛЖ по данным магнитно-резонансной томографии. Гипертрофия ЛЖ на ЭКГ, удлинение интервала QT указывают на повышенный риск тахикардий; повышенный уровень креатинина, повышенная жесткость сосудов, атеросклероз, измеряемый с помощью CAC и ABI, и воспаление были одними из главных предикторов СН. Повышенное пульсовое давление и увеличенное соотношение объема талии и бедер оказались одними из основных факторов риска возникновения СН, что отражает роль ожирения и артериальной гипертензии в развитии СН.

Факторы, влияющие на исходы

Из вышесказанного следует, что переменные, связанные с визуализирующими маркерами, ABI и сывороточными биомаркерами, имели прогностическую значимость от средней до высокой, в то время как опросники и воздействие лекарств имели меньшее значение. Компоненты ЭКГ, относящиеся к сегментам ST, имели промежуточную значимость, в то время как другие показатели ЭКГ имели прогностическую значимость от низкой до средней. Интересно, что только первые 5-6 переменных, перечисленных алгоритмом RSF, давали C-индексы $>0,75$ для прогнозирования ССЗ, ИБС, приступов СН и ФП, что отражает важность NT-proBNP (СН и ФП) и CAC (ИБС

и ССЗ). Однако для прогнозирования смертности от всех причин и инсульта при С-индексе >0,75 требовалась более широкая группа переменных.

Сравнение моделей прогнозирования

В табл. 2 приведены С-индекс и BS для 8 протестированных моделей с применением тестовых наборов данных. Стандартные методы Cox, LASSO-Cox, AIC-Cox не сходились при включении всех 735 переменных, и потому не удалось рассчитать BS и С-индекс. При использовании моделей для получения стабильного и высокого С-индекса для метода RSF было необходимо менее 20 переменных.

Для всех исходов модель RSF с 735 предикторами показала очень высокий С-индекс и низкий BS. Модель RSF-20 была сопоставима и в некоторых случаях даже превосходила модель RSF со всеми предикторами, но обе модели RSF превзошли AIC-Cox с более высоким С-индексом и более низким BS. Использование RSF для отбора переменных в топ-20 (RSF-20) и последующее применение LASSO-Cox, AIC-Cox с обратным шагом привело к тому, что в конечные модели для большинства результатов было отобрано меньшее количество переменных.

Показатели С-индекса у моделей Cox, LASSO-Cox и AIC-Cox с обратным шагом были высокими, а BS – низкими и очень похожи на показатели модели RSF-20. В целом значения С-индекса были выше для прогнозирования краткосрочных событий по сравнению с долгосрочными. Модели МО, которые включали биомаркеры и показатели субклинического заболевания, оказались, как и ожидалось, лучше, чем модели ACC/ANA ASCVD (Американский колледж кардиологии / Американская ассоциация кардиологов (American Heart Association) – атеросклеротические сердечно-сосудистые

Таблица 2
Количество переменных и эффективность (С-индекс и балл Брайера) для каждой из протестированных моделей, а также для оценки риска в конце периода наблюдения* [7]

Table 2
Number of variables and performance (C-index and Brier score) for each of the models tested, as well as for the risk assessment at the end of the follow-up period* [7]

Модели	DTH	STRK	CHD	CVD	HF	AF	DTH	STRK	CHD	CVD	HF	AF
	C-index at 12 years						Brier score at 12 years					
RSF со всеми ковариациями	0.86	0.77	0.81	0.81	0.84	0.82	0.083	0.031	0.067	0.083	0.035	0.039
RSF с 20 лучшими ковариантами	0.84	0.75	0.80	0.80	0.84	0.75	0.076	0.030	0.065	0.079	0.033	0.038
AIC-Cox с прямым выбором	0.78	0.70	0.74	0.74	0.78	0.79	0.088	0.032	0.069	0.087	0.035	0.037
Cox с лучшими 20 RSF-ковариациями	0.80	0.66	0.75	0.76	0.81	0.78	0.086	0.031	0.070	0.087	0.035	0.037
LASSO-Cox с лучшими 20 RSF	0.80	0.67	0.75	0.76	0.82	0.78	0.086	0.031	0.070	0.087	0.33	0.38
AIC-Cox обратная выборка с использованием 20 лучших ковариантов RSF	0.80	0.68	0.75	0.76	0.80	0.78	0.086	0.031	0.069	0.087	0.35	0.038

Примечания: AIC-Cox – информационный критерий Акаике для регрессии Кокса; AF – фибрилляция предсердий; CHD – ишемическая болезнь сердца; CVD – сердечно-сосудистые заболевания; DTH – смерть; HF – сердечная недостаточность; STRK – инсульт.

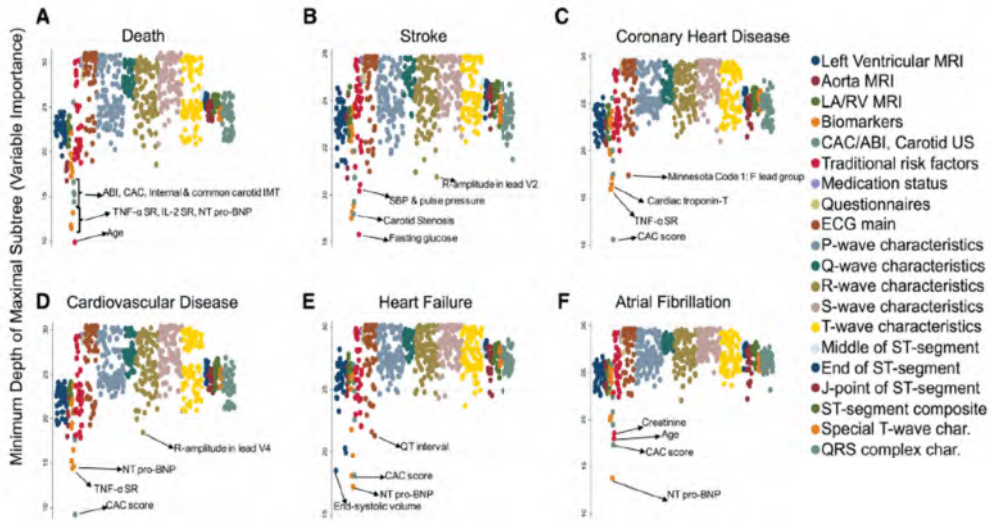


Рис. 3. Графики показывают важность каждой из 735 переменных, использованных в анализе. Цвет точек обозначает категорию или тип измерения. В условных обозначениях справа указаны категории фенотипов, упорядоченные слева направо на отдельных графиках. Более низкие значения отражают большую важность соответствующей переменной. ABI – лодыжечно-плечевой индекс; CAC – показатель кальция в коронарных артериях; IL-2 SR – интерлейкин-2 растворимый рецептор; IMT – толщина интимы-медиа; NT-proBNP – N-концевой натрийуретический пептид про-B-типа; CAD – систолическое артериальное давление; TNF-α SR – растворимый рецептор фактора некроза тканей-α [7]

Fig. 3. Graphs reflecting the importance of each of the 735 variables used in the analysis. The color of the dots represents the category or type of measurement. The legend on the right shows phenotype categories, ordered from left to right in the individual graphs. Lower values reflect greater importance of the corresponding variable. ABI – ankle-brachial index; CAC – coronary artery calcium; IL-2 SR – soluble interleukin-2 receptor; IMT – intima-media thickness; NT-proBNP – N-terminal pro-B-type natriuretic peptide; SBP – systolic blood pressure; TNF-α SR – soluble tissue necrosis factor-α receptor [7]

заболевания; индекс C: 0,73, BS: 0,11) и Фрамингемские (индекс C: 0,73, BS: 0,089) баллы риска для прогнозирования CCЗ.

При сравнении показателей риска для конкретной популяции модель RSF-20, которая включала биомаркеры и показатели субклинического заболевания, оказалась лучше, чем оценка риска ИБС по шкале MESA (C-индекс: 0,79, BS: 0,074) для случаев прогнозирования ХСН [18] (табл. 2). Эффективность модели RSF-20 для прогнозирования ХСН была выше, чем оценка риска MESA HF [19] (C-индекс: 0,80, BS: 0,038).

■ ОБСУЖДЕНИЕ

Результаты этого исследования показывают, что методы МО хорошо подходят для достоверного прогнозирования риска в крупномасштабных эпидемиологических исследованиях с широким фенотипированием. Метод прогнозирования риска, основанный на RSF, обеспечивает лучшее предсказание событий по сравнению со стандартными показателями риска. Основанные на RSF методы выбора переменных и методы регрессии Кокса также позволили улучшить прогнозирование результатов

без проблем с переобучением и несоответствием при учете нелинейностей. Отметим важность глубокого фенотипирования с применением субклинических маркеров, определяемых с помощью визуализации, электрокардиографии и биохимии крови, о чем свидетельствует их присутствие в списках 20 лучших предикторов прогнозирования развития ССЗ.

Интересно, что закономерности прогнозирования варьируются в зависимости от исходов конкретного заболевания. Действительно, если возраст, воспаление, сердечный стресс и сосудистые заболевания являются основными факторами, влияющими на прогнозирование смерти в исследовании MESA, то нарушение метаболизма глюкозы и артериальная гипертензия играют ведущую роль в прогнозировании инсульта, а субклинические маркеры атеросклероза занимают центральное место в прогнозировании общих сердечно-сосудистых событий, будь то заболевания сердца (ИБС) или системного кровообращения. Более того, при возникновении СН и ФП комбинация маркеров, отражающих повышенную нагрузку на камеры сердца в сочетании с нарушением электрической регуляции, является важным фактором, определяющим исход.

Еще одной важной закономерностью стала слабая роль некоторых традиционных факторов риска ССЗ, таких как пол, расовая/этническая принадлежность и воздействие терапии (использование лекарств) среди основных факторов, влияющих на исход заболевания. Меньшая, чем ожидалось, представленность традиционных факторов риска может быть обусловлена тем фактом, что они имеют основное значение для возникновения, поддержания и прогрессирования ССЗ, следовательно, они являются неотъемлемыми компонентами других фенотипов, в частности субклинических, которые более отдалены от начала заболевания, но ближе к неблагоприятным исходам. Несмотря на то что некоторые из этих факторов риска не вошли в топ-20, они по-прежнему имеют решающее значение для медицинской практики, особенно для профилактики заболеваний.

Машинное обучение и глубокое фенотипирование

Применение инструментов МО для диагностики ССЗ по-прежнему вызывает споры, хотя в последние годы все чаще используются визуализирующие тесты, ЭКГ-обследования и лабораторные исследования [20]. В большинстве случаев даже при выявлении нескольких маркеров не все они используются для диагностики [21]. Например, многие клиницисты часто игнорируют показатели региональной функции, полученные с помощью визуализации, большую часть панелей биомаркеров или сигналы ЭКГ. В точной медицине понимание использования фенотипических данных и методов анализа уже полученной информации имеет первостепенное значение. Методы МО, и в частности RSF, и раньше использовались для прогнозирования риска ССЗ [22–31]. В этом исследовании удалось привлечь данные с глубоким фенотипированием для прогнозирования результатов в популяционном исследовании, учитывающем время до наступления события [32, 33]. Преимуществом является то, что методы могут регулярно расширяться и совершенствоваться с учетом новых данных и нелинейности взаимосвязей.

Эта работа также подтверждает влияние определенных маркеров и факторов риска на сердечно-сосудистые события. Например, что важность маркеров TNF- α SR и IL-2 SR, которые до сих пор недооценивались, выходит на первый план. В этом

отношении МО открывает возможность выявления новых взаимосвязей, которые не основаны на гипотезах и предварительных предположениях. Выявление эффективных маркеров заболеваний и неизвестных механизмов может быть полезным для разработки эффективных стратегий скрининга и определения конкретных целей по снижению рисков.

Еще одним преимуществом МО является способность распознавать наилучшие предикторы в рамках предметной области (анкеты, визуализация и т. д.) и их важность по сравнению с предикторами из других предметных областей. Такой подход к идентификации биомаркеров может быть особенно полезен в группах среднего риска, где субклинический риск, лежащий в основе традиционных сердечно-сосудистых факторов риска, не очевиден.

■ ЗАКЛЮЧЕНИЕ

В популяции с широким спектром фенотипирования, не имеющей ССЗ на исходном этапе, используя случайные выборки, продемонстрировано эффективное прогнозирование сердечно-сосудистого риска для конкретных исходов, включая смерть, инсульт, сердечно-сосудистые события, случаи СН и ФП. Воспаление, субклинический атеросклероз, повреждение миокарда и нагрузка на камеры сердца были одними из наиболее важных факторов, влияющих на все исходы.

Работа представляет основу применения больших данных для получения достоверного прогноза риска, идентификации биомаркеров и выработки гипотез на основе данных. MESA, предназначенный для изучения прогрессирования субклинического заболевания с проявлением симптомов и исходов, был проведен среди людей среднего возраста, у которых на начальном этапе не было ССЗ, и поэтому результаты не могут распространяться на другие исследуемые группы населения. Здесь не включали генетические данные, хотя ясно, что выявление взаимодействия между феноменом и геномом, а также оценка их совокупной способности к прогнозированию потенциально могут улучшить результаты. Следовательно, необходима дальнейшая работа, включая валидацию в других популяциях, поскольку наборы данных для обучения и тестирования были получены только из популяции MESA.

■ ЛИТЕРАТУРА/REFERENCES

1. Sung JJ, Stewart CL, Freedman B. Artificial intelligence in health care: preparing for the fifth Industrial Revolution. *The Medical Journal of Australia*. 2020 Sep 1;213(6):253. <https://doi.org/10.5694/mja2.50755>
2. Goodfellow I, Bengio Y, Courville A. (2016) *Deep learning*. MIT Press.
3. Lloyd-Jones DM. Cardiovascular risk prediction. *Circulation*. 2010 Apr 19;121(15):1768–77. <https://doi.org/10.1161/circulationaha.109.849166>
4. D'Agostino RB, Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for use in Primary care. *Circulation*. 2008 Jan 23;117(6):743–53. <https://doi.org/10.1161/circulationaha.107.699579>
5. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Journal of the American College of Cardiology*. 2013 Nov 12;63(25):2935–59. <https://doi.org/10.1016/j.jacc.2013.11.005>
6. Lloyd-Jones DM, Wilson PWF, Larson MG, et al. Framingham risk score and prediction of lifetime risk for coronary heart disease. *The American Journal of Cardiology*. 2004 Jun 24;94(1):20–4. <https://doi.org/10.1016/j.amjcard.2004.03.023>
7. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning. *Circulation Research*. 2017 Aug 10;121(9):1092–101. <https://doi.org/10.1161/circresaha.117.311312>
8. Adilova FT, Trigulova RK, Davronov RR. Artificial intelligence models for predicting cardiovascular diseases in people with type 2 diabetes mellitus: analytical review. *Journal of cardiorespiratory research*. 2024;5(2):33–40. <http://dx.doi.org/10.26739>
9. Adilova FT, Trigulova RK, Davronov RR. Comparative assessment of cardiovascular disease risk prediction in people with type 2 diabetes mellitus: indicators, geography, AI models 2024. *Journal of cardiorespiratory research*. 2024;5(3):19–36. (in Russian) <http://dx.doi.org/10.26739/2181-0974-2024-3>

10. Bild DE, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol*. 2002;156:871–881.
11. Gorodeski EZ, Ishwaran H, Kogalur UB, et al. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the women's health initiative. *Circ Cardiovasc Qual Outcomes*. 2011;4:521–532.
12. Akaike H. Likelihood of a model and information criteria. *J Econom*. 1981;16:3–14.
13. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385–395.
14. Harrell FE, Jr, Califf RM, Pryor DB, et al. Evaluating the yield of medical tests. *JAMA*. 1982;247:2543–2546.
15. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1–3.
16. Cox DR. Regression models and life-tables. *J R Stat Soc Series B*. 1972;34:187–220.
17. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. *Annals Appl Stat*. 2008;2:841–860.
18. McClelland RL, Jorgensen NW, Budoff M, et al. 10-Year Coronary heart Disease risk Prediction using coronary artery calcium and traditional risk factors. *Journal of the American College of Cardiology*. 2015 Oct 1;66(15):1643–53. <https://doi.org/10.1016/j.jacc.2015.08.035>
19. Chahal H, Bluemke DA, Wu CO, et al. Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis. *Heart*. 2014 Nov 7;101(11):58–64. <https://doi.org/10.1136/heartjnl-2014-305697>
20. Andrus BW, Welch HG. Medicare services provided by cardiologists in the United States: 1999–2008. *Circulation Cardiovascular Quality and Outcomes*. 2012 Jan 1;5(1):31–6. <https://doi.org/10.1161/circoutcomes.111.961813>
21. Lanktree MB, Hassell RG, Lahiry P, et al. Phenomics: expanding the role of clinical evaluation in genomic studies. *J Investig Med*. 2010;58:700–706. doi: 10.231/JIM.0b013e3181d844f7
22. Deo RC. Machine learning in medicine. *Circulation*. 2015 Nov 16;132(20):1920–30. <https://doi.org/10.1161/circulationaha.115.001593>
23. Ishwaran H, Blackstone EH, Pothier CE, et al. Relative risk forests for exercise heart rate recovery as a predictor of mortality. *J Am Stat Assoc*. 2004;99:591–600.
24. Inuzuka R, Diller GP, Borgia F, et al. Comprehensive use of cardiopulmonary exercise testing identifies adults with congenital heart disease at increased mortality risk in the medium term. *Circulation*. 2011 Dec 7;125(25):250–9. <https://doi.org/10.1161/circulationaha.111.058719>
25. Hsieh E, Gorodeski EZ, Blackstone EH, et al. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation Cardiovascular Quality and Outcomes*. 2010 Nov 24;4(1):39–45. <https://doi.org/10.1161/circoutcomes.110.939371>
26. Sitar-tăut A, Zdrenghea D, Pop D, et al. Using machine learning algorithms in cardiovascular disease risk evaluation. *Age*. 2009;1:4.
27. Colombet I, Ruelland A, Chatellier G, et al. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp*. 2000:156.
28. Park GM, Han S, Kim SH, et al. Model for assessing cardiovascular risk in a Korean population. *Circulation Cardiovascular Quality and Outcomes*. 2014 Oct 29;7(6):944–51. <https://doi.org/10.1161/circoutcomes.114.001305>
29. Shardell MD, Alley DE, Hicks GE, et al. Low-serum carotenoid concentrations and carotenoid interactions predict mortality in US adults: the Third National Health and Nutrition Examination Survey. *Nutrition Research*. 2011 Mar 1;31(3):178–89. <https://doi.org/10.1016/j.nutres.2011.03.003>
30. Rizza S, Copetti M, Rossi C, et al. Metabolomics signature improves the prediction of cardiovascular events in elderly subjects. *Atherosclerosis*. 2013 Nov 21;232(2):260–4. <https://doi.org/10.1016/j.atherosclerosis.2013.10.029>
31. Rebholz CM, Grams ME, Matsushita K, et al. Change in multiple filtration markers and subsequent risk of cardiovascular disease and mortality. *Clin J Am Soc Nephrol*. 2015;10:941–948.
32. Yoo I, Alafaireet P, Marinov M, et al. Data mining in Healthcare and Biomedicine: A Survey of the literature. *Journal of Medical Systems*. 2011 May 2;36(4):2431–48. <https://doi.org/10.1007/s10916-011-9710-5>
33. Benjamin I, Brown N, Burke G, et al. American Heart Association Cardiovascular Genome-Phenome Study. *Circulation*. 2014 Nov 20;131(1):100–12. <https://doi.org/10.1161/circulationaha.114.014190>