DOI 10.37469/0507-3758-2025-71-4-OF-2258

Вопросы онкологии, 2025. Том 71, № 4
УДК 616-006.6+616-079
DOI 10.37469/0507-3758-2025-71-4-OF-2258

© *А.Д. Ермак[1], Д.В. Гаврилов[1], Р.Э. Новицкий[1], А.В. Гусев[2,3], Ю.И. Комаров[4], А.Е. Андрейченко[1]*

Использование машинного обучения для прогнозирования онкологических заболеваний на основе данных электронных медицинских карт: автоматизированный подход к скринингу*

[1]Общество с ограниченной ответственностью «К-Скай», г. Петрозаводск, Российская Федерация
[2]Федеральное государственное бюджетное учреждение «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Министерства здравоохранения Российской Федерации, Москва, Российская Федерация
[3]Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», Москва, Российская Федерация
[4]Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр онкологии имени Н.Н. Петрова» Министерства здравоохранения Российской Федерации, Санкт-Петербург, Российская Федерация

© *Andrey D. Ermak[1], Denis V. Gavrilov[1], Roman E. Novitskiy[1], Alexander V. Gusev[2,3], Yuriy I. Komarov[4], Anna E. Andreychenko[1]*

Machine Learning for Cancer Risk Prediction Using Electronic Health Records: An Automated Screening Framework

[1]K-SkAI LLC, Petrozavodsk, the Russian Federation
[2]Federal Research Institute for Health Organization and Informatics, Moscow, the Russian Federation
[3]Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, the Russian Federation
[4]N.N. Petrov National Medical Research Center of Oncology, St. Petersburg, the Russian Federation

Введение. Своевременная диагностика онкологических заболеваний повышает выживаемость пациентов и снижает затраты на здравоохранение за счет сокращения числа госпитализаций и повышения шансов на ремиссию. Сохраняется необходимость в практичных и интерпретируемых инструментах скрининга, которые могут эффективно способствовать раннему выявлению пациентов с онкологическими заболеваниями, для своевременного вмешательства.

Цель. Разработка и внешняя валидация моделей машинного обучения для прогнозирования вероятности развития онкологических заболеваний в течение 18 мес. на основе данных реальной клинической практики.

Материалы и методы. В исследовании использовались анонимизированные данные электронных медицинских карт 1,3 млн пациентов из 36 регионов Российской Федерации. В качестве предикторов рассмотрены пол, возраст, среднее изменение массы тела за месяц, скорость оседания эритроцитов, гемоглобин крови, индекс массы тела и история клинически значимых сопутствующих заболеваний. Целевое событие представлено любым онкологическим заболеванием, определенным по кодам группы С МКБ-10 у 177 384 пациентов. Для сравнения использовались модели Logistic Regression, LGBMClassifier, Random Forest, Linear Discriminant Analysis и Naive Bayes. Внешняя валидация проводилась на данных из регионов с различным географическим происхождением (29 681 и 25 145 пациентов).

Introduction. Timely cancer diagnosis significantly improves patient survival rates while reducing healthcare costs through decreased hospitalizations and increased likelihood of remission. There remains an urgent need for practical, clinically interpretable screening tools capable of effectively identifying at-risk patients to enable early intervention

Aim. To develop and externally validate machine learning models for predicting 18-month cancer risk using real-world clinical data.

Materials and Methods. The study analyzed anonymized electronic health records (EHR) from 1.3 million patients across 36 Russian regions. We examined multiple predictors including sex, age, monthly weight change rate, erythrocyte sedimentation rate, hemoglobin levels, body mass index, and clinically significant comorbidities. The primary outcome was any cancer diagnosis classified under ICD-10 codes C00-C96, which occurred in 177,384 patients. We conducted comparative analysis of five machine learning approaches: Logistic Regression, LightGBM Classifier, Random Forest, Linear Discriminant Analysis, and Naïve Bayes. External validation was performed using two independent geographically distinct patient cohorts (n=29,681 and n=25,145) to evaluate model generalizability across diverse populations.

*Статья содержит онлайн-приложение, в котором размещены дополнительные материалы https://voprosyonkologii.ru/index.php/journal/article/view/4-25-Machine-Learning

Результаты. Модель на основе LGBMClassifier продемонстрировала лучшие результаты с AUROC 0,807 (95 % ДИ 0,798–0,815) при внутреннем тестировании, а также на внешних данных, взятых из отдельного региона и отдельного временного промежутка (0,794 (95 % ДИ 0,786–0,800) и 0,790 (95 % ДИ 0,782–0,798) соответственно).

Заключение. Новый подход с использованием модели машинного обучения, подготовленной на простых и распространенных клинических, лабораторных и анамнестических признаках, продемонстрировал эффективность и практичность применения как на внешних данных, так и по сравнению с предыдущими исследованиями.

Ключевые слова: онкологические заболевания; прогнозные модели; машинное обучение; ранняя диагностика

Для цитирования: Ермак А.Д., Гаврилов Д.В., Новицкий Р.Э., Гусев А.В., Комаров Ю.И., Андрейченко А.Е. Использование машинного обучения для прогнозирования онкологических заболеваний на основе данных электронных медицинских карт: автоматизированный подход к скринингу. *Вопросы онкологии*. 2025; 71(4): 914-926.-DOI: 10.37469/0507-3758-2025-71-4-OF-2258

Results. The LightGBM classifier achieved superior performance, demonstrating an AUROC of 0.807 (95 % CI 0.798-0.815) during internal validation. The model maintained strong discrimination on two independent external validation sets: 0.794 (95 % CI 0.786-0.800) for geographically distinct data and 0.790 (95 % CI 0.782-0.798) for temporally distinct data.

Conclusion. Our machine learning approach utilizing routinely available clinical, laboratory, and anamnestic features proved both effective and practical. The model outperformed existing benchmarks from prior studies while demonstrating consistent performance across external validation cohorts.

Keywords: cancer; predictive models; machine learning; early cancer detection

For Citation: Andrey D. Ermak, Denis V. Gavrilov, Roman E. Novitskiy, Alexander V. Gusev, Yuriy I. Komarov, Anna E. Andreychenko. Machine learning for cancer risk prediction using electronic health records: An automated screening framework. *Voprosy Onkologii = Problems in Oncology*. 2025; 71(4): 914-926.-DOI: 10.37469/0507-3758-2025-71-4-OF-2258

✉ Контакты: Ермак Андрей Дмитриевич, aermak@webiomed.ru

Введение

Своевременная диагностика злокачественных новообразований (ЗНО) приводит к более высоким шансам выздоровления, а также снижению смертности и сопутствующих расходов при ведении пациентов. На ранних стадиях ЗНО требуется меньше сложных схем лечения и госпитализаций, что приводит к снижению затрат на здравоохранение.

Задача определения риска ЗНО у симптомных пациентов и оценка риска их развития у бессимптомных активно исследуется с использованием данных различных биобанков и эпидемиологических исследований [1]. На основе таких работ создаются и периодически актуализируются калькуляторы рисков [2, 3]. Для них предсказательные алгоритмы строятся как на основе прямых статистических методов, так и при помощи методов машинного обучения (МО). Исследование Kulm и соавт. показало, что дискриминативная способность линейных моделей при использовании отобранных десяти предикторов для каждой разновидности ЗНО сравнима с показателями моделей МО и имеющихся калькуляторов рисков [4].

Единственная универсальная модель МО для оценки риска развития в целом ЗНО была описана в работе Miotto R. в 2016 г. [5]. Данная разработка имеет нейросетевую архитектуру и использует для прогнозирования векторное представление пациента на основе данных его электронной медицинской карты. Однако такое представление медицинских сведений невозможно интерпретировать специалистом, что ограничивает возможность использования этого инструмента в реальной клинической практике.

Таким образом, выявление пациентов группы высокого риска развития ЗНО важно для здравоохранения, а создание нового скринингового инструмента для диагностирования таких заболеваний в течение ближайшего времени, обладающего простотой и интерпретируемостью, является актуальной задачей.

Материалы и методы

Источник данных

Проведено многоцентровое когортное ретроспективное исследование с использованием базы данных платформы Webiomed (https: //webiomed.ru/), содержащей обезличенные электронные медицинские карты 50 млн пациентов из 36 регионов Российской Федерации. Информация о каждом пациенте в базе данных представлена в виде множества записей с учетом временных изменений и включает клинические данные о состоянии здоровья, лабораторные и инструментальные исследования, зарегистрированные заболевания. Такой подход позволяет анализировать динамику данных, выявлять тенденции и углубленно исследовать процесс лечения пациента во времени.

Выбор предикторов

На основании анализа научной литературы [6–9] был осуществлен отбор предикторов, наиболее часто упоминаемых в современных исследованиях в качестве классических факторов риска и характерных признаков онкологических заболеваний. Учитывая, что целью исследования было создание универсального инструмента скрининга, применимого на всех этапах оказания медицинской помощи пациентам, в создании модели использовались рутинные клинические и лабораторные параметры, часто применяемые во

врачебной практике и обладавшие значительной заполненностью в используемой базе данных (онлайн-прил. 1).

Критерии включения пациентов и определение целевого события

Для исследования были использованы данные о пациентах с подтвержденным ЗНО (наличие кода из группы С по МКБ-10) и без онкологических заболеваний, но с периодом наблюдения в базе данных не менее 18 мес. Для отобранных пациентов анализировались исторические данные: у пациентов с ЗНО — за последние 18 мес. до постановки диагноза, у пациентов без онкологических заболеваний — за все время наблюдения в базе данных.

Общий период наблюдения за каждым пациентом в базе данных разделялся на цифровые профили — интервалы длительностью 6 мес. с сохранением даты начала и окончания всех интервалов. Для каждого предиктора определялось наличие в базе данных значений, соответствующих определенному цифровому профилю. Правила формирования значений признаков для отдельных цифровых профилей описаны ниже. После распределения значений признаков по профилям в качестве источника соответствующей пациенту записи в итоговом наборе данных (строки со значениями признаков) использовался самый заполненный профиль, а датой прогноза считалась дата окончания интервала, соответствующего этому цифровому профилю.

Можно выделить две группы записей, полученных по итогу описанного алгоритма — записи на основании цифровых профилей пациентов, у которых в течение 18 мес. после определенной нами даты прогноза в листе окончательных диагнозов пациента было зафиксировано появление онкологического диагноза (записи с целевым событием), а также записи пациентов без истории онкологических заболеваний (записи без целевого события). Период прогноза в 18 мес. был выбран на основании исследований, которые показывают, что чаще всего активные симптомы, которые можно выявить с помощью скрининга или диагностики, проявляются в течение 1–1,5 лет после первых проявлений ЗНО.

Формирование значений признаков в пределах цифровых профилей

Для признаков масса тела, рост, гемоглобин крови, скорость оседания эритроцитов (СОЭ) в пределах отдельного цифрового профиля формировались значения, только если их дата извлечения находилась между началом и окончанием соответствующего этому профилю временного интервала. В случае, когда данному условию удовлетворяло несколько значений признака — для формирования записи в наборе данных использовалось ближайшее к дате прогноза.

Пол пациента определен у всех пациентов в базе данных Webiomed. Значение признака возраст соответствовало разнице между датой рождения пациента и датой начала соответствующего цифрового профиля, определенной в годах. Среди извлеченных значений признака «Клинически значимое сопутствующее заболевание» присутствовали только True, отражающие наличие у соответствующего кода по МКБ-10 из списка определенных медицинским экспертом клинически важных для прогноза заболеваний (онлайн-прил. 3).

Для расчета признака индекс массы тела (ИМТ) использовались значения роста и массы тела. В случае отсутствия хотя бы одного из этих признаков в пределах цифрового профиля ИМТ оставляли незаполненным. Признак среднее изменение массы тела за месяц рассчитывался только в случае наличия среди исторических данных пациента еще одного значения массы тела, извлеченного не ранее чем за 12 мес. и не позднее чем за месяц до уже определенного, отнесенного к соответствующему цифровому профилю. При выполнении этого условия рассчитывалась абсолютная разница между двумя значениями массы тела у пациента, которая далее делилась на количество месяцев, прошедших между их датами извлечения. Таким образом получали усредненную динамику изменения массы тела пациента за месяц.

Выделение отдельных наборов для этапов исследования

Сформированный набор данных включал записи от 1 293 583 уникальных пациентов в возрасте на дату прогноза от 18 до 99 лет, за период с 2000 по 2024 гг., из 36 регионов Российской Федерации.

Для внешней валидации разработанной модели были отложены две отдельные выборки. Согласно рекомендациям TRIPOD [10, 11], для надежной оценки обобщающей способности модели важно использовать данные из разных географических регионов или временных периодов, отличающихся от периода сбора данных для разработки. Для внешней валидации были применены оба подхода: 29 681 запись отдельного региона (за период 2020–2024 гг.) и 25 145 записей за 2022 г. из другого региона Российской Федерации («Внешний набор данных № 1» и «Внешний набор данных № 2» соответственно). Записи из последнего, полученные за другие года, использовались в разработке моделей МО.

Данные, использовавшиеся для разработки моделей, для проведения отдельных этапов были рандомно разделены на «Набор данных для обучения» (80 %) и «Набор данных для внутреннего тестирования» (20 %) со стратификацией по целевому событию. Выделение записей на

этом этапе проводилось при обязательном условии: информация о каждом отдельном пациенте присутствовала строго только в одном из итоговых наборов, что исключало утечку данных [12]. Общий дизайн исследования представлен на рис. 1.

Статистический анализ

Для статистического анализа, обучения и анализа моделей МО использовали язык программирования Python версии 3.9.16. Результаты анализа количественных данных представлены в виде медианы с указанием межквартильного размаха, а категориальных — в виде долей. Сравнение количественных переменных между группами с целевым событием и без него проводилось с использованием теста Манна — Уитни, категориальных — с помощью $\chi 2$. Значение $p < 0,05$ принималось за статистически значимое.

Анализ пропущенных данных проводился с использованием теста Little's MCAR Test, который позволяет оценить гипотезу о случайности пропущенных значений [13]. Значение $p < 0,05$ свидетельствовало о том, что механизм пропущенных значений не является полностью случайным.

Для исследования эффективности моделей использовали набор классических метрик для бинарной классификации [14]: площадь под характеристической кривой (AUROC), чувствительность, специфичность, точность, сбалансированная точность, прогностические ценности положительного (ПЦПР) и отрицательного результатов (ПЦОР), а также площадь под кривой ПЦПР-чувствительности (AUPRC). Доверительные интервалы значений выбранных метрик оценивались с помощью метода бутстрап путем случайной генерации тысячи псевдовыборок из 20 тыс. записей [15]. В качестве порогов активации использовали максимум индекса Юдена, а также целевые уровни прогностических ценностей отрицательного (0,99) и положительного (0,5) результата, полученные при внутреннем тестировании [16]. Относительную значимость предикторов определяли по методу Шепли [17]. Калибровка итоговых моделей оценивалась с помощью показателя Brier и калибровочных кривых [18].

Обработка данных

При обработке наборов данных к пропущенным значениям в количественных признаках, а также значениям, выходящим за установленные медицинским экспертом границы, применялось заполнение фиксированной константой «–10 000», медианой или средним, определенными по данным в наборе для обучения [19]. Заполнение пропусков в признаке «Клинически значимое сопутствующее заболевание» проводилось с помощью нулей. На этапе масштабирования использовали гистограммную нормализацию, стандартизацию или данные в изначальной размерности [20, 21]. Для коррекции дисбаланса записей по целевому событию использовались алгоритмы Random Undersampling, Random Oversampling и Synthetic Minority Oversampling Technique либо данные оставляли с изначальным балансом [22].

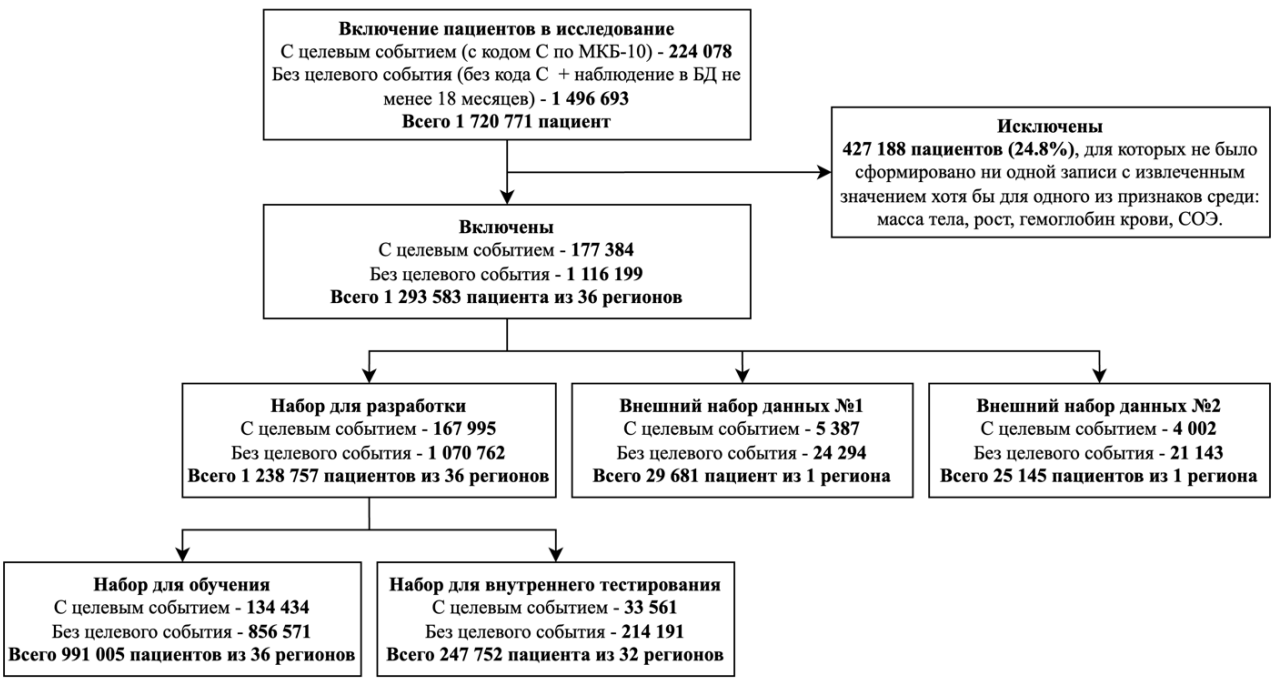

Рис. 1. Дизайн исследования: формирование наборов данных для обучения и валидации
Fig. 1. Study design: Data cohort formation for model training and validation

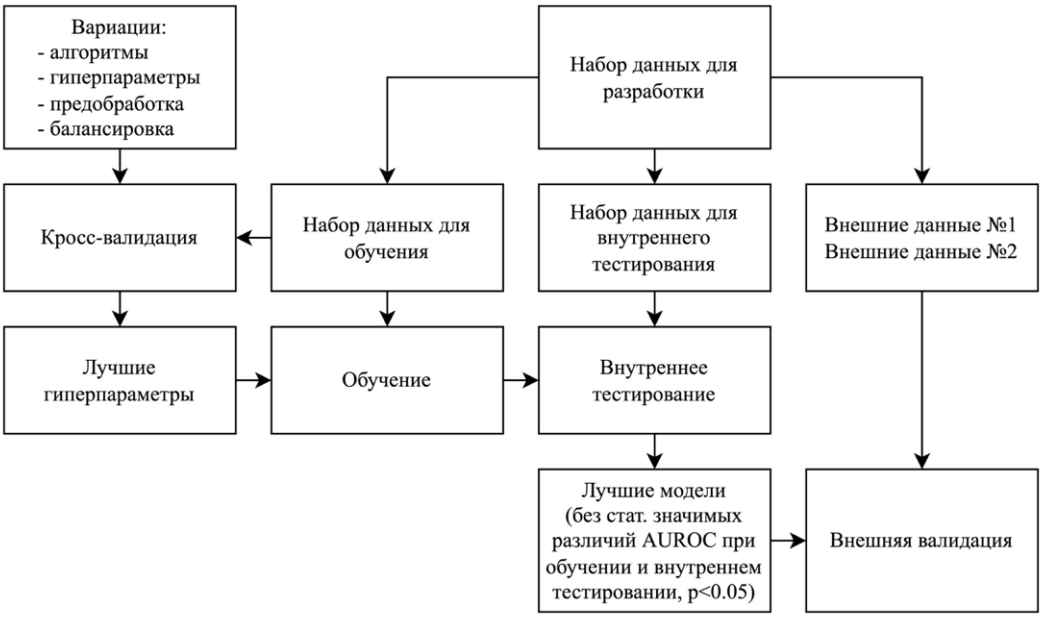

Рис. 2. Алгоритм обучения и оценки дискриминативной способности моделей
Fig. 2. A pipeline for model training and evaluation

Моделирование

В качестве потенциальных архитектур были исследованы: логистическая регрессия с L2 регуляризацией, ансамблевые алгоритмы на основе деревьев решений с градиентным бустингом — LGBMClassifier [23], и бэггингом — Random Forest [24], а также линейный дискриминантный анализ и наивный байесовский классификатор. На основе каждой архитектуры последовательно обучали модели при использовании всех возможных подходов к обработке данных с применением различных комбинаций заполнения пропусков, масштабирования и коррекции дисбаланса.

Перед непосредственным обучением моделей определялись оптимальные гиперпараметры для всех архитектур при соответствующей обработке данных. Для этого проводили кросс-валидацию на наборе для обучения с использованием пяти фолдов и стратификацией по целевому событию. Поиск гиперпараметров был построен на алгоритме Optuna [25] с использованием 100 итераций по предложенным сеткам гиперпараметров и AUROC в качестве целевой метрики для максимизации.

На следующем этапе модели обучались на всех записях из обучающего набора при оптимальных гиперпараметрах, калибровались с использованием изотонической регрессии и анализировались на записях из набора данных для внутреннего тестирования.

До внешней валидации допускались те модели, при исследовании которых доверительный интервал разницы AUROC, полученной при обучении и внутреннем тестировании, включал в себя ноль [26]. В ситуации, когда, по итогам обучения на данных в разных обработке, для одной архитектуры было получено несколько моделей, удовлетворяющих этому условию, предпочтение отдавалось модели с наибольшим значением AUROC при тестировании. Выбор итоговой модели основывался на максимальном значении AUROC по итогам проведенной внешней валидации при условии пересечения ДИ метрики при внутреннем тестировании и внешней валидации. Схема обучения и оценки дискриминативной способности моделей представлена на рис. 2.

Результаты

Описательная статистика

Заполненность признаков в наборе данных до разделения его на выборки для разработки и внешней валидации отражена на рис. 3, а. У всех пациентов были известны пол и возраст на дату прогноза, для количественных признаков — гемоглобин крови, СОЭ, среднее изменение массы тела за месяц, ИМТ — были извлечены соответствующие значения из базы данных по меньшей мере в 40 % записей; графики распределения их значений представлены в онлайн-прил. 2.

При проведении Little's MCAR теста получено значение $p < 0,05$. Результат свидетельствовал о том, что пропуски в данных не являлись полностью случайными, что соответствовало нашим ожиданиям. Действительно, ИМТ и среднее изменение массы тела за месяц связаны с заполненностью массы тела, тогда как гемоглобин крови и СОЭ происходят из одного источника —

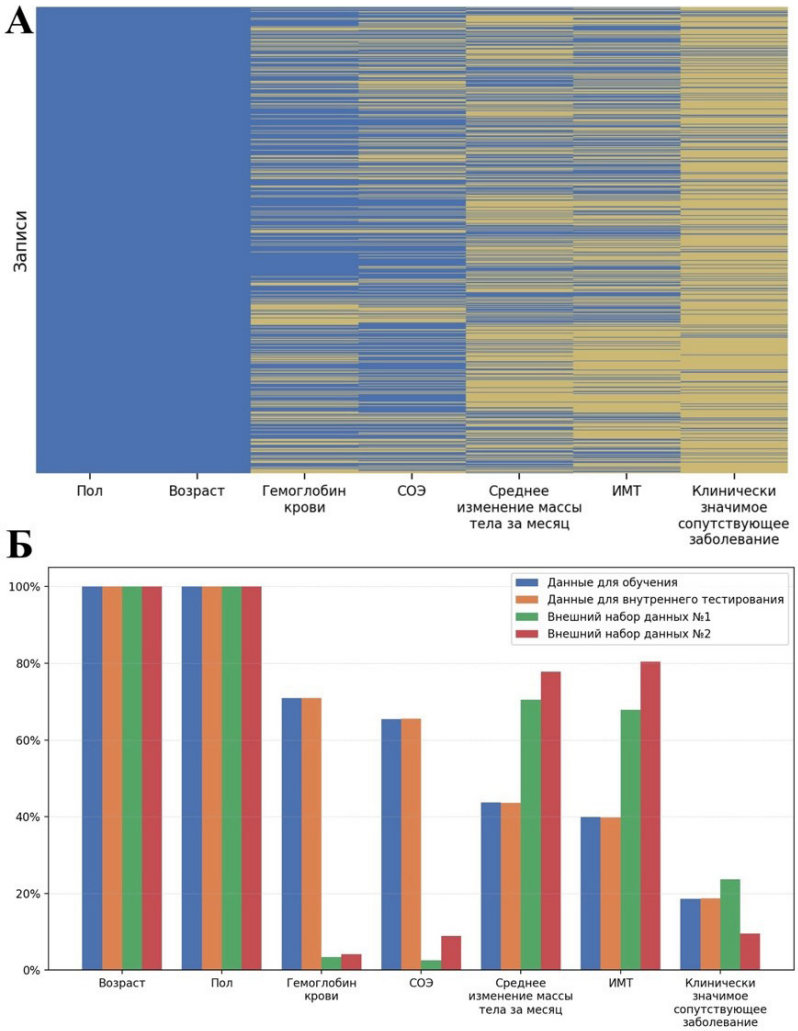

Рис. 3. Заполненность признаков: А — среди всех сформированных записей (поля, выделенные синим цветом, обозначают заполненные значения признаков; поля, выделенные желтым цветом, обозначают пропуски), Б — по отдельным выборкам. Для признака «Клинически значимое сопутствующее заболевание» отражена заполненность значениями True
Fig. 3. Data completeness analysis: (A) Full dataset feature completeness (blue = observed values, yellow = missing data). For "Clinically significant comorbidity", only "True" values are shown. (Б) Feature completeness across distinct patient subsets (with "Clinically significant comorbidity" reporting only "True" values).

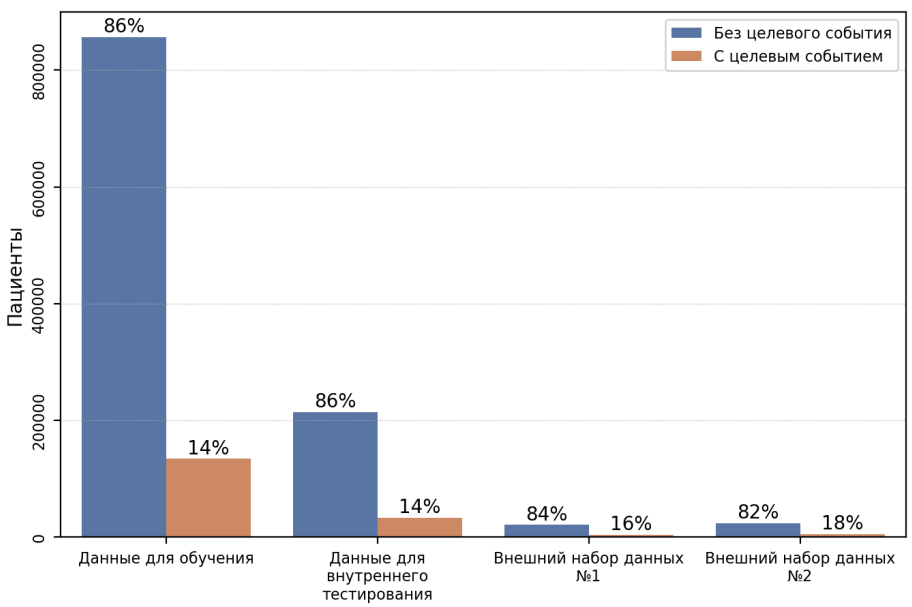

Рис. 4. Распределение записей по целевому событию в отдельных выборках
Fig. 4. Distribution of records by the target event across separate sets

результатов общего анализа крови. Для дальнейшего изучения были проведены статистические сравнения между записями с и без пропущенных значений по упомянутым четырем признакам. Во всех случаях не было выявлено значимых различий между группами. Это подтвердило отсутствие систематических искажений и оправдало применение простых методов заполнения, не требующих сложного моделирования.

После исключения всех записей для внешней валидации был сформирован набор данных для разработки, в который вошли 167 995 (14 %) записей с целевым событием и 1 070 762 (86 %) записи без целевого события. При сравнении записей с и без целевого события в этом наборе не было выявлено статистически значимой разницы в распределении только одного признака — ИМТ (онлайн-прил. 1). Записи пациентов с целевым событием характеризовались более высокими значениями возраста и СОЭ, среди них было больше мужчин, чаще встречались пониженный уровень гемоглобина крови и тенденция к снижению массы тела.

Сформированные наборы для проведения внешней валидации и разработки моделей характеризовались схожим балансом (рис. 4). Однако в этих выборках была выявлены значительные отличия в заполненности количественных признаков (рис. 3, б) и распределении их значений (онлайн-приложение 1), что было отмечено нами как положительный аспект для анализа работы моделей на данных, отличающихся от данных при обучении.

Результаты внутреннего тестирования и внешней валидации

По итогам внутреннего тестирования, ни одна модель на основе Random Forest не показала стабильности с включением нуля в доверительный интервал разницы AUROC, полученной при обучении и внутреннем тестировании. Из

Таблица 1. Параметры наилучшей предобработки данных для соответствующих алгоритмов согласно результатам проведения внутреннего тестирования

Модель	Заполнение пропусков	Масштабирование	Коррекция дисбаланса
LogisticRegression	Фиксированное значение (–10 000)	Стандартизация	Не проводилась
GaussianNB	Фиксированное значение (–10 000)	Стандартизация	SMOTE
Linear Discriminant Analysis	Фиксированное значение (–10 000)	Стандартизация	ROS
LGBMClassifier	Фиксированное значение (–10 000)	Не проводилось	ROS

Table 1. Optimal data preprocessing parameters for the algorithms based on testing results

Model	Missing Value Imputation	Scaling	Imbalance Correction
LogisticRegression	Fixed value (–10,000)	Standardization	Not performed
GaussianNB	Fixed value (–10,000)	Standardization	SMOTE
Linear Discriminant Analysis	Fixed value (–10,000)	Standardization	ROS
LGBMClassifier	Fixed value (–10,000)	Not performed	ROS

Таблица 2. Значения AUROC [95 % ДИ] на наборах для обучения, внутреннего тестирования и внешней валидации для отобранных моделей. Выполнена сортировка по значению метрики при проведении внешней валидации

Модель	Набор для обучения	Набор для внутреннего тестирования	Внешние данные № 1	Внешние данные № 2
Logistic Regression	0,756 [0,747, 0,765]	0,758 [0,749, 0,767]	0,745 [0,738, 0,753]	0,731 [0,724, 0,739]
GaussianNB	0,765 [0,756, 0,775]	0,768 [0,76, 0,777]	0,769 [0,761, 0,777]	0,758 [0,749, 0,766]
Linear Discriminant Analysis	0,765 [0,757, 0,773]	0,768 [0,759, 0,776]	0,766 [0,757, 0,773]	0,756 [0,749, 0,764]
LGBM Classifier	0,808 [0,800, 0,816]	0,807 [0,798, 0,815]	0,794 [0,786, 0,8]	0,790 [0,782, 0,798]

Table 2. AUROC [95 % CI] for the results of training, evaluation, and validation of the selected models. The results were ranked according to the performance metric during external validation

Model	Training set	Internal test set	External data No 1	External data No 2
Logistic Regression	0.756 [0.747, 0.765]	0.758 [0.749, 0.767]	0.745 [0.738, 0.753]	0.731 [0.724, 0.739]
GaussianNB	0.765 [0.756, 0.775]	0.768 [0.76, 0.777]	0.769 [0.761, 0.777]	0.758 [0.749, 0.766]
Linear Discriminant Analysis	0.765 [0.757, 0.773]	0.768 [0.759, 0.776]	0.766 [0.757, 0.773]	0.756 [0.749, 0.764]
LGBM Classifier	0.808 [0.800, 0.816]	0.807 [0.798, 0.815]	0.794 [0.786, 0.8]	0.790 [0.782, 0.798]

моделей на основании оставшихся архитектур до внешней валидации были допущены наилучшие, показавшие наибольшее значение AUROC при внутреннем тестировании. Подходы к заполнению пропусков, масштабированию и коррекции дисбаланса, при которых были получены эти модели, представлены в табл. 1.

Значения AUROC, полученные на наборах для внешней валидации, внутреннего тестирования и обучения для отобранных моделей, представлены в табл. 2. Модель на основе LGBMClassifier продемонстрировала лучшую дискриминативную способность, а также стабильность на внешних данных. Значение метрики AUROC составило 0,807 (95 % ДИ 0,798–0,815) при внутреннем тестировании, 0,794 (95 % ДИ 0,786–0,800) при использовании внешнего набора данных № 1 и 0,790 (95 % ДИ 0,782–0,798) — для внешнего набора данных № 2.

Для разделения записей на три группы риска на наборе данных для тестирования нами дополнительно рассчитаны два порога активации в зависимости от целевых ПЦОР и ПЦПР. ROC-кривая модели LGBMClassifier при проведении внутреннего тестирования с указанием трех порогов представлена на рис. 5. На данных для внутреннего тестирования точность модели с порогом классификации 0,021, при котором достигнуто целевое значение ПЦОР на внутреннем тестировании (0,99), составила 0,334 (95 % ДИ 0,329–0,339). Чувствительность при этом составила 0,985 (95 % ДИ 0,980–0,990), а специфичность — 0,232 (95 % ДИ 0,226–0,238). При использовании второго порога (0,391) с ожидаемой ПЦПР при тестировании, равной 0,5, метрики качества были следующими: точность — 0,865 (95 % ДИ 0,861–0,868), чувствительность — 0,293 (95 % ДИ 0,276–0,310), специфичность — 0,954 (95 % ДИ 0,951–0,957).

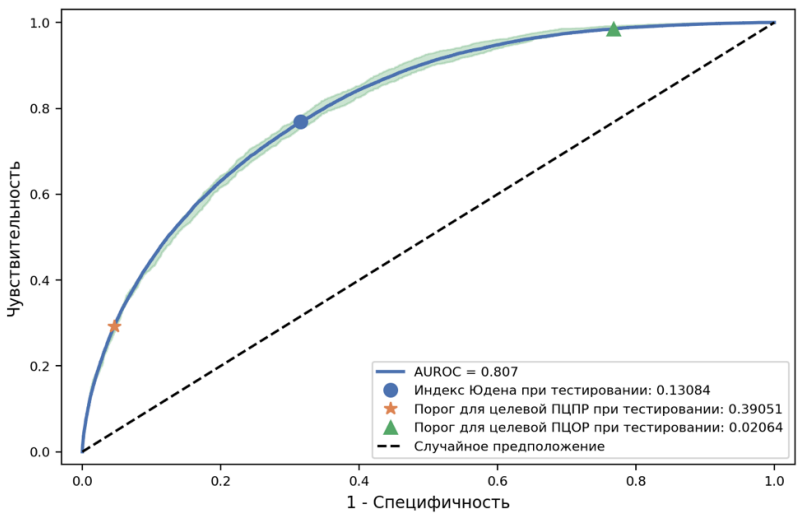

Рис. 5. ROC-кривая с 95 %-ным ДИ для модели LGBMClassifier с порогами на данных для внутреннего тестирования
Fig. 5. ROC curve with 95% confidence intervals for the LightGBM classifier model, showing decision thresholds on internal validation data.

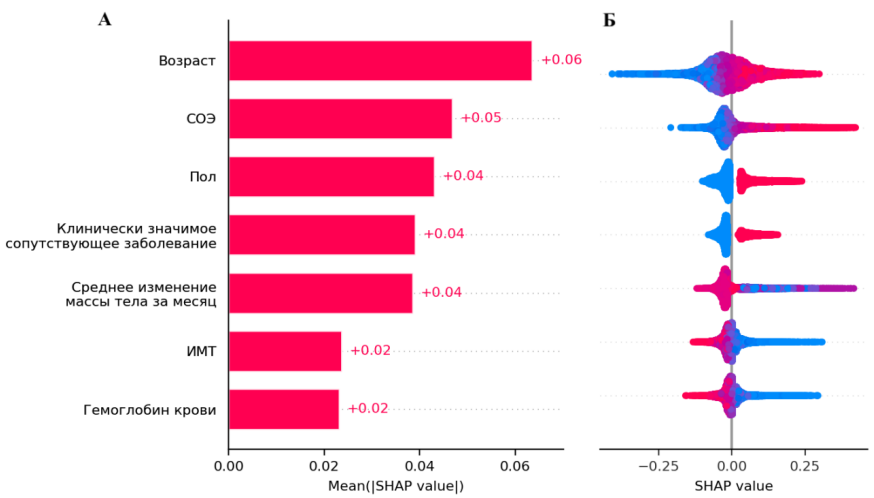

Рис. 6. А — средняя значимость признаков для модели LGBMClassifier. Б — распределение абсолютных значений значимости признаков по всем записям. Для признака «пол» красным цветом отмечены пациенты мужского пола, синим — женского, для признака «клинически значимое сопутствующее заболевание» — пациенты с соответствующими заболеваниями в анамнезе и без соответственно
Fig. 6. А — Mean feature importance scores for the LightGBM classifier model. Б — Distribution of absolute feature importance values across all observations. For the “Gender” feature, male and female patients are denoted by red and blue markers respectively. For “Clinically significant comorbidity”, red indicates patients with relevant medical history, while blue indicates those without

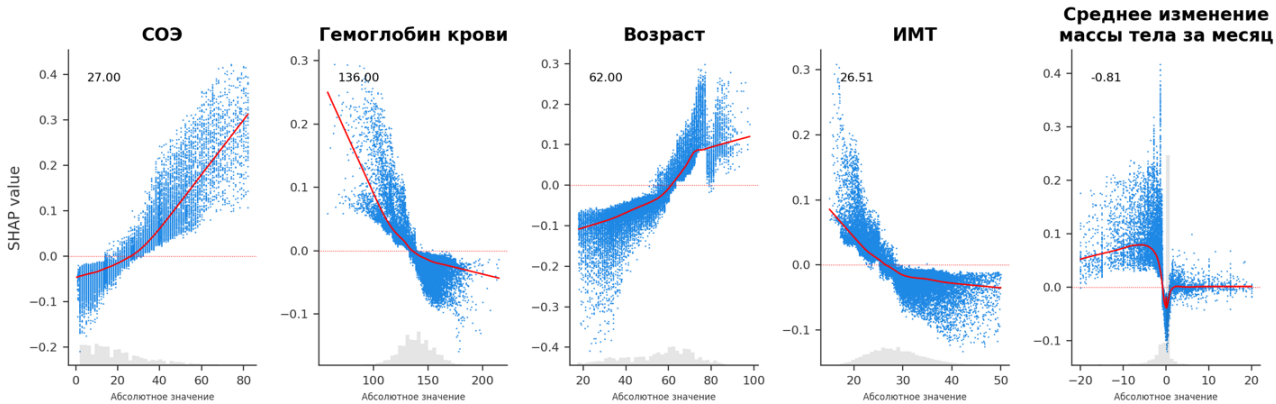

Рис. 7. Диаграммы рассеивания абсолютных значений количественных признаков и соответствующих им чисел Шепли
Fig. 7. Scatter plots showing how absolute values of quantitative features correlate with their corresponding Shapley values

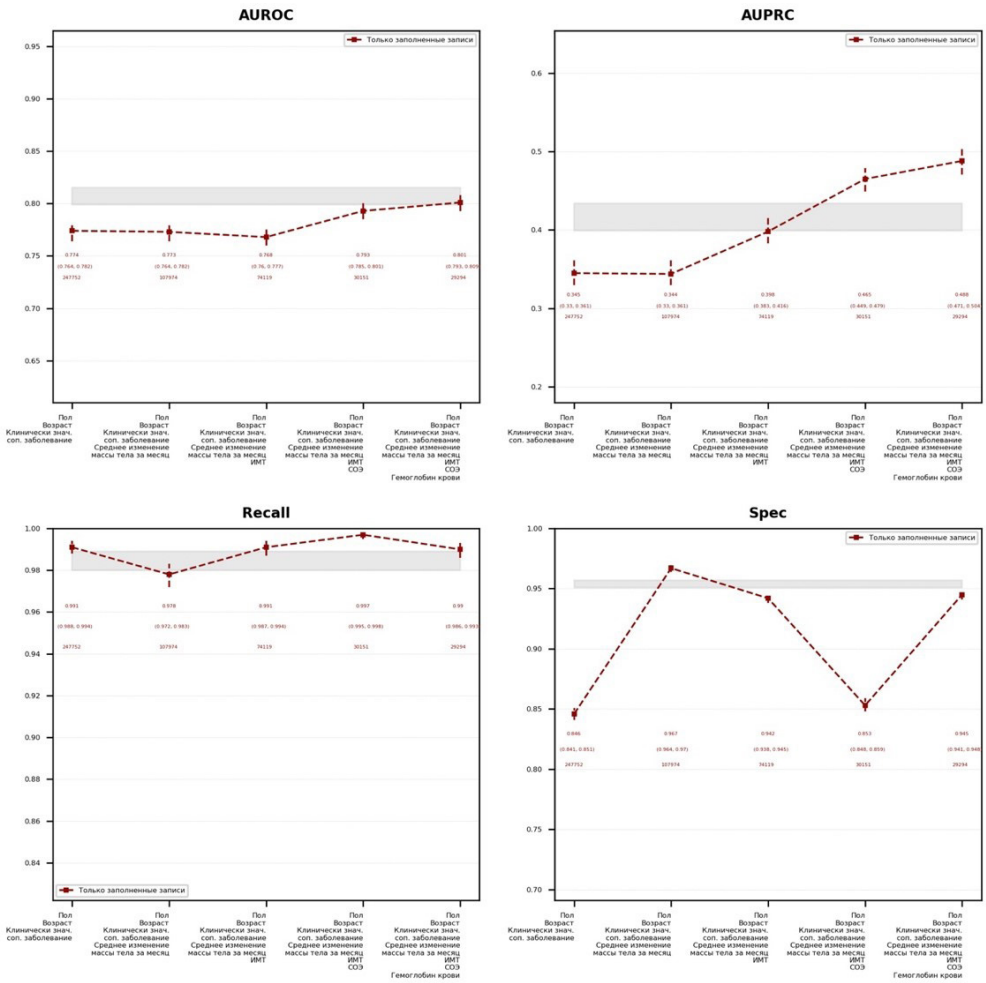

Рис. 8. Динамика метрик (95 % ДИ) модели LGBMClassifier при использовании выборок с различной заполненностью признаков. Размеры соответствующих выборок указаны под ДИ. Серая полоса на каждом графике иллюстрирует доверительный интервал для значений соответствующей метрики, полученный при анализе работы модели на всем наборе данных для внутреннего тестирования без выделения только заполненных записей
Fig. 8. Performance metric trends (95% CI) for the LGBMClassifier across datasets with varying feature completeness. Sample sizes are annotated below each confidence interval. Gray bands represent baseline metrics from the complete internal validation dataset (including records with missing values)

Полная характеристика эффективности итоговой модели после применения упомянутых порогов классификации, а также максимума индекса Юдена, рассчитанных на тестовом наборе, представлена в онлайн-прил. 4. Средняя абсолютная ошибка калибровочной кривой равнялась 4,34 % при показателе Brier наборе данных для внутреннего тестирования, равном 0,097.

Исследование значимости признаков

Значимость признаков для итоговой модели, определенная по методу Шепли, показана на рис. 6. Значимость соответствует клинической

практике: разработанный нами инструмент считает, что такие характеристики пациента, как пожилой возраст, мужской пол, наличие клинически значимого сопутствующего заболевания в анамнезе, сниженные уровни ИМТ и гемоглобина крови, а также повышенный уровень СОЭ и тенденция к снижению массы тела значительно повышают риск онкологического заболевания в течение последующих 18 мес.

Для определения пороговых значений количественных признаков, которые модель считает критичными для определения риска, мы провели дополнительный анализ их распределения. Были построены диаграммы рассеивания абсолютных значений количественных признаков в зависимости от соответствующих им чисел Шепли (рис. 7). Было отмечено, что повышение СОЭ выше 27 мм/ч, возраст старше 62 лет, снижение уровня гемоглобина крови ниже 136 г/л и ИМТ ниже 26,5 кг/м^2, а также тенденция к потере массы тела более чем 800 г в месяц ведут к повышению риска.

Результаты интерпретации значений признаков на примере пациентов из набора данных для внутреннего тестирования с самым высоким и самым низким риском диагностирования онкологического заболевания в течение 18 мес. представлены в онлайн-прил. 5.

Исследование применимости модели в условиях различной заполненности данных

Использование модели на сторонних данных предполагает определение требований к ним в виде списка обязательно присутствующих признаков. Для разработки такого списка были рассчитаны метрики качества итоговой модели на специально сформированных выборках из набора данных для внутреннего тестирования. Для каждой точки на рис. 8 использовались те записи, в которых были заполнены значения признаков из списка, соответствующего этой точке по оси абсцисс. Значения оставшихся предикторов, использовавшихся в нашем исследовании и не входящих в соответствующий список, заменялись пропусками. Таким образом имитировалась ситуация подачи на модель лишь ограниченно заполненных данных.

Формирование первой такой искусственной выборки было проведено с использованием в качестве обязательных признаков пол и возраст, так как они были заполнены во всех записях, и клинически значимое сопутствующее заболевание. Так, в случае пропусков в нем подразумевалось отсутствие соответствующих заболеваний в анамнезе пациента (онлайн-прил. 3), что заполнялось нулями. Порядок добавления следующих признаков к списку обязательных определялся их значимостью согласно числам Шепли. По мере увеличения числа обязательных признаков, выборка, соответствующая новым, более строгим критериям отбора, уменьшалась, о чем свидетельствуют указанные под каждой точкой объемы данных.

Тенденции изменения метрик на сформированных выборках представлены на рис. 8. AUROC и чувствительность (Recall, при использовании порога для целевой ПЦОР) сохраняли стабильность вне зависимости от количества подаваемых признаков. В то же время AUPRC показала нарастающий тренд с увеличением объема информации о пациенте; при этом с повышением ее значения в максимальной точке на 17 % по сравнению с изначальным набором данных для внутреннего тестирования, обладавшего «естественной» заполненностью. Специфичность (Spec, при использовании порога для целевой ПЦПР) не продемонстрировала четкой динамики и варьировалась от 0,846 до 0,967 в зависимости от исследуемого набора.

Обсуждение

Методы МО хорошо зарекомендовали себя в разработке прогнозных инструментов для определения развития и исходов различных многофакторных заболеваний. При этом в качестве предикторов могут используются различные клинические, демографические, лабораторные и инструментальные параметры, которые можно отслеживать и контролировать. Уклон большинства современных работ обращен в сторону прогнозирования риска и диагностирования отдельных подтипов ЗНО, в то время как целью нашего исследования было создание универсального инструмента скрининга, позволяющего в дальнейшем значительно корректировать тактику ведения отдельного пациента в зависимости от его персонального риска диагностирования любого злокачественного новообразования в будущем.

Значение AUROC при проведении тестирования существующих аналогичных инструментов прогнозирования находилось в интервале 0,684–0,722 [4]. Это свидетельствует о том, что полученные нами результаты внутреннего тестирования и внешней валидации модели на основе LGBMClassifier по уровню целевой метрики превосходят разработки зарубежных авторов. С использованием двух порогов активации также было предложено разделение пациентов на три группы риска в зависимости от выхода модели. При бинаризации на уровне первого порога (0,021), установленного для достижения ожидаемой ПЦОР (0,99), модель демонстрирует высокую чувствительность. Это позволяет надежно идентифицировать пациентов с низким риском, которые не нуждаются в активном диагностическом поиске и мониторинге. Использование вто-

рого порога (0,391) характеризовалось высоким значением специфичности и ПЦПР (0,5). Он позволяет с уверенностью определить высокий риск развития онкологического заболевания у пациента, что указывает на необходимость проведения для него дополнительных исследований.

Использование разработок на основе МО в реальной клинической практике предполагает возможность их работы с данными, значительно отличающимися от данных для обучения. Результаты нашей модели при валидации показали ее стабильность в подобных ситуациях, однако, несмотря на использование термина «внешние данные», эти выборки для внешней валидации также были получены из нашей базы данных, хоть и отобраны на основании рекомендаций TRIPOD. Проведенный эксперимент с формированием подвыборок с искусственной заполненностью показал относительную стабильность метрик AUROC и чувствительности (с порогом 0,021) вне зависимости от количества подаваемых на модель признаков. Однако при исследовании AUPRC отмечена выраженная динамика, заключающаяся в повышении способности модели выделять пациентов с высоким риском развития целевого события с увеличением объема информации о нем на момент прогноза. Динамика специфичности при пороге 0,391 не имела определенного тренда и, вероятнее всего, была связана с самим значением порога для максимизации ПЦПР на наборе данных для внутреннего тестирования, обладавшего «естественной» (в соответствии с реальной клинической практикой) заполненностью, соответствующей нашей базе данных. Несмотря на факт, что наименьшее полученное значение специфичности было 0,846, данная находка открывает перспективу для новых исследований, в которых стоит изучить возможность определения различных порогов бинаризации для одной модели в зависимости от заполненности входных данных.

Важным выводом является то, что половина пациентов, у которых развилось онкологическое заболевание, оно было диагностировано в течение двух месяцев с момента прогноза. Это предполагает, что значительная часть этих случаев, вероятно, была выявлена на поздних стадиях. К сожалению, отсутствие в наборе данных такой информации помешало нам провести детальную оценку того, как стадийность может повлиять на результаты работы модели.

Проведенный анализ чисел Шепли для разработанной модели показал высокую интерпретируемость ее работы. Результаты исследования значимости отдельных количественных признаков, а также сравнение определенных у них критически важных значений для прогноза со значениями, использующимися в реальной клинической практике (табл. 3), показали, что методы МО обладают яркой способностью к извлечению зависимостей из медицинских данных, совпадающих с медицинской логикой, и могут быть использованы для определения новых, ранее не определенных порогов при анализе других лабораторных и инструментальных предикторов.

Таблица 3. Определение критических значений для изменения общего риска развития онкологического заболевания согласно медицинской логике и логике МО

Основа	Критическое значение	
	Медицинские знания	LGBM Classifier
Среднее изменение массы тела за месяц, кг/мес.	Потеря более 5 кг за 6 мес. или 10 кг за год ~= –0.8	≤ –0,8
Гемоглобин крови, г/л	< 120 (для мужчин) < 115 (для женщин)	< 136
СОЭ, мм/ч	>50 (старше 50 лет) >15 (для мужчин младше 50 лет) >20 (для женщин младше 50 лет)	>27
ИМТ, кг/м^2	< 25	< 26,5

Table 3. Definition of critical values for shifts in the overall cancer risk according to medical logic and ML logic

Feature	Critical values	
	Clinical data	LGBM Classifier
Mean monthly weight change, kg/month	Loss of more than 5 kg in 6 months or 10 kg in a year ~= <= –0.8	<= –0.8
Blood hemoglobin level, g/L	< 120 (for men) < 115 (for women)	< 136
ESR, mm/h	> 50 (for individuals over 50 years old) > 15 (for men under 50 years old) > 20 (for women under 50 years old)	> 27
BMI, kg/m^2	< 25	< 26.5

Все упомянутые факты подчеркивают значительный потенциал разработанной модели и МО в целом как дополнительного инструмента для скрининга пациентов на предмет оценки риска развития ЗНО. Учитывая ее устойчивость, продемонстрированную во внешней валидации, модель может быть использована в практических проспективных исследованиях. Однако одной из ключевых задач в прогнозировании является определение оптимального порядка последующих обследований для диагностики конкретных онкологических заболеваний. Это пилотное исследование направлено на оценку применимости методов искусственного интеллекта для решения онкологических задач. Следующим этапом станет разработка модели машинного обучения на основе многометочной классификации. Такая модель может обеспечить более точную оценку риска отдельных видов ЗНО и улучшит персонализированные стратегии по тактике ведения.

Заключение

В результате исследования была успешно разработана и валидирована модель МО на основе LGBMClassifier для прогнозирования развития ЗНО в течение 18 мес., метрики которой соответствовали результатам инструментов, представленных в ранее опубликованных исследованиях. Результаты внешней валидации подтвердили, что модель сохраняет свою стабильность при обработке данных из других регионов и временных периодов, что, в сочетании с ее качественными характеристиками, свидетельствует о возможности ее применения в реальной клинической практике. Простота интерпретируемости прогноза, а также подход с выделением трех групп риска значительно увеличивают ценность нашей разработки, позволяя формировать группы пациентов с реально высоким или низким риском диагностирования онкологического заболевания в течение следующих 18 мес. с даты прогноза.

Конфликт интересов
Авторы заявляют об отсутствии конфликта интересов.
Conflict of interest
The authors declare no conflict of interest.

Соблюдение прав пациентов и правил биоэтики
Для сбора данных компанией-разработчиком платформы Webiomed были подписаны соглашения с соответствующими операторами персональных медицинских данных на их обезличивание на стороне оператора и передачу результатов в платформу Webiomed, в том числе для научно-исследовательских целей. Поскольку анализировались обезличенные медицинские данные, информированное добровольное согласие не использовалось.
Compliance with patient rights and principles of bioethics
The Webiomed platform developer signed agreements with medical data operators to anonymize the data on their side and share it with Webiomed for research purposes. Since the data was anonymized, informed consent was not required.

Источник финансирования
Исследование не имело спонсорской поддержки.
Funding
The work was performed without external funding.

Участие авторов
Авторы декларируют соответствие своего авторства международным критериям ICMJE.
Ермак А.Д. — обработка материала, анализ и интерпретация данных, написание текста статьи;
Гаврилов Д.В. — обзор публикаций по теме статьи, анализ и интерпретация данных, научное редактирование;
Новицкий Р.Э. — разработка дизайна исследования, сбор материала исследования;
Гусев А.В. — разработка дизайна исследования, научное редактирование;
Комаров Ю.И. — научное редактирование;
Андрейченко А.Е. — анализ и интерпретация данных, научное редактирование.
Все авторы одобрили финальную версию статьи перед публикацией, выразил(и) согласие нести ответственность за все аспекты работы, подразумевающую надлежащее изучение и решение вопросов, связанных с точностью или добросовестностью любой части работы.
Authors' contributions
The authors declare the compliance of their authorship according to the international ICMJE criteria.
Ermak A.D. — data curation, data analysis and interpretation, original draft preparation;
Gavrilov D.V. — literature review, data analysis and interpretation, review and editing;
Novitskiy R.E. — study conceptualization, data collection;
Gusev A.V. — study conceptualization, reviewing and editing;
Komarov Yu.I. — reviewing and editing;
Andreychenko A.E. — data analysis and interpretation, reviewing and editing.
All authors have approved the final version of the article before publication, agreed to assume responsibility for all aspects of the work, implying proper review and resolution of issues related to the accuracy or integrity of any part of the work.

ЛИТЕРАТУРА / REFERENCES

1. Usher-Smith J., Emery J., Hamilton W., et al. Risk prediction tools for cancer in primary care. *British Journal of Cancer*. 2015; 113(12): 1645–50.-DOI: 10.1038/bjc.2015.409.
2. Chiang P.C., Glance D., Walker J., et al. Implementing a qcancer risk tool into general practice consultations: An exploratory study using simulated consultations with Australian general practitioners. *Br J Cancer*. 2015; 112(S1): S77–83.-DOI: 10.1038/bjc.2015.46.
3. Hippisley-Cox J., Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: Prospective cohort study. *BMJ Open*. 2015; 5(3): e007825.-DOI: 10.1136/bmjopen-2015-007825.
4. Kulm S., Kofman L., Mezey J., Elemento O. Simple linear cancer risk prediction models with novel features outperform complex approaches**.** *JCO Clin Cancer Inform.* 2022; (6).-DOI: 10.1200/CCI.21.00166.
5. Miotto R., Li L., Kidd B.A., Dudley J.T. Deep patient: an unsupervised representation to predict the future of patients

from the electronic health records. *Sci Rep.* 2016; (6).-DOI: 10.1038/srep26094.

6. Watson J., Salisbury C., Banks J., et al. Predictive value of inflammatory markers for cancer diagnosis in primary care: a prospective cohort study using electronic health records. *Br J Cancer.* 2019; 120(11): 1045–51.-DOI: 10.1038/s41416-019-0458-x.
7. Nicholson B.D., Hamilton W., Sullivan J.O', et al. Weight loss as a predictor of cancer in primary care: A systematic review and meta-analysis**.** 2018; 68(670): e311–22, *Royal College of General Practitioners*.-DOI: 10.3399/bjgp18X695801.
8. Hung N., et al. Risk of cancer in patients with iron deficiency anemia: A nationwide population-based study. *PLoS One*. 2015; 10(3): e0119647.-DOI: 10.1371/journal.pone.0119647.
9. Star J., et al. Updated review of major cancer risk factors and screening test use in the United States, with a focus on changes during the COVID-19 pandemic. 2023; 32(7): 879–88. *American Association for Cancer Research Inc*.-DOI: 10.1158/1055-9965.EPI-23-0114.
10. Collins G.S., Reitsma J.B., Altman D.G., Moons K.G.M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med.* 2015; 13(1): 1.-DOI: 10.1186/s12916-014-0241-z.
11. Collins G.S., Moons K.G.M., Dhiman P., et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024; e078378.-DOI: 10.1136/bmj-2023-078378.
12. Kapoor S., Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023; 4(9): 100804.-DOI: 10.1016/j.patter.2023.100804.
13. Li C. Little′s test of missing completely at random. *The Stata Journal*. 2013; 13(4): 795–809.-DOI: 10.1177/1536867X1301300407.
14. Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag*. 2009; 45(4): 427–37.-DOI: 10.1016/j.ipm.2009.03.002.
15. Zoubir A., Iskandler D. Bootstrap methods and applications. *IEEE Signal Processing Magazine*. 2007; 24(4): 10–9.-DOI: 10.1109/msp.2007.4286560.
16. Fischer G., Evans A.T. SpPin and SnNout are not enough. It's Time to fully embrace likelihood ratios and probabilistic reasoning to achieve diagnostic excellence. *J Gen Intern Med*. 2023; 38(9): 2202–4.-DOI: 10.1007/s11606-023-08177-5.
17. Lundberg S.M., Erion G., Chen H., et al. From local explanations to global understanding with explainable AI for trees**.** *Nat Mach Intell.* 2020; 2(1): 56–67.-DOI: 10.1038/s42256-019-0138-9.
18. Van Calster B., McLernon D.J., van Smeden M., et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019; 17(1).-DOI: 10.1186/s12916-019-1466-7.
19. Ding Y., Simonoff J. An investigation of missing data methods for classification trees applied to binary response data. *J Mach Learn Res*. 2010; 11: 131-70.-DOI: 10.5555/1756006.1756012.
20. Cao X.H., Stojkovic I., Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics*. 2016; 17(1): 359.-DOI: 10.1186/s12859-016-1236-x.
21. de Amorim L.B.V., Cavalcanti G.D.C., Cruz R.M.O. The choice of scaling technique matters for classification performance. *Appl Soft Comput*. 2023; 133: 109924.-DOI: 10.1016/j.asoc.2022.109924.
22. Weiss G.M. Foundations of imbalanced learning. In: He H., Ma Y., eds. Imbalanced learning: foundations, algorithms, and applications. Hoboken (NJ): John Wiley & Sons. 2013; 13-41.-ISBN: 9781118074626.
23. Ke G., Meng Q., Finley T., et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst.* 2017.
24. Breiman L. Random forests. *Mach Learn*. 2001; 45(1): 5-32.-DOI: 10.1023/A:1010933404324.
25. Akiba T., Sano S., Yanase T., et al. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019; 2623-31.-DOI: 10.1145/3292500.3330701.
26. Wester D.B. Comparing treatment means: overlapping standard errors, overlapping confidence intervals, and tests of hypothesis. *Biom Biostat Int J*. 2018; 7(1): 73-85.-DOI: 10.15406/bbij.2018.07.00192.

Поступила в редакцию / Received / 22.01.2025
Прошла рецензирование / Reviewed / 19.03.2025
Принята к печати / Accepted for publication / 20.03.2025

Сведения об авторах / Author Information / ORCID

Андрей Дмитриевич Ермак / Andrey D. Ermak / ORCID: https://orcid.org/0000-0002-0513-8557.
Денис Владимирович Гаврилов / Denis V. Gavrilov / ORCID: https://orcid.org/0000-0002-8745-857X.
Роман Эдвардович Новицкий / Roman E. Novitskiy / ORCID: https://orcid.org/0000-0002-2350-977X.
Александр Владимирович Гусев / Alexander V. Gusev / ORCID: https://orcid.org/0000-0002-7380-8460.
Юрий Игоревич Комаров / Yuriy I. Komarov / ORCID: https://orcid.org/0000-0003-3256-0451.
Анна Евгеньевна Андрейченко / Anna E. Andreychenko / ORCID: https://orcid.org/0000-0001-6359-0763.