

МОДЕЛЬ ДЛЯ ПРОГНОЗИРОВАНИЯ СМЕРТИ У ВЗРОСЛЫХ ПАЦИЕНТОВ В ТЕЧЕНИЕ 10 ЛЕТ

А.Н. КАФТАНОВ¹, А.Е. АНДРЕЙЧЕНКО², А.Д. ЕРМАК¹, Д.В. ГАВРИЛОВ¹,
А.В. ГУСЕВ³, Р.Э. НОВИЦКИЙ¹

¹ ООО «К-СКАЙ», г. Петрозаводск, Россия;

² ФГАОУ ВО «Национальный исследовательский университет ИТМО», г. Санкт-Петербург, Россия;

³ ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Министерства здравоохранения Российской Федерации, г. Москва, Россия.

УДК 303.092.6

DOI: 10.21045/2782-1676-2025-5-2-4-16

Аннотация

Введение. Выявление факторов риска и прогнозирование смерти от различных причин являются важными вопросами медицины. С точки зрения профилактического направления важно выявлять пациентов с высоким риском смерти, так как раннее обнаружение и лечение заболеваний эффективно повышают продолжительность жизни. **Цель исследования:** разработать универсальную модель прогнозирования смерти у взрослых пациентов в течение 10 лет и сравнить предсказательную способность прогноза смерти в многочисленной современной когорте модели МО (деревья решений) с обычной моделью логистической регрессии Кокса. **Материалы и методы.** Источником данных для исследования являлась база данных платформы прогнозной аналитики Webiomed компании ООО «К-Скай». В исследование было включено 1129268 записей 201985 пациентов в возрасте от 18 лет. Изучено 177 прогностических признаков, из которых в результате многоступенчатого отбора для моделирования выбрано 12. Для моделирования использовались два алгоритма анализа выживаемости: CoxPHFitter и RandomSurvivalForest. С помощью моделей определялась вероятность наступления смерти в течение 1, 3, 5 и 10 лет. **Результаты.** По результатам тестирования обе модели показали хорошие результаты по предсказанию смерти. Однако лучший результат был получен у модели RandomSurvivalForest. Метрики лучшей модели с 95% доверительным интервалом для предсказания смерти в течение 10 лет: Площадь под ROC кривой 0,921 (0,914–0,929), Точность 0,849 (0,84–0,858), Чувствительность 0,813 (0,795–0,83), Специфичность 0,871 (0,859–0,882), Индекс соответствия 0,867 (0,861–0,874), Прогностическая ценность положительного результата 0,791 (0,776–0,806), Прогностическая ценность отрицательного результата 0,886 (0,876–0,895). **Заключение.** Было показано, что модели машинного обучения хорошо предсказывают смертельные исходы, демонстрируя высокую дискриминацию и точность классификации. Их использование может помочь выявлять пациентов высокого риска с целью формирования решения о политике действий для предотвращения смерти.

Ключевые слова: оценка риска смерти, искусственный интеллект, машинное обучение, прогнозирование смерти, анализ выживаемости.

Для цитирования: Кафтанов А.Н., Андрейченко А.Е., Ермак А.Д., Гаврилов Д.В., Гусев А.В., Новицкий Р.Э. Модель для прогнозирования смерти у взрослых пациентов в течение 10 лет. *Общественное здоровье*. 2025; 2(5):4–16, DOI: 10.21045/2782-1676-2025-5-2-4-16

Контактная информация: Кафтанов Алексей Николаевич, e-mail: akaftanov@webiomed.ru

Финансирование: исследование не имело спонсорской поддержки.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Статья поступила в редакцию: 23.04.2025. **Статья принята к печати:** 10.06.2025. **Дата публикации:** 25.06.2025.

UDC 303.092.6

DOI: 10.21045/2782-1676-2025-5-2-4-16

A MODEL FOR PREDICTING DEATH IN ADULT PATIENTS WITHIN 10 YEARS

A.N. Kaftanov¹, A.E. Andreychenko², A.D. Ermak¹, D.V. Gavrilov¹, A.V. Gusev³, R.E. Novitskiy¹

¹ K-Skai, Petrozavodsk, Russia;

² ITMO University, Saint-Petersburg, Russia;

³ Russian Research Institute of Health, Moscow, Russia.

Abstract

Introduction. The identification of risk factors and the prediction of mortality from various causes are important issues in medicine. From a preventive perspective, it is crucial to identify patients at high risk of death, as early detection and treatment of diseases effectively increase life expectancy. **The purpose of the study:** to develop a universal model for predicting death in adult patients within 10 years and to compare the predictive ability of predicting death in a large contemporary

© А.Н. Кафтанов, А.Е. Андрейченко, А.Д. Ермак, Д.В. Гаврилов, А.В. Гусев, Р.Э. Новицкий, 2025 г.

cohort of the machine learning model (decision trees) with a Cox regression. *Materials and methods.* The data source for the study was the database of the Webiomed predictive analytics platform. The study included 1,129,268 records of 201,985 patients aged 18 years and older. 177 predictive features were investigated, of which 12 were selected for modelling as a result of a multi-stage selection process. Two survival analysis algorithms, CoxPHFitter and RandomSurvivalForest, were used for modelling. The models were used to determine the probability of death within 1, 3, 5 and 10 years. *Results.* Both models performed well in predicting death, however, the best result was obtained by the RSF model. Metrics of the best model with 95% CI for predicting death within 10 years: AUC0.921 (0.914–0.929), Accuracy 0.849 (0.84–0.858), Sensitivity 0.813 (0.795–0.83), Specificity 0.871 (0.859–0.882), Concordance index 0.867 (0.861–0.874), Positive predictive value 0.791 (0.776–0.806), Negative Predictive Value 0.886 (0.876–0.895). *Conclusion.* Machine learning models predict mortality outcomes well, demonstrating high discrimination and classification accuracy. Their use may help to identify high-risk patients to inform decisions to prevent death.

Keywords: risk assessment of death, artificial intelligence, machine learning, death prediction, survival analysis.

For citation: Kaftanov A.N., Andreychenko A.E., Ermak A.D., Gavrilov D.V., Gusev A.V., Novitskiy R.E. A model for predicting death in adult patients within 10 years. Public health. 2025; 2(5):4–16, DOI: 10.21045/2782-1676-2025-5-2-4-16

For correspondence: Alexey N. Kaftanov, e-mail: akaftanov@webiomed.ru

Funding: the study had no sponsorship.

Conflict of interests: the authors declare that there is no conflict of interests.

Аннотации на испанском и французском языках приводятся в конце статьи

ВВЕДЕНИЕ

Выявление факторов риска и прогнозирование смерти от различных причин являются важными вопросами медицины. С точки зрения профилактического направления важно выявлять пациентов с высоким риском смерти, так как раннее обнаружение и лечение заболеваний эффективно повышают продолжительность жизни.

Исторически факторы риска определялись с помощью установления связи между каждым предиктором и смертью [1]. Затем для прогнозирования смерти и выявления факторов риска стали использоваться многомерные линейные модели, такие как логистический регрессионный анализ [1, 2]. Хотя линейные модели исторически были популярны из-за их интерпретируемости [1], существовала необходимость в разработке методов оценки риска с более высокой прогностической способностью.

В последние годы в здравоохранении стали широко применяться методы машинного обучения (МО), которые часто позволяют достигнуть более высокой прогностической точности по сравнению с традиционными методами прогноза [1, 2], поскольку они могут учитывать нелинейное взаимодействие между предикторами в дополнение к линейным отношениям [1]. В эпоху больших данных имеется большой оптимизм по поводу того, что МО может потенциально произвести революцию в медицинской помощи, предложить подходы для диагностической и прогностической

оценки наравне с врачами или даже превосходя их [3].

Актуальная проблема применения некоторых методов МО состоит в том, что созданные с их помощью прогнозные модели представляют собой «черные ящики», где неизвестно влияние предикторов на прогноз. В этой связи были предложены подходы к интерпретации результатов анализа МО, например, значения Шепли или SHAP-значения [1, 2]. В итоге можно сказать, что SHAP-значения дают возможность заглядывать в «чёрные ящики» моделей МО. Это позволяет получить представление о том, как именно каждый из прогнозных признаков воздействует на выходные результаты моделей, что повышает доверие к ним при применении в условиях реальной клинической практики.

Таким образом, МО имеет потенциал для улучшения прогнозирования исходов большой сложности и многофакторной причинности, таких как летальные исходы.

Цель исследования: разработать универсальную модель прогнозирования смерти у взрослых пациентов в течение 10 лет и сравнить предсказательную способность прогноза смерти в многочисленной современной когорте модели МО (деревья решений) с обычной моделью логистической регрессии Кокса.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования: работа выполнена в формате ретроспективного когортного

исследования с использованием вторичного источника данных (интегрированные медицинские карты пациентов).

Источником данных для исследования является база данных (БД) платформы прогнозной аналитики Webiomed компании ООО «К-Скай». Данное программное обеспечение предназначено для автоматизации процесса сбора «сырых» медицинских данных из медицинских информационных систем лечебно-профилактических учреждений. В платформе накапливаются следующие обезличенные данные электронных медицинских карт (далее – ЭМК): протоколы врачебных осмотров, данные лабораторной и инструментальной диагностики, коды и дополнительные характеристики диагностированных заболеваний.

Для извлечения информации из ЭМК используются: структурированные запросы к базе данных, язык PL/pgSQL, используемый в СУБД PostgreSQL, библиотеки языка Python (pandas, numpy, matplotlib, scipy, re), методы NLP, поиск по тексту (регулярные выражения), алгоритмический поиск по условию (параметры лабораторных данных), группировка и агрегация формализованных данных разных типов (текстовые, численные), парсинг данных в формате json, pivot table Python (методы сводных таблиц).

Для исследования использовались Python 3.9.0, Jupyter Notebook 6.5.4., библиотеки языка Python: pandas, numpy, matplotlib, scipy, re, sklearn, seaborn, mlxtend, lifelines, scikit-survival.

Для реализации задач исследования использовались обезличенные медицинские записи из лечебно-профилактических медицинских организаций на основании Соглашения о передаче обезличенных данных.

Популяция исследования. Критерии включения:

1. Известен пол и дата рождения.
2. Возраст на дату начала наблюдения не менее 18 лет.
3. Определение класса 1 (пациенты с целевым событием): а) наличие документа «Свидетельство о смерти»; б) указание даты смерти в документе «Свидетельство о смерти»; в) абсолютная разница между датой смерти и датой последнего документа у пациента не превышает 7 дней; г) смерть в течение 11 лет после даты начала наблюдения.
4. Определение класса 0 (пациенты без целевого события): а) отсутствие документа «Свидетельство о смерти» и даты смерти;

б) наблюдение за пациентом в используемой базе данных не менее 11 лет.

Соответствовали критериям включения 201 985 пациентов, 1 129 268 записей:

- Класс 0: 133 372 пациентов, 963 933 записей.
- Класс 1: 68 613 пациентов, 165 335 записей.

Под записью подразумевается информация о ежегодном изменении состоянии здоровья пациента с момента начала наблюдения и до целевого события или до достижения 11 лет. Таким образом, у пациента могло быть от 1-ой, если в течение всего периода наблюдения значения предикторов не изменялись, до 11 записей, если каждый год наблюдения изменялось значение хотя бы одного предиктора.

Полученный набор данных разделен:

1. 20% пациентов случайным образом было выделено в набор для отбора признаков: 40 397 пациентов, 226 114 записей:

- Класс 0: 26 674 пациентов, 192 766 записей.
- Класс 1: 13 723 пациентов, 33 348 записей.

2. 20% пациентов выделено в тестовый набор: 40 397 пациентов, 225 809 записей:

- Класс 0: 26 675 пациентов, 192 877 записей.
- Класс 1: 13 722 пациентов, 32 932 записей.

3. 60% пациентов выделено в обучающий набор: 121 191 пациентов, 677 345 записей:

- Класс 0: 80 023 пациентов, 578 290 записей.
- Класс 1: 41 168 пациентов, 99 055 записей.

Для расчета метрик точности предсказаний моделей для отдельных целевых периодов (1, 3, 5 и 10 лет) тренировочный и тестовый наборы преобразовывались следующим образом:

1) если у записи пациента из класса 1 смерть наступила позже целевого периода, то такая запись пациента переводилась в класс 0 для данного целевого периода;

2) если у записи пациента из класса 0 период наблюдения (длительность между датой записи и датой окончания наблюдения за пациентом) был меньше целевого периода, то такая запись пациента удалялась из набора для данного целевого периода.

Клинические данные и отбор признаков.

Изначально было отобрано 177 признаков, которые присутствовали в БД: пол и возраст, основные факторы риска смерти по данным ВОЗ (11 признаков) [4], ТОП-10 причин смертности среди всех неинфекционных хронических заболеваний по данным Росстата [5], 50 наиболее часто встречающихся в БД Webiomed признаков, 88 признаков из литературных источников, которые не были включены ранее, [1–3, 6–11] и 16 признаков, включенных медицинскими

экспертами. Период сбора клинических данных, вошедших в набор: 08.11.1999 г. – 27.04.2024 г.

Отбор признаков проводился в 4 этапа:

1. Исключение признаков, заполненных менее 10%: исключено 22 признака.

2. Исключение коррелирующих признаков: исключено 4 признака.

3. Многофакторный анализ с учетом изменения признаков с течением времени с помощью алгоритма CoxTimeVaryingFitter (библиотека lifelines): исключено 25 признаков.

4. Отбор лучших признаков с помощью алгоритма SequentialFeatureSelection (библиотека mlxtend) моделью LGBMClassifier на выделенном наборе данных: отобрано 8 признаков.

5. К этим признакам медэкспертом было добавлено 4 признака, которые признаны значимыми для прогноза смерти.

Таким образом, для моделирования было отобрано 12 признаков (таблица 1). Стремление к использованию небольшого количества признаков было продиктовано целью создать универсальную модель.

Заполнение пропущенных значений и устранение выбросов. Стратегия заполнения пропусков:

1. Поля с диагнозами заполнялись «True» для всех последующих периодов после периода первичной регистрации диагноза.

2. Для категориальных признаков пропущенные значения заполнялись «False».

3. Для численных признаков отсутствующие значения заполнялись – 10000.

Устранение выбросов проводилось согласно ГОСТ Р ИСО 16269–4–2017: наблюдения, лежащие за пределами верхней ($Q3 + k(Q3 - Q1)$) и нижней ($Q1 - k(Q3 - Q1)$) внешних границ на диаграмме «ящик с усами» при $k = 3$, рассматривались как явные выбросы. Далее установленные пределы были согласованы медицинскими экспертами с точки зрения того, что пределы не идут в разрез с клинической практикой.

Моделирование. Для моделирования использовалось два алгоритма анализа выживаемости: CoxPHFitter (CP) (библиотека lifelines) [12] и RandomSurvivalForest (RSF) (библиотека scikit-survival) [13]. Методы моделей CP и RSF predict_survival_function возвращают функцию выживаемости, которая, в свою очередь, является вероятностью того, что интересующее событие не наступит до определенного момента времени [14].

Функция выживаемости (S) рассчитывается по формуле:

$$S(t) = \Pr(T \geq t).$$

Вероятность наступления смерти до определенного момента времени (F) рассчитывается по формуле:

$$F(t) = 1 - S(t).$$

С помощью моделей определялась вероятность наступления смерти в течение 1, 3, 5 и 10 лет.

Подбор оптимальных параметров моделей проводился с помощью метода Randomized-SearchCV и техники кросс-валидации из библиотеки Scikit-learn.

Таблица 1

Признаки, отобранные для моделирования

№ п.п.	Название признака
1	АГ/СД/Ожирение
2	Возраст
3	Гематокрит
4	Злокачественные новообразования
5	ИМ/ОНМК
6	Наркотические заболевания
7	Общий белок в моче
8	Пол
9	Сердечная недостаточность
10	Фиброз и цирроз печени (кроме алкогольного)
11	Холестерин липопротеинов высокой плотности (ХС-ЛПВП)
12	SpO2 (Сатурация)

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

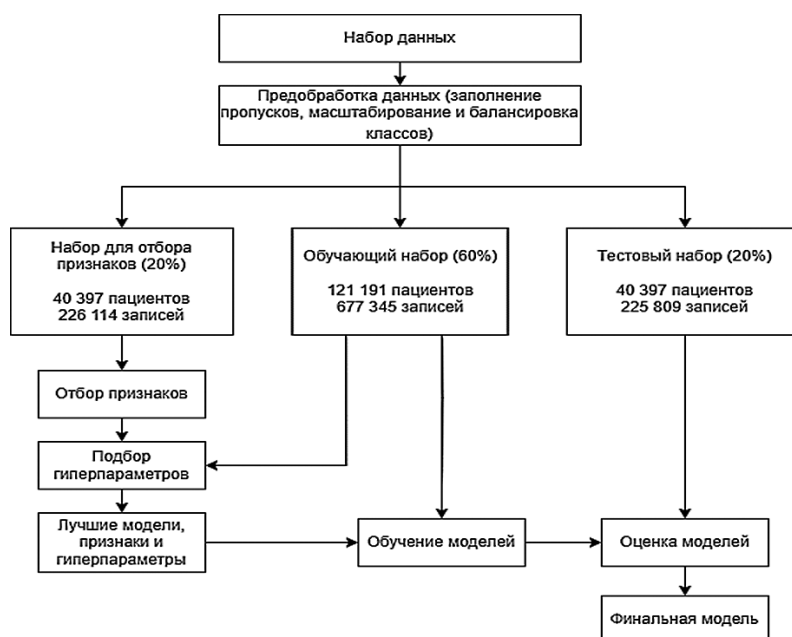


Рис. 1. **Схема исследования: составление наборов данных и разработка моделей**

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

Оценка точности разработанных моделей проводилась по следующим расчетным показателям: Площадь под ROC кривой (AUC / Area Under Curve), Площадь под Precision – Recall кривой (AUPRC), Точность (Accuracy), Полнота или Чувствительность (Recall / True positive rate / Sensitivity), Специфичность (Specificity), Прогностическая ценность положительного результата (ПЦПР) (Precision / Positive predictive value / PPV), Прогностическая ценность отрицательного результата (ПЦОР) (Negative Predictive Value / NPV), Индекс соответствия (Concordance Index), Оценка Брайера (Brier score), Сбалансированная точность (Balanced accuracy).

Общий вид исследования представлен на рис. 1.

Статистический анализ. Количественные признаки представлены в виде числа наблюдений, среднего арифметического, 95% доверительного интервала (ДИ) для среднего (рассчитанного на основе нормального распределения), стандартного отклонения, медианы, минимума и максимума.

Качественные признаки представлены в виде абсолютных частот (количества наблюдений), относительных частот (процентов).

Для проверки гипотез о равенстве средних значений количественных нормально распределенных признаков в зависимости от категории использовался t-тест Стьюдента.

Для оценки различий между двумя независимыми выборками по уровню количественных ненормально распределенных признаков в зависимости от категории в независимых выборках использовался непараметрический U-критерий Манна – Уитни.

Для оценки значимости различий между количеством исходов или качественных характеристик выборки, попадающих в каждую категорию, и теоретическим количеством, которое можно ожидать в изучаемых группах при справедливости нулевой гипотезы (оценка связи между категориальными признаками) использовался непараметрический, одновыборочный тест хи-квадрат Пирсона.

РЕЗУЛЬТАТЫ

Обученная модель RSF на наборе для тестирования продемонстрировала точность по метрике площадь под ROC кривой для 1, 3, 5, 10-летнего прогноза смерти 0,915, 0,898, 0,899, 0,921 соответственно. Обученная модель CP на наборе для тестирования продемонстрировала точность по метрике площадь под ROC кривой для 1, 3, 5, 10-летнего прогноза смерти 0,882, 0,872, 0,880, 0,913 соответственно. Результаты на тестовом наборе получились сопоставимыми с результатами при обучении (рис. 2), о чем говорит пересечение

ДИ метрик на обучающем и тестовом наборах. Однако, лучший результат показала модель RSF, так как была получена более точная предсказательная способность для краткосрочного прогноза по сравнению с моделью СР. С целью оценки прогнозирования целевого события или его отсутствия для конкретного пациента были рассчитаны метрики ПЦПР и ПЦОР обеих моделей: наилучший результат ПЦПР по порогу Юдена показала модель RSF со значениями для 10-летнего прогноза 0,791, и наилучший результат ПЦОР по порогу Юдена для модели RSF и модели СР был аналогичен (0,992 и 0,989 соответственно)

для 1-летнего прогноза. Полное описание метрик моделей представлено в таблицах 2 и 3. В таблице 4 представлены 99% ДИ разницы средних значений метрик Площадь под ROC кривой и Площадь под Precision – Recall кривой моделей, которые показывают, что метрики статистически значимо выше у модели RSF.

Наиболее значимыми признаками как для модели RSF так и для модели СР по векторам Шепли стали возраст, мужской пол, заболевания сосудисто-метаболического профиля (АГ/СД/ожирение). Значимость всех признаков по 2 моделям представлена на рис. 3 и 4.

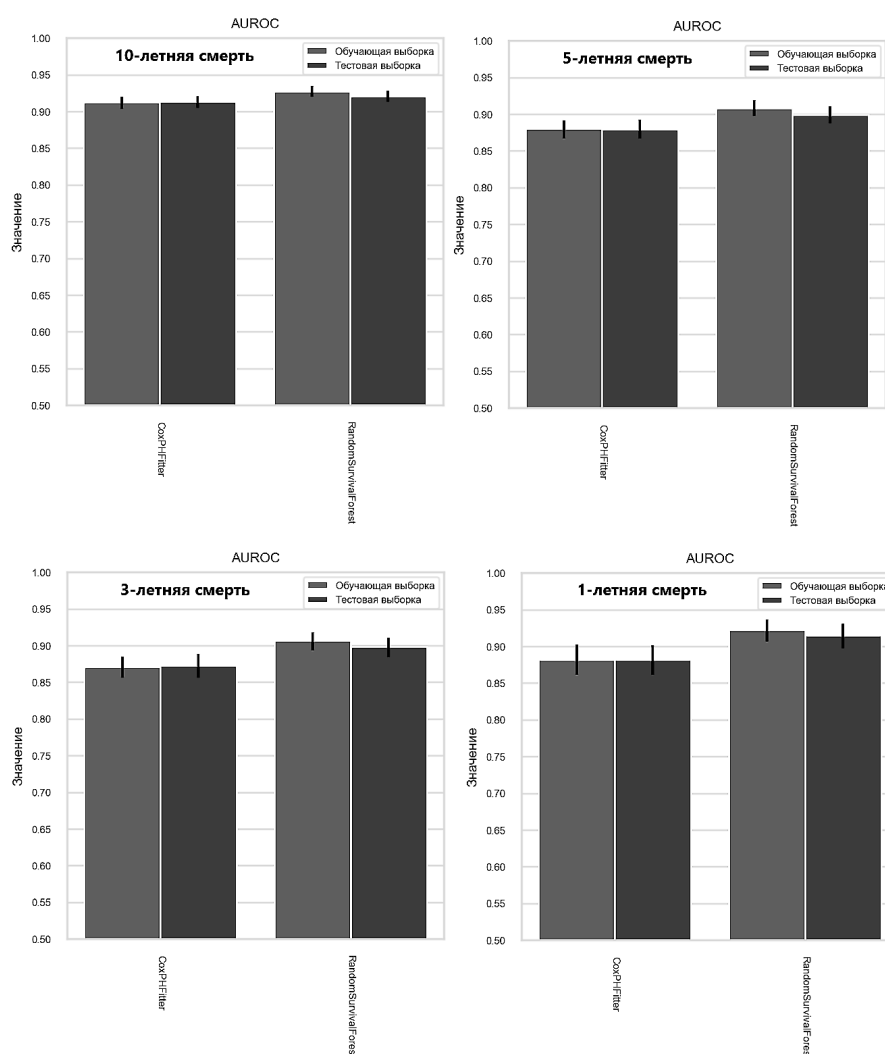


Рис. 2. Площадь под ROC-кривой для моделей RSF и СР на обучающем и тестовом наборах данных

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

Таблица 2

Метрики модели RandomSurvivalForest на тестовом наборе

Период	Тип порога	Порог классификации	TN	FP	FN	TP	AUROC	Доля правильных ответов (Accuracy)	BAS	Полнота (Recall)	Специфичность (Spec)	ПЦПР (Prec)	ПЦОР	Somers' Index	Brier Score
10-летняя смерть	Индекс Юдена	0,172	47789	7063	6169	26744		0,849 [0,84 0,858]	0,842 [0,832 0,852]	0,813 [0,795 0,83]	0,871 [0,859 0,882]	0,791 [0,776 0,806]	0,886 [0,876 0,895]		
	ПЦПР – порог	0,64	54420	432	20566	12347	0,921 [0,914 0,929]	0,761 [0,752 0,769]	0,684 [0,673 0,694]	0,375 [0,355 0,396]	0,992 [0,989 0,995]	0,966 [0,953 0,978]	0,726 [0,719 0,733]	0,867 [0,861 0,874]	0,139 [0,135 0,145]
	ПЦОР – порог	0,025	18219	36633	368	32545		0,578 [0,568 0,589]	0,66 [0,652 0,669]	0,989 [0,983 0,993]	0,332 [0,316 0,349]	0,47 [0,465 0,477]	0,98 [0,971 0,988]		
5-летняя смерть	Индекс Юдена	0,137	112493	23213	5421	23226		0,826 [0,815 0,836]	0,82 [0,805 0,834]	0,811 [0,784 0,836]	0,829 [0,817 0,841]	0,5 [0,481 0,519]	0,954 [0,948 0,96]		
	ПЦПР – порог	0,65	134736	970	21709	6938	0,899 [0,887 0,911]	0,862 [0,856 0,868]	0,618 [0,602 0,634]	0,242 [0,213 0,274]	0,993 [0,99 0,995]	0,877 [0,835 0,917]	0,861 [0,856 0,866]	0,881 [0,87 0,891]	0,087 [0,083 0,091]
	ПЦОР – порог	0,018	40396	95310	348	28299		0,418 [0,406 0,43]	0,643 [0,635 0,651]	0,988 [0,981 0,994]	0,298 [0,283 0,312]	0,229 [0,225 0,233]	0,991 [0,986 0,996]		
3-летняя смерть	Индекс Юдена	0,114	146781	29729	4250	18150		0,829 [0,819 0,84]	0,821 [0,805 0,838]	0,81 [0,78 0,842]	0,832 [0,821 0,843]	0,379 [0,363 0,398]	0,972 [0,968 0,976]		
	ПЦПР – порог	0,58	175081	1429	17737	4663	0,898 [0,885 0,911]	0,904 [0,899 0,908]	0,6 [0,584 0,616]	0,208 [0,178 0,24]	0,992 [0,989 0,995]	0,765 [0,698 0,829]	0,908 [0,905 0,911]	0,888 [0,875 0,901]	0,066 [0,063 0,069]
	ПЦОР – порог	0,01	45073	131437	189	22211		0,338 [0,327 0,35]	0,623 [0,616 0,631]	0,992 [0,982 0,998]	0,255 [0,243 0,269]	0,145 [0,142 0,147]	0,996 [0,992 0,999]		
1-летняя смерть	Индекс Юдена	0,055	178111	36904	1485	9309		0,83 [0,819 0,841]	0,845 [0,821 0,867]	0,862 [0,816 0,904]	0,828 [0,818 0,839]	0,201 [0,189 0,215]	0,992 [0,989 0,994]		
	ПЦПР – порог	0,4	213549	1466	8960	1834	0,915 [0,898 0,931]	0,954 [0,951 0,957]	0,582 [0,558 0,606]	0,17 [0,121 0,218]	0,993 [0,991 0,996]	0,556 [0,446 0,673]	0,96 [0,958 0,962]	0,915 [0,898 0,931]	0,035 [0,034 0,037]
	ПЦОР – порог	0,004	78819	136196	114	10680		0,396 [0,384 0,409]	0,678 [0,668 0,686]	0,989 [0,975 1,]	0,367 [0,354 0,379]	0,073 [0,071 0,074]	0,999 [0,997 1,]		

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

Таблица 3

Метрики модели SoxRNFitter на тестовом наборе

Период	Тип порога	Порог классификации	TN	FP	FN	TP	AUROC	Доля правильных ответов (Accuracy)	BAS	Полнота (Recall)	Специфичность (Spec)	ПЦПР (Prec)	ПЦОР	Concordance Index	Brier Score
10-летняя смерть	Индекс Юдена	0,184	46493	8359	5724	27189		0,84 [0,83 0,849]	0,837 [0,827 0,847]	0,826 [0,81 0,843]	0,848 [0,836 0,86]	0,765 [0,75 0,779]	0,89 [0,881 0,9]		
	ПЦПР – порог	0,45	54519	333	23667	9246	0,913 [0,905 0,921]	0,727 [0,719 0,734]	0,637 [0,627 0,648]	0,281 [0,261 0,302]	0,994 [0,991 0,996]	0,965 [0,95 0,98]	0,697 [0,692 0,703]	0,863 [0,855 0,87]	0,179 [0,175 0,182]
	ПЦОР – порог	0,06	12319	42533	254	32659		0,512 [0,503 0,521]	0,608 [0,601 0,616]	0,992 [0,988 0,996]	0,225 [0,21 0,238]	0,434 [0,43 0,439]	0,98 [0,969 0,99]		
5-летняя смерть	Индекс Юдена	0,152	109242	26464	5643	23004		0,805 [0,793 0,816]	0,804 [0,789 0,818]	0,803 [0,775 0,829]	0,805 [0,793 0,818]	0,465 [0,447 0,483]	0,951 [0,944 0,957]		
	ПЦПР – порог	0,45	134644	1062	23693	4954	0,88 [0,867 0,892]	0,849 [0,844 0,854]	0,583 [0,57 0,595]	0,173 [0,149 0,197]	0,992 [0,989 0,995]	0,823 [0,76 0,879]	0,85 [0,846 0,854]	0,862 [0,851 0,874]	0,106 [0,104 0,109]
	ПЦОР – порог	0,048	26881	108825	250	28397		0,336 [0,326 0,346]	0,595 [0,588 0,601]	0,991 [0,984 0,997]	0,198 [0,186 0,209]	0,207 [0,204 0,21]	0,991 [0,984 0,997]		
3-летняя смерть	Индекс Юдена	0,116	140698	35812	4479	17921		0,797 [0,786 0,808]	0,799 [0,781 0,815]	0,8 [0,769 0,831]	0,797 [0,785 0,808]	0,334 [0,318 0,348]	0,969 [0,964 0,974]		
	ПЦПР – порог	0,39	175059	1451	19435	2965	0,872 [0,858 0,887]	0,895 [0,891 0,899]	0,562 [0,549 0,577]	0,132 [0,107 0,162]	0,992 [0,989 0,994]	0,671 [0,589 0,753]	0,9 [0,897 0,903]	0,863 [0,849 0,877]	0,079 [0,077 0,081]
	ПЦОР – порог	0,03	25672	150838	131	22269		0,241 [0,231 0,25]	0,57 [0,563 0,576]	0,994 [0,988 1,]	0,145 [0,135 0,156]	0,129 [0,127 0,13]	0,995 [0,989 1,]		
1-летняя смерть	Индекс Юдена	0,057	172872	42143	1927	8867		0,805 [0,795 0,815]	0,813 [0,786 0,836]	0,821 [0,77 0,87]	0,804 [0,794 0,815]	0,174 [0,163 0,186]	0,989 [0,986 0,992]		
	ПЦПР – порог	0,23	213362	1653	9662	1132	0,882 [0,86 0,902]	0,95 [0,947 0,953]	0,549 [0,531 0,569]	0,105 [0,071 0,146]	0,992 [0,99 0,995]	0,406 [0,293 0,52]	0,957 [0,955 0,959]	0,882 [0,86 0,902]	0,04 [0,039 0,041]
	ПЦОР – порог	0,013	30780	184235	54	10740		0,184 [0,175 0,194]	0,569 [0,562 0,575]	0,995 [0,983 1,]	0,143 [0,134 0,154]	0,055 [0,054 0,056]	0,998 [0,994 1,]		

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

Таблица 4

99% ДИ разницы средних значений AUROC и AUPRC моделей на тестовом наборе

Период	Модель	AUROC			AUPRC		
		Среднее значение	99% ДИ разницы средних	Значимость различий	Среднее значение	99% ДИ разницы средних	Значимость различий
10-летняя смерть	RandomSurvivalForest	0,921 [0,913 0,929]	0,0074– 0,0084	Различие значимо	0,89 [0,88 0,901]	0,0137– 0,015	Различие значимо
	CoxPHFitter	0,913 [0,905 0,921]			0,875 [0,864 0,887]		
5-летняя смерть	RandomSurvivalForest	0,899 [0,887 0,911]	0,0188– 0,0202	Различие значимо	0,709 [0,679 0,735]	0,0614– 0,0649	Различие значимо
	CoxPHFitter	0,88 [0,867 0,892]			0,645 [0,614 0,678]		
3-летняя смерть	RandomSurvivalForest	0,898 [0,884 0,913]	0,0254– 0,0272	Различие значимо	0,588 [0,549 0,629]	0,0899– 0,0948	Различие значимо
	CoxPHFitter	0,872 [0,858 0,887]			0,495 [0,456 0,535]		
1-летняя смерть	RandomSurvivalForest	0,915 [0,897 0,931]	0,0315– 0,0337	Различие значимо	0,391 [0,336 0,454]	0,1084– 0,1146	Различие значимо
	CoxPHFitter	0,882 [0,86 0,902]			0,279 [0,237 0,338]		

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

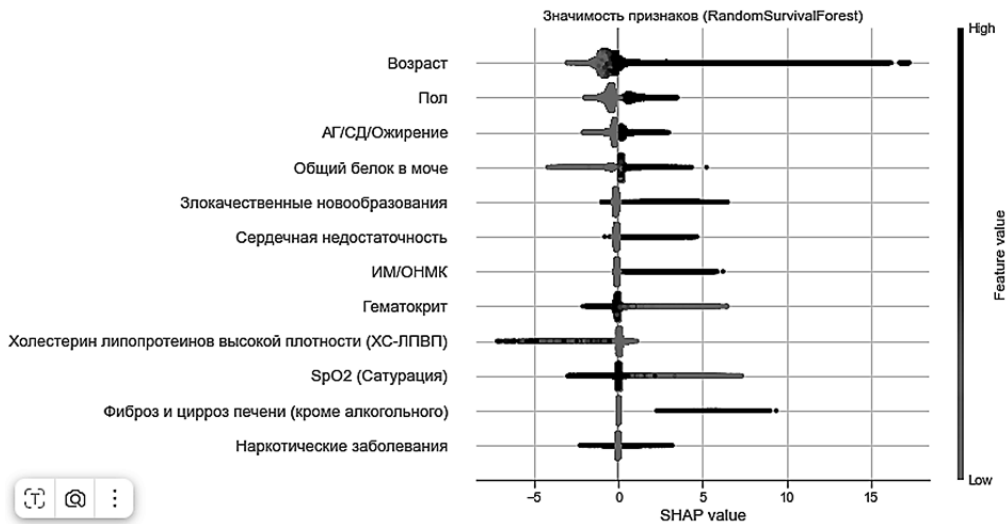


Рис. 3. Значимость признаков модели RandomSurvivalForest

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

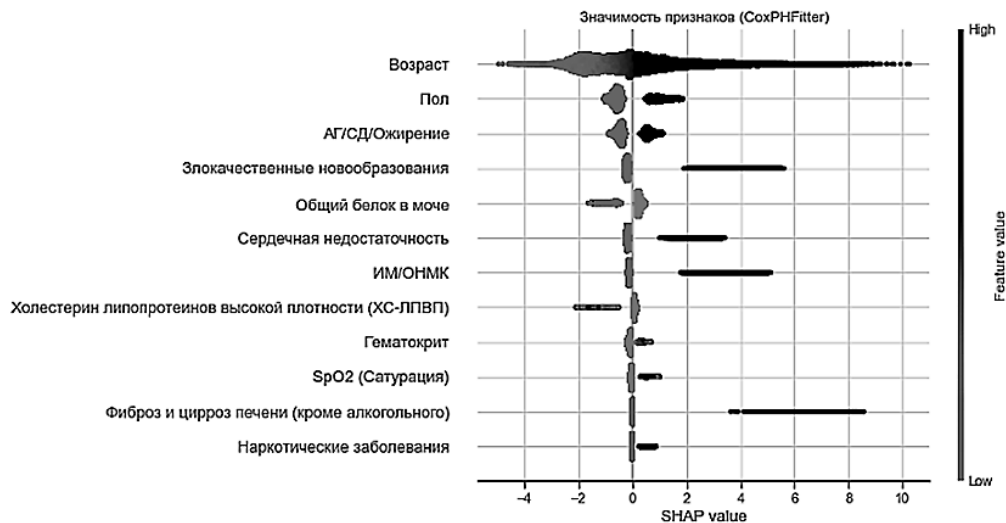


Рис. 4. Значимость признаков модели CoxPHFitter

Источник: составлено авторами по собственным данным, актуальным на 28.11.2024 г.

ОБСУЖДЕНИЕ

Возможность правильно определять пациентов с высоким риском смерти, которые могут получить пользу от усиленных профилактических мер, является одной из основных проблем в медицине. Хотя знания об общепризнанных факторах риска смерти дают основу для общих рекомендаций по профилактике и являются эффективными инструментами для популяционных прогнозов, они мало эффективны для индивидуальной оценки риска, что подчеркивает необходимость более точной оценки прогноза для конкретного пациента.

В более ранних исследованиях сообщалось, что модели на основе МО обладают более высокой прогностической способностью, чем обычные регрессионные модели, как при оценке риска смерти [3], так и риска сердечно-сосудистых заболеваний [15].

В данном исследовании авторами были разработаны две прогностические модели смерти в течение 10 лет на основе алгоритмов анализа выживаемости CoxPHFitter и RandomSurvivalForest. Исследование показало, что метод RSF, основанный на МО, продемонстрировал более высокую прогностическую способность по сравнению с моделью логистической регрессии Кокса. Это указывает на то, что МО может быть полезным для оценки рисков смерти пациентов и улучшения программ медицинских осмотров. Важность объясняющих переменных, оцененная с помощью значений SHAP, была схожей для обеих моделей.

Метрики точности рассматриваемых моделей сопоставимы с аналогичными метриками в более ранних работах по прогнозированию риска смерти (таблица 5).

Такие факторы для прогноза, как возраст, пол, уровень ЛПВП, артериальная гипертензия, сахарный диабет, ожирение, злокачественные новообразования, фиброз и цирроз печени, наркотические заболевания (употребление алкоголя и наркотиков), ИМ и ОНМК в анамнезе являются общепризнанными факторами риска смерти, поэтому результаты данного исследования представляются обоснованными. Общий белок в моче, гематокрит и сатурация были выделены в качестве факторов высокого риска только с помощью методов МО. Хотя механизмы, по которым эти факторы связаны со смертью, до конца не ясны, они могут повысить прогностическую способность для выявления лиц с высоким риском смерти.

Сильной стороной данного исследования является то, что, во-первых, исследование проведено на широкой популяции, отобранной из базы данных, содержащей 374 086 938 медицинских документов 39 935 673 пациентов из 38 субъектов Российской Федерации. Во-вторых, проведен анализ 177 различных предикторов. В-третьих, исследование проводилось на данных, полученных в реальных клинических условиях, что обеспечивает широкую репрезентативность рассматриваемой популяции пациентов. И, наконец, несомненным достоинством этого исследования, проведенного на основе

Таблица 5

Сравнение метрик лучшей модели с литературными данными

Метрика	Webiomed	Liu	Moll	Siga	Lee	Weng	Argyridou	Motwani	Kawano	Qiu
Популяция	Пациенты от 18 лет	Пациенты с заболеваниями коронарных артерий и фибрилляцией предсердий	Пациенты с заболеваниями легких	Пациенты на гемодиализе	Пациенты с диабетом	Пациенты 40–69 лет	Пациенты 38–73 лет	Пациенты с заболеваниями коронарных артерий	Пациенты 40–74 лет	Любые пациенты
Прогноз	10 лет	4 года	3 150 дней	2 года	11 лет	10 лет	10 лет	5 лет	5 лет	10 лет
AUC	0,921	0,732	-	0,78	0,86	0,79	-	0,79	0,811	0,92
accuracy	0,849	0,936	-	-	-	-	-	-	0,908	-
sensitivity	0,813	-	-	0,72	0,87	0,65	-	-	0,445	-
specificity	0,871	0,699	-	0,69	-	0,78	-	-	-	-
c-index	0,867	-	0,731	-	0,87	-	0,722	-	-	-
ПЦПР	0,791	-	-	0,70	-	-	-	-	-	-
ПЦОР	0,886	-	-	0,71	-	-	-	-	-	-

Источник: составлено авторами по данным литературы [1–3, 6–11].

анализа ЭМК, по сравнению с использованием традиционных методов сбора, очистки и систематизации данных, является высокая скорость получения результатов и относительно небольшие затраты ресурсов. Также, насколько авторам известно, данное исследование является первым в России, в котором рассматривался вопрос создания моделей прогнозирования смерти на основе МО.

Ограничения исследования. При проведении исследования источником данных являлись ЭМК, поэтому потенциальными ограничениями были:

- Пропуски данных и некачественное заполнение экранных форм пользователями (ошибки ввода данных).
- Неструктурированные записи представляют собой большинство данных в ЭМК, и извлечение информации из неструктурированных данных ведет к повышению риска возникновения ошибок.
- Риск неполноты информации о пациентах, связанный с тем, что пациенты могут наблюдаться в амбулаторных медицинских организациях, передающих сведения в платформу Webiomed, а лечиться в стационарах, не передающих сведения в Webiomed, и наоборот.
- Использовались не все признаки, которые по данным литературы влияли на смертность из-за ограничений извлечения признаков платформой или встречаемости в ЭМК.

Например, не оценивались образ жизни, диета и физическая активность, хотя считается, что диета и физическая активность напрямую влияют на риск смерти от всех причин [7].

Устранение имеющихся ограничений в дальнейшем при переобучении модели позволит улучшить качество прогноза.

ЗАКЛЮЧЕНИЕ

Авторами настоящего исследования была создана простая в использовании и понятная модель прогнозирования риска смерти у взрослых пациентов в течение 10 лет с соблюдением компромисса между сложностью сбора признаков и производительностью модели. В этом исследовании наглядно показано, что модели МО хорошо предсказывают смерть людей, демонстрируя высокую дискриминацию и точность классификации. Их использование в условиях реальной клинической практики может помочь выявлять пациентов с высоким риском смерти, назначать профилактические мероприятия и осуществлять мониторинг, что в конечном счете будет положительно сказываться на сокращении предотвратимой смертности.

Исследования для прогнозирования с использованием сложных моделей МО, вероятно,

будут расширяться в ближайшие годы, поскольку методы МО продолжают быстро развиваться. Эта работа позволяет предположить, что

использование МО должно более активно применяться при разработке моделей для прогноза или диагностики.

ЛИТЕРАТУРА/ REFERENCES

1. Qiu W., Chen H., Dincer A.B., Lundberg S., Kaeberlein M., Lee S.I. Interpretable machine learning prediction of all-cause mortality. *Commun Med (Lond)*. 2022 Oct 3;2:125. <https://doi.org/10.1038/s43856-022-00180-x>
2. Kawano K., Otaki Y., Suzuki N., Fujimoto S., Iseki K., Moriyama T., Yamagata K., Tsuruya K., Narita I., Kondo M., Shibagaki Y., Kasahara M., Asahi K., Watanabe T., Kenta T. Prediction of mortality risk of health checkup participants using machine learning-based models: the J-SHC study. *Sci Rep*. 2022 Aug 19;12(1):14154. <https://doi.org/10.1038/s41598-022-18276-8>.
3. Weng S.F., Vaz L., Qureshi N., Kai J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One*. 2019 Mar 27;14(3):e0214365. <https://doi.org/10.1371/journal.pone.0214365>.
4. Global health risks: mortality and burden of disease attributable to selected major risks. – Geneva: World health organization, cop. 2009. ISBN 978-92-4-156387-1.
5. Федеральная служба государственной статистики: официальный сайт. – Москва, 2024. – URL: <https://rosstat.gov.ru/folder/12781> (Дата обращения: 06.05.2024).
6. Motwani M., Dey D., Berman D.S., Germano G., Achenbach S., Al-Mallah M.H., Andreini D., Budoff M.J., Cademartiri F., Callister T.Q., Chang H.J., Chinnaiyan K., Chow B.J., Cury R.C., Delago A., Gomez M., Gransar H., Hadamitzky M., Hausleiter J., Hindoyan N., Feuchtner G., Kaufmann P.A., Kim Y.J., Leipsic J., Lin F.Y., Maffei E., Marques H., Pontone G., Raff G., Rubinshtein R., Shaw L.J., Stehli J., Villines T.C., Dunning A., Min J.K., Slomka P.J. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J*. 2017 Feb 14;38(7):500–507. <https://doi.org/10.1093/eurheartj/ehw188>.
7. Argyridou S., Zaccardi F., Davies M.J., Khunti K., Yates T. Walking pace improves all-cause and cardiovascular mortality risk prediction: A UK Biobank prognostic study. *Eur J Prev Cardiol*. 2020 Jul;27(10):1036–1044. <https://doi.org/10.1177/2047487319887281>.
8. Moll M., Qiao D., Regan E.A., Hunninghake G.M., Make B.J., Tal-Singer R., McGeachie M.J., Castaldi P.J., San Jose Estepar R, Washko GR, Wells JM, LaFon D, Strand M, Bowler R.P., Han M.K., Vestbo J., Celli B., Calverley P., Crapo J., Silverman E.K., Hobbs B.D., Cho M.H. Machine Learning and Prediction of All-Cause Mortality in COPD. *Chest*. 2020 Sep;158(3):952–964. <https://doi.org/10.1016/j.chest.2021.03.045>.
9. Siga M.M., Ducher M., Florens N., Roth H., Mahloul N., Fouque D., Fauvel J.P. Prediction of all-cause mortality in haemodialysis patients using a Bayesian network. *Nephrol Dial Transplant*. 2020 Aug 1;35(8):1420–1425. <https://doi.org/10.1093/ndt/gfz295>.
10. Lee S., Zhou J., Leung K.S.K., Wu W.K.K., Wong W.T., Liu T., Wong I.C.K., Jeevaratnam K., Zhang Q., Tse G. Development of a predictive risk model for all-cause mortality in patients with diabetes in Hong Kong. *BMJ Open Diabetes Res Care*. 2021 Jun;9(1):e001950. <https://doi.org/10.1136/bmjdr-2020-001950>.
11. Liu X., Jiang J., Wei L., Xing W., Shang H., Liu G., Liu F. Prediction of all-cause mortality in coronary artery disease patients with atrial fibrillation based on machine learning models. *BMC Cardiovasc Disord*. 2021 Oct 16;21(1):499. <https://doi.org/10.1186/s12872-021-02314-w>.
12. lifelines: official site. – 2024. – URL: https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html#lifelines.fitters.coxph_fitter.SemiParametricPHFitter.predict_survival_function (Date of address: 15.05.2024).
13. Scikit-survival: official site. – 2024. – URL: https://scikit-survival.readthedocs.io/en/stable/api/generated/sksurv.linear_model.CoxnetSurvivalAnalysis.html#sksurv.linear_model.CoxnetSurvivalAnalysis.predict_survival_function (Date of address: 15.05.2024).
14. Ping Wang, Yan Li and Chandan K. Reddy. 2019. Machine Learning for Survival Analysis: A Survey. *ACM Comput. Surv.* 51, 6, Article 110 (November 2019), 36 pages. <https://doi.org/10.1145/3214306>.
15. Unnikrishnan P., Kumar D.K., Poosapadi Arjunan S., Kumar H, Mitchell P., Kawasaki R. Development of Health Parameter Model for Risk Prediction of CVD Using SVM. *Comput Math Methods Med*. 2016 Aug 9. 2016;2016:3016245. <https://doi.org/10.1155/2016/3016245>.

ES

Modelo para la predicción de la muerte en pacientes adultos por un periodo de 10 años

A.N. Kaftanov, A.E. Andreichenko, A.D. Ermak, D.V. Gavrilov, A.V. Gusev, R.E. Novitsky

Anotación

Introducción. La identificación de factores de riesgo y la predicción de la muerte por diversas causas son cuestiones importantes en medicina. Desde una perspectiva preventiva, es crucial identificar a los pacientes con alto riesgo de muerte, ya que la detección y tratamiento tempranos de enfermedades aumentan eficazmente la esperanza de vida. **Objetivo del estudio:** desarrollar un modelo universal para

FR

Un modèle pour prédire la mort des patients adultes dans les 10 ans à venir

A.N. Kaftanov, A.E. Andreychenko, A.D. Ermak, D.V. Gavrilov, A.V. Gusev, R.E. Novitsky

Annotation

Introduction. L'identification des facteurs de risque et la prédiction des décès dus à diverses causes sont des questions importantes en médecine. Du point de vue de la prévention, il est important d'identifier les patients présentant un risque élevé de décès, car la détection et le traitement précoces des maladies augmentent efficacement l'espérance de vie. **Objectif de l'étude:** développer un modèle

la predicción de la muerte en pacientes adultos durante 10 años y comparar la capacidad predictiva del pronóstico de muerte en una amplia cohorte moderna del modelo de árboles de decisión (MO) con el modelo convencional de regresión logística de Cox. *Materiales y métodos.* La fuente de datos para el estudio fue la base de datos de la plataforma de análisis predictivo Webiomed de la empresa LLC «K-Sky». Se incluyeron en el estudio 1,129,268 registros de 201,985 pacientes mayores de 18 años. Se estudiaron 177 características predictivas, de las cuales se seleccionaron 12 para la modelización tras un proceso de selección en múltiples etapas. Se utilizaron dos algoritmos de análisis de supervivencia para la modelización: CoxPHFitter y RandomSurvivalForest. A través de los modelos, se determinó la probabilidad de muerte en 1, 3, 5 y 10 años. *Resultados.* Los resultados de las pruebas mostraron que ambos modelos presentaron buenos resultados en la predicción de la muerte. Sin embargo, el mejor resultado se obtuvo con el modelo RandomSurvivalForest. Métricas del mejor modelo con un intervalo de confianza del 95% para la predicción de muerte a 10 años: Área bajo la curva ROC0.921 (0.914–0.929), Precisión 0.849 (0.84–0.858), Sensibilidad 0.813 (0.795–0.83), Especificidad 0.871 (0.859–0.882), Índice de concordancia 0.867 (0.861–0.874), Valor predictivo positivo 0.791 (0.776–0.806), Valor predictivo negativo 0.886 (0.876–0.895). *Conclusión.* Se demostró que los modelos de aprendizaje automático predicen bien los desenlaces fatales, mostrando alta discriminación y precisión en la clasificación. Su uso puede ayudar a identificar a los pacientes de alto riesgo con el fin de tomar decisiones sobre políticas de acción para prevenir la muerte.

Palabras clave: evaluación del riesgo de muerte, inteligencia artificial, aprendizaje automático, predicción de la muerte, análisis de supervivencia.

universel pour prédire la mort chez les patients adultes sur une période de 10 ans et comparer la capacité de prédiction de la mort dans une grande cohorte contemporaine du modèle MO (arbres de décision) avec le modèle conventionnel de régression logistique de Cox. *Matériaux et méthodes.* La source de données pour l'étude était la base de données de la plateforme d'analyse prédictive Webiomed de K-Sky LLC. 1 129 268 dossiers de 201 985 patients âgés de 18 ans et plus ont été inclus dans l'étude. 177 caractéristiques prédictives ont été étudiées, dont 12 ont été sélectionnées pour la modélisation à la suite d'une sélection en plusieurs étapes. Deux algorithmes d'analyse de survie, CoxPHFitter et RandomSurvivalForest, ont été utilisés pour la modélisation. Les modèles ont été utilisés pour déterminer la probabilité de décès à 1, 3, 5 et 10 ans. *Résultats.* D'après les résultats des tests, les deux modèles ont bien prédit le décès. Cependant, le meilleur résultat a été obtenu par le modèle RandomSurvivalForest. La métrique du meilleur modèle avec un intervalle de confiance de 95% pour prédire la mort dans les 10 ans: Surface sous la courbe ROC0,921 (0,914–0,929), Précision 0,849 (0,84–0,858), Sensibilité 0,813 (0,795–0,83), Spécificité 0,871 (0,859–0,882), Indice d'adéquation 0,867 (0,861–0,874), Valeur pronostique du résultat positif 0,791 (0,776–0,806), Valeur pronostique du résultat négatif 0,886 (0,876–0,895). *Conclusion.* Les modèles d'apprentissage automatique se sont avérés efficaces pour prédire les résultats en matière de mortalité, en faisant preuve d'une grande précision de discrimination et de classification. Leur utilisation peut aider à identifier les patients à haut risque afin de prendre une décision sur une politique d'action pour prévenir la mort.

Mots clés: évaluation du risque de mortalité, intelligence artificielle, apprentissage automatique, prédiction des décès, analyse de survie.

ИНФОРМАЦИЯ ОБ АВТОРАХ / ABOUT THE AUTORS

Кафтанов Алексей Николаевич – кандидат медицинских наук, аналитик данных, ООО «К-СКАЙ», г. Петрозаводск, Россия.
Alexey N. Kaftanov – PhD in Medical sciences, data analyst, K-Skai, Petrozavodsk, Russia.
E-mail: akaftanov@webiomed.ru; ORCID: 0000-0001-6898-8009; SPIN-код: 8768–2083

Андрейченко Анна Евгеньевна – кандидат физико-математических наук, ведущий научный сотрудник, ФГАОУ ВО «Национальный исследовательский университет ИТМО», г. Санкт-Петербург, Россия.
Anna E. Andreychenko – PhD in Physics and Mathematics sciences, lead research scientist, ITMO University, St. Petersburg, Russia.
E-mail: andreychenko.a@gmail.com; ORCID: 0000-0001-6359-0763; SPIN-код: 6625–4186

Ермак Андрей Дмитриевич – аналитик данных, ООО «К-СКАЙ», г. Петрозаводск, Россия.
Andrey D. Ermak – data analyst, K-Skai, Petrozavodsk, Russia.
E-mail: aermak@webiomed.ru; ORCID: 0000-0002-0513-8557

Гаврилов Денис Владимирович – руководитель медицинского направления, ООО «К-СКАЙ», г. Петрозаводск, Россия.
Denis V. Gavrilov – Head of the medical department, K-Skai, Petrozavodsk, Russia.
E-mail: dgavrilov@webiomed.ai; ORCID: 0000-0002-8745-857X; SPIN-код: 2860–6040

Гусев Александр Владимирович – кандидат технических наук, эксперт по искусственному интеллекту, ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Министерства здравоохранения Российской Федерации, г. Москва, Россия.
Aleksandr V. Gusev – PhD in Engineering sciences, artificial intelligence expert, Russian Research Institute of Health, Moscow, Russia.
E-mail: agusev@webiomed.ai; ORCID: 0000-0002-7380-8460; SPIN-код: 9160–7024

Новицкий Роман Эдвардович – генеральный директор, ООО «К-СКАЙ», г. Петрозаводск, Россия.
Roman E. Novitskiy – General manager, K-Skai, Petrozavodsk, Russia.
E-mail: roman@webiomed.ru; ORCID: 0000-0002-2350-977X; SPIN-код: 8309–1740