

РАЗРАБОТКА И ВАЛИДАЦИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ, ПРОГНОЗИРУЮЩИХ РИСК ГОСПИТАЛИЗАЦИИ ПАЦИЕНТОВ С САХАРНЫМ ДИАБЕТОМ В ТЕЧЕНИЕ ПОСЛЕДУЮЩИХ 12 МЕСЯЦЕВ



© А.Е. Андрейченко^{1*}, А.Д. Ермак¹, Д.В. Гаврилов¹, Р.Э. Новицкий¹, А.В. Гусев^{2,3}

¹ООО «К-Скай», Петрозаводск

²Центральный научно-исследовательский институт организации и информатизации здравоохранения, Москва

³Научно-практический клинический центр диагностики и телемедицинских технологий, Москва

ОБОСНОВАНИЕ. Заболеваемость сахарным диабетом (СД) как в Российской Федерации, так и во всем мире неуклонно растет последние десятилетия. Стабильный популяционный рост и современные эпидемиологические характеристики СД приводят к колоссальным экономическим расходам и значительному социальному ущербу во всем мире. Заболевание зачастую приобретает прогрессирующее течение с развитием характерных осложнений, при этом значительно повышая вероятность госпитализации. Создание и внедрение модели прогнозирования госпитализаций пациентов с СД в круглосуточный стационар позволит персонифицировать оказание медицинской помощи и оптимизировать нагрузку на всю систему здравоохранения.

ЦЕЛЬ. Разработка и валидация моделей прогнозирования госпитализаций пациентов с СД по поводу самого заболевания и его осложнений с использованием алгоритмов машинного обучения и данных реальной клинической практики.

МАТЕРИАЛЫ И МЕТОДЫ. По сведениям из деперсонифицированных электронных медицинских карт, полученных из платформы Webiomed, была проанализирована 170 141 запись 23 742 пациентов с СД. В качестве потенциальных факторов прогноза отобраны анамнестические, конституциональные, клинические, инструментальные и лабораторные данные, широко используемые в рутинной врачебной практике — всего 33 признака. Для создания моделей применялась логистическая регрессия (LR), методы градиентного бустинга (LightGBM, XGBoost, CatBoost), методы, основанные на деревьях решений (RandomForest и ExtraTrees), а также алгоритм на основе нейронных сетей (Multi-layer Perceptron).

РЕЗУЛЬТАТЫ. Наилучшие результаты и устойчивость к внешним данным показала модель LightGBM со значением целевой метрики AUC 0.818 (95% ДИ 0,802–0,834) при внутреннем тестировании и 0,802 (95% ДИ 0,773–0,832) при внешней валидации.

ЗАКЛЮЧЕНИЕ. Метрики полученной итоговой модели превосходили опубликованные ранее аналоги. Результаты внешней валидации показали относительную устойчивость модели к новым данным из другого региона, что в совокупности с показателями качества отражает возможность ее использования в реальной клинической практике.

КЛЮЧЕВЫЕ СЛОВА: сахарный диабет; госпитализация; прогнозные модели; машинное обучение; искусственный интеллект.

DEVELOPMENT AND VALIDATION OF MACHINE LEARNING MODELS TO PREDICT UNPLANNED HOSPITALIZATIONS OF PATIENTS WITH DIABETES WITHIN THE NEXT 12 MONTHS

© Anna E. Andreychenko^{1*}, Andrey D. Ermak¹, Denis V. Gavrilov¹, Roman E. Novitskiy¹, Alexander V. Gusev^{2,3}

¹K-SkAI LLC; Petrozavodsk, Russia

²Federal Research Institute for Health Organization and Informatics, Moscow, Russia

³Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia

BACKGROUND: The incidence of diabetes mellitus (DM) both in the Russian Federation and in the world has been steadily increasing for several decades. Stable population growth and current epidemiological characteristics of DM lead to enormous economic costs and significant social losses throughout the world. The disease often progresses with the development of specific complications, while significantly increasing the likelihood of hospitalization. The creation and inference of a machine learning model for predicting hospitalizations of patients with DM to an inpatient medical facility will make it possible to personalize the provision of medical care and optimize the load on the entire healthcare system.

AIM: Development and validation of models for predicting unplanned hospitalizations of patients with diabetes due to the disease itself and its complications using machine learning algorithms and data from real clinical practice.

MATERIALS AND METHODS: 170,141 depersonalized electronic health records of 23,742 diabetic patients were included in the study. Anamnestic, constitutional, clinical, instrumental and laboratory data, widely used in routine medical practice, were considered as potential predictors, a total of 33 signs. Logistic regression (LR), gradient boosting methods (LightGBM, XGBoost, CatBoost), decision tree-based methods (RandomForest and ExtraTrees), and a neural network-based algorithm (Multi-layer Perceptron) were compared. External validation was performed on the data of the separate region of Russian Federation.



RESULTS: The best results and stability to external validation data were shown by the LightGBM model with an AUC of 0.818 (95% CI 0.802–0.834) in internal testing and 0.802 (95% CI 0.773–0.832) in external validation.

CONCLUSION: The metrics of the best model were superior to previously published studies. The results of external validation showed the relative stability of the model to new data from another region, that reflects the possibility of the model's application in real clinical practice.

KEYWORDS: *diabetes mellitus; hospitalization; predictive models; machine learning; artificial intelligence.*

ОБОСНОВАНИЕ

Сахарный диабет (СД) — это распространенное хроническое заболевание, развивающееся в результате неэффективной выработки инсулина клетками поджелудочной железы или снижения чувствительности периферических тканей к его биологическому действию [1]. Само хроническое течение СД оказывает значительное влияние на качество жизни пациентов, требует оказания специализированной медицинской помощи с формированием навыков по регулярному самоконтролю гликемии и приему назначенных препаратов.

В настоящее время в отечественной эндокринологической практике используется классификация СД Всемирной организации здравоохранения 1999 г. с выделением четырех основных групп: СД 1 и 2 типов, гестационный СД и другие специфические типы СД [2]. Несмотря на отличительные механизмы развития самого заболевания, у всех этих состояний в конечном счете можно выделить общность патогенеза осложнений, связанных с хронической гипергликемией [1]. Потому заболевание приобретает прогрессирующее течение с развитием характерных осложнений, значительно повышая вероятность госпитализации пациентов с СД.

Распространенность СД как в Российской Федерации, так и во всем мире неуклонно растет последние десятилетия [3]. Согласно отчету Международной диабетической федерации, ожидается, что число пациентов, страдающих этим заболеванием, увеличится к 2045 г. до 700 млн, что составит 10,9% всего взрослого населения [4]. Стабильный популяционный рост и современные эпидемиологические характеристики СД приводят к колоссальным экономическим расходам и значительному социальному ущербу во всем мире. К примеру, национальные данные Южной Кореи свидетельствуют о 17-кратном увеличении ежегодной стоимости оказания помощи пациентам с СД в 2019 г. в сравнении с 2002-м [5]. Согласно проведенному исследованию Bommer и соавт., в 2015 г. 1,8% мирового ВВП приходилось на лечение и профилактику СД, что составило 1,31 триллиона долларов [6]. При этом на примере австралийской системы здравоохранения было показано, что более 40% этих средств выделяется на оказание специализированной помощи в круглосуточных стационарах [7]. В этой связи создание и внедрение многофакторной персонифицированной модели прогнозирования госпитализаций пациентов с СД в круглосуточный стационар по поводу самого заболевания и его осложнений является актуальной задачей, решение которой позволит значительно улучшить качество оказания медицинской помощи и снизить нагрузку на всю систему здравоохранения за счет тщательного наблюдения за пациентами из группы риска, раннего превентивного вмешательства в течение

их заболевания, а также рационального планирования штатной численности и коечного фонда медицинских организаций.

В научной литературе опубликованы результаты нескольких исследований, связанных с разработкой предиктивных моделей госпитализаций пациентов с СД. Большое внимание авторов было обращено к проблеме повторных госпитализаций при СД 2 типа (СД2). Так, в публикации Tsai-Chung Li и соавт. [8] был описан процесс создания модели на основе регрессии Кокса для предсказания таких госпитализаций в течение 1, 3, 5 и 8 лет с использованием для обучения 21 769 записей пациентов. По результатам проведенного внутреннего тестирования (10 884 записи), целевая метрика AUC составила 0,80, 0,77, 0,76 и 0,74 соответственно. Brisimi и соавт. для решения той же задачи использовали 26 478 записей пациентов, разделенных в отношении 60/40 на обучающий набор и НД для внутреннего тестирования, причем при применении на последней финальной модели на основе Gradient Tree Boosting метрика AUC составила 0,7272 [9]. Отличительными особенностями публикации Naohui Lu и соавт. являются исследование госпитализаций пациентов вне зависимости от типа СД и большой размер исследуемой выборки — 97 108 пациентов [10]. Однако полученная по итогу обучения ансамблевая модель на основе Random Forest, AdaBoost, XGBoost, and CatBoost показала низкую точность при использовании кросс-валидации (AUC 0.6736). Активное повсеместное использование нейросетей не оставило без внимания и современную эндокринологию: ряд публикаций был посвящен использованию различных нейросетевых архитектур в решении задачи прогнозирования повторных госпитализаций пациентов с СД. Так, в работе Ameen A. Hai и соавт. [11] использование нейросети LSTM показало лучшие результаты в сравнении с классическими методами машинного обучения (МО) (RandomForest, Logistic Regression и AdaBoost) со значением AUC 0.79. В то же время, согласно результатам Bhuvan и соавт. [12], в их работе RandomForest показал сопоставимое с нейронной сетью значение целевой метрики качества — AUC 0,65 и 0,654 соответственно. Поскольку качество и эффективность разработанных на сегодняшний день моделей прогнозирования варьируется в широких пределах, а возможность их практического применения сомнительна ввиду отсутствия информации о результатах валидации на внешних данных, вопрос исследований в данной области остается актуальным.

ЦЕЛЬ ИССЛЕДОВАНИЯ

Разработка и валидация моделей прогнозирования госпитализаций пациентов с СД по поводу самого заболевания и его осложнений с использованием алгоритмов МО и данных реальной клинической практики.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования

Проведено многоцентровое ретроспективное обсервационное исследование.

Место и время проведения исследования

В качестве источника информации использовалась база данных платформы прогнозной аналитики Webiomed, содержащая обезличенные формализованные данные электронных медицинских карт (ЭМК) 11,6 миллиона пациентов из 25 регионов Российской Федерации. Медицинские данные и машиночитаемые значения признаков собирались из ЭМК с помощью различных технологий извлечения информации из неструктурированных либо слабоструктурированных текстовых данных.

Исследуемые популяции

На основе имеющейся базы данных был сформирован НД, включающий информацию о 170 141 записи 23 742 пациентов с СД, которым была оказана медицинская помощь в период с 2002 по 2023 гг., при этом 90% данных представлены периодом 2015–2023 гг. Под записью понимается случай получения медицинской помощи по любым

поводам в ЭМК пациента с СД, который включает в себя врачебные осмотры, инструментальные и лабораторные исследования, кодировку заболеваний. В качестве первичного критерия включения в исследование использовался факт наличия у пациента в перечне диагнозов в ЭМК кода МКБ-10, относящегося к СД (E10–E14, включая все подкоды). Для выявления общей группы высокого риска госпитализаций по данной нозологии пациенты не делились на типы СД.

Способ формирования выборки из изучаемой популяции

Полученный НД содержал информацию о пациентах из 11 регионов Российской Федерации (Ямало-Ненецкий автономный округ, Республики Бурятия, Карелия, Крым, Саха (Якутия), Волгоградская, Кировская, Курганская, Московская и Челябинская области, Пермский край). Разделение на основании территориального распределения по типу 2b рекомендаций TRIPOD является более сильным дизайном внешней валидации предиктивных моделей, чем случайное разбиение (тип 2a) согласно данным рекомендациям [13]. Поэтому из НД были отложены 27 135 записей из республики Карелия для формирования выборки для внешней валидации. Общий дизайн исследования представлен на рис. 1. Итоговый набор

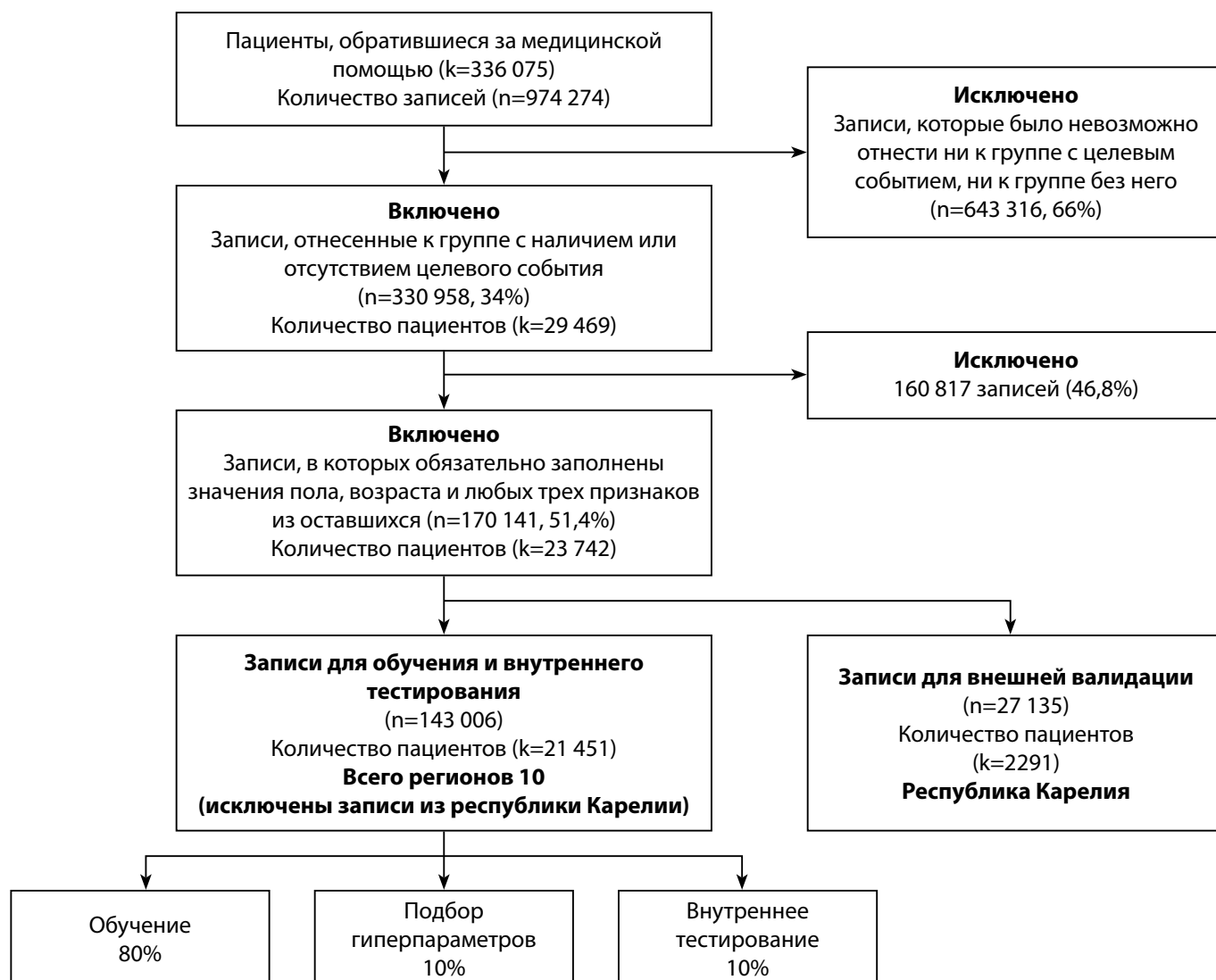


Рисунок 1. Дизайн исследования.

для разработки случайным образом делили на обучающую выборку (80%), выборку для подбора оптимальных гиперпараметров (10%) и выборку для внутреннего тестирования (10%). Выборки для внутреннего тестирования и внешней валидации не участвовали в обучении моделей.

Методы

Предикторы

На момент исследования платформа Webiomed поддерживала автоматическое извлечение более 2900 машиночитаемых признаков из электронных медицинских карт, включающих социодемографические, антропометрические, лабораторные, инструментальные, физикальные и анамнестические данные. В процессе формирования набора для каждой записи, соответствующей отдельному законченному случаю лечения, заполнялись значения признаков при их наличии у пациента. Количественные признаки с заполненностью менее чем в 10% записей итогового НД не использовались в разработке моделей. Отбор бинарных (логических) и категориальных признаков проводился медицинским экспертом на основании их клинической важности и частоты встре-

чаемости в итоговом НД. Всего 86 признаков было выбрано в качестве предикторов госпитализаций пациентов с СД для первого этапа моделирования (Приложение 1). На основании абсолютных значений векторов Шепли [14], полученных по результатам обучения алгоритмов на первоначальном НД, проводили дополнительную селекцию, отбирая для дальнейшего анализа и уточнения модели признаки с наибольшими абсолютными значениями и в совокупности формирующие 95% суммарного вектора Шепли всех исходных 86 признаков.

Целевое событие

В качестве целевого события был взят факт госпитализации пациента в круглосуточный стационар в течение 12 месяцев после обращения за медицинской помощью, и основной диагноз закрытия госпитализации относился к СД или его осложнению. Виды госпитализаций не учитывались ввиду ограничений первичных данных. Схема алгоритма отнесения записей пациента к группе с наличием или отсутствием целевого события представлена на рис. 2. Коды МКБ-10, использовавшиеся для отбора пациентов с СД, а также наличия или отсутствия целевого события представлены в табл. 1.

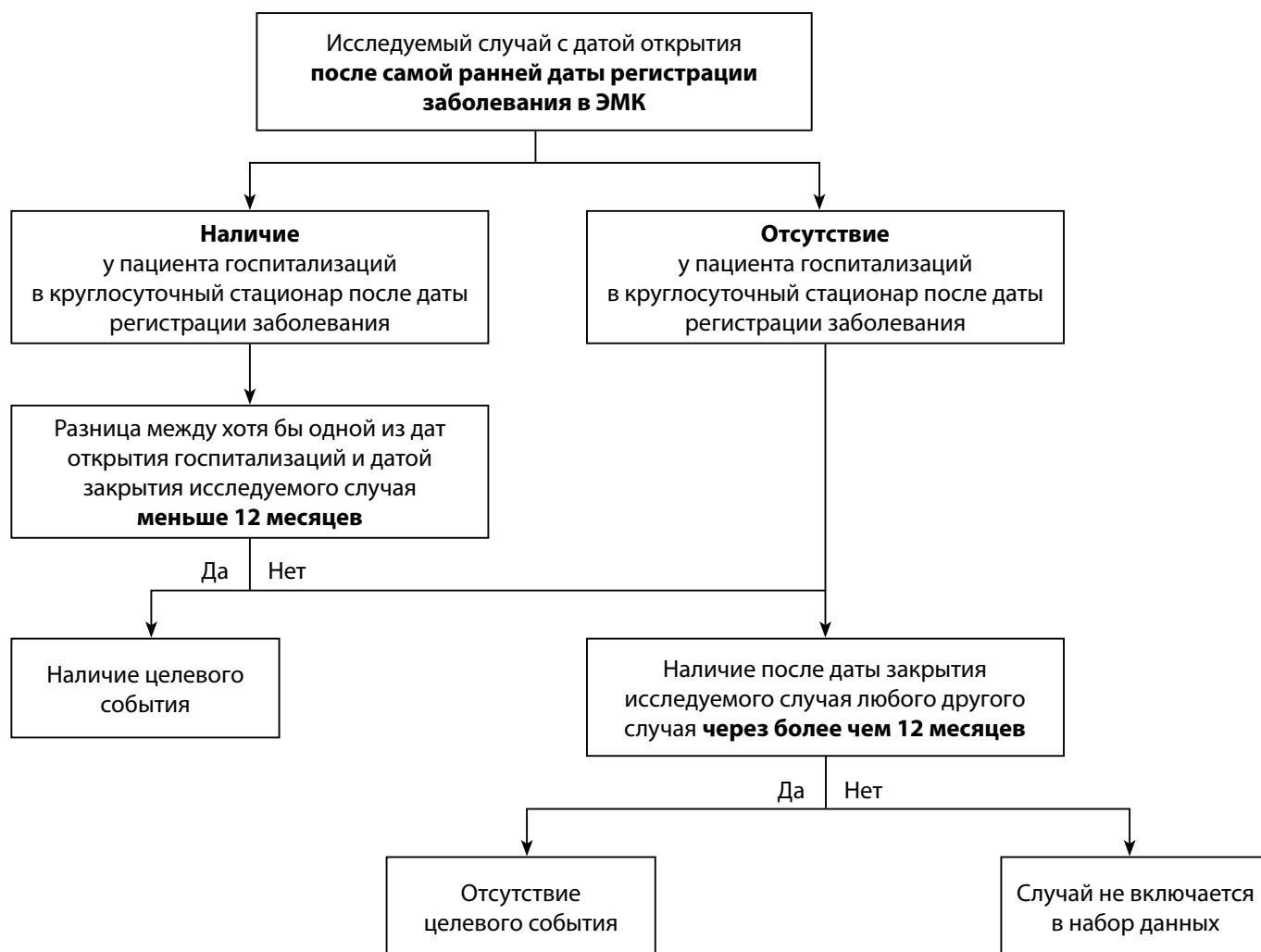


Рисунок 2. Алгоритм отнесения случая/записи к классу с наличием или отсутствием целевого события.

Примечание. Целевое событие — госпитализация в круглосуточный стационар в течение 12 месяцев после закрытия случая прогноза по поводу заболевания СД и/или его осложнений, представленных в табл. 1; ЭМК — электронная медицинская карта.

Таблица 1. Коды МКБ-10, наличие которых в электронных медицинских картах учитывалось на разных этапах формирования набора данных

Группа		Коды МКБ
Критерий включения пациентов в исследование		E10*, E11*, E12*, E13*, E14*
Отнесение госпитализации к целевому событию	Основное заболевание	E10*, E11*, E12*, E13*, E14*
	Метаболические осложнения	E15*, E16*, E87.2, R73
	Кардиологические осложнения	I20*, I21*, I22*, I23*, I24*, I25*, I50*
	Нефрологические осложнения	N17*, N18*, N19*, Z49, Z99.2, T82.4, Z94.0, T86.1, N08.3, N04.9, N05.9, N26, N28.9
	Цереброваскулярные осложнения	I61*, I63*, I64*, I65*, I66*, I69*, G45*
	Офтальмологические осложнения	H43.1, H35.8, H35.0, H35.2, H36.0, H28.0
	Диабетическая нейропатия	G59.0, G63.2, G99.0, G73.0, G62.9, G63.3
	Периферическая ангиопатия	Z89*, R02, L97, L98.4, I79.2, I73.9, I70.2, L00, L01, L02, L03, L04, L05, L08, M72.5, M72.6, A48.0
	Другие осложнения	M14.2, M14.6

Примечание. * — включая все подкоды.

Статистический анализ

Статистический анализ и построение моделей МО выполняли на языке программирования Python, версия 3.9. Нормальность распределения количественных переменных оценивали при помощи критерия Колмогорова-Смирнова, их данные отражены в виде медианы с указанием межквартильного размаха, категориальных — в виде долей (N, %). Сравнение количественных переменных в группах с целевым событием и без проводили при помощи теста Манна-Уитни, категориальных — при помощи χ^2 , значение $p < 0,05$ считалось статистически значимым.

В качестве алгоритмов МО применяли логистическую регрессию (LR), деревья решений с подходами к построению на основе градиентного бустинга (LightGBM, XGBoost, CatBoost) и бэггинга (RandomForest и ExtraTrees) и алгоритм на основе нейронных сетей (Multi-layer Perceptron). Внутреннюю валидацию полученных моделей проводили на тестовом НД. Критериями оценки эффективности и качества моделей служили AUC, чувствительность (Recall), специфичность, точность (Accuracy), прогностические ценности положительного (Precision) и отрицательного результатов, матрица ошибок и калибровочные кривые [15, 16]. Доверительные интервалы значений выбранных статистик оценивались с помощью метода бутстрап путем случайной генерации тысячи псевдовыборок [17]. В качестве порогов активации использовали максимум индекса Юдена, а также целевые уровни прогностических ценностей отрицательного (0,99) (ПЦОР) и положительного (0,65) результата (ПЦПР). Относительную значимость предикторов, используемых моделями машинного обучения, определяли по методу Шепли [14].

Разработка моделей

Первый этап обучения моделей проводили на сформированном НД для обучения с использованием всех признаков, изначально включенных в набор (прил. 1). Определение оптимальных гиперпараметров моделей проводилось с целью максимизации значения целевой метрики (AUC) на отдельном НД для определения гиперпараметров (рис. 1).

Модели с оптимальными гиперпараметрами проходили внутреннее тестирование. По итогам внутреннего тестирования до внешней валидации допускалось по одной модели от каждой архитектуры, показавшей максимальное значение AUC на тестовом НД при обязательном пересечении доверительных интервалов данной метрики с метрикой, полученной на НД для обучения (рис. 3). По результатам внешней валидации выбиралась модель с наилучшим результатом для последующего уточнения предикторов.

На основании значений усредненных векторов Шепли [14] были отобраны предикторы с наибольшими абсолютными значениями векторов Шепли и в совокупности вносящие 95% кумулятивного вклада в результат работы наилучшей модели по итогам проведенной внешней валидации на первом этапе.

Второй этап обучения проводился с использованием только отобранных признаков и определенных на первом этапе оптимальных гиперпараметров для каждой из архитектур. Все полученные модели также проходили внутреннее тестирование и внешнюю валидацию, подбор гиперпараметров на данном этапе не проводился. Выбор итоговой модели основывался на максимальном значении AUC при внешней валидации, а также при пересечении доверительных интервалов метрики и минимальном различии порогов активации по методу

максимизации индекса Юдена при внутреннем тестировании и внешней валидации. Алгоритм разработки и выбора итоговой модели представлен на рис. 3.

Этическая экспертиза

Для сбора данных компанией-разработчиком платформы Webiomed были подписаны соглашения с соот-

ветствующими 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.

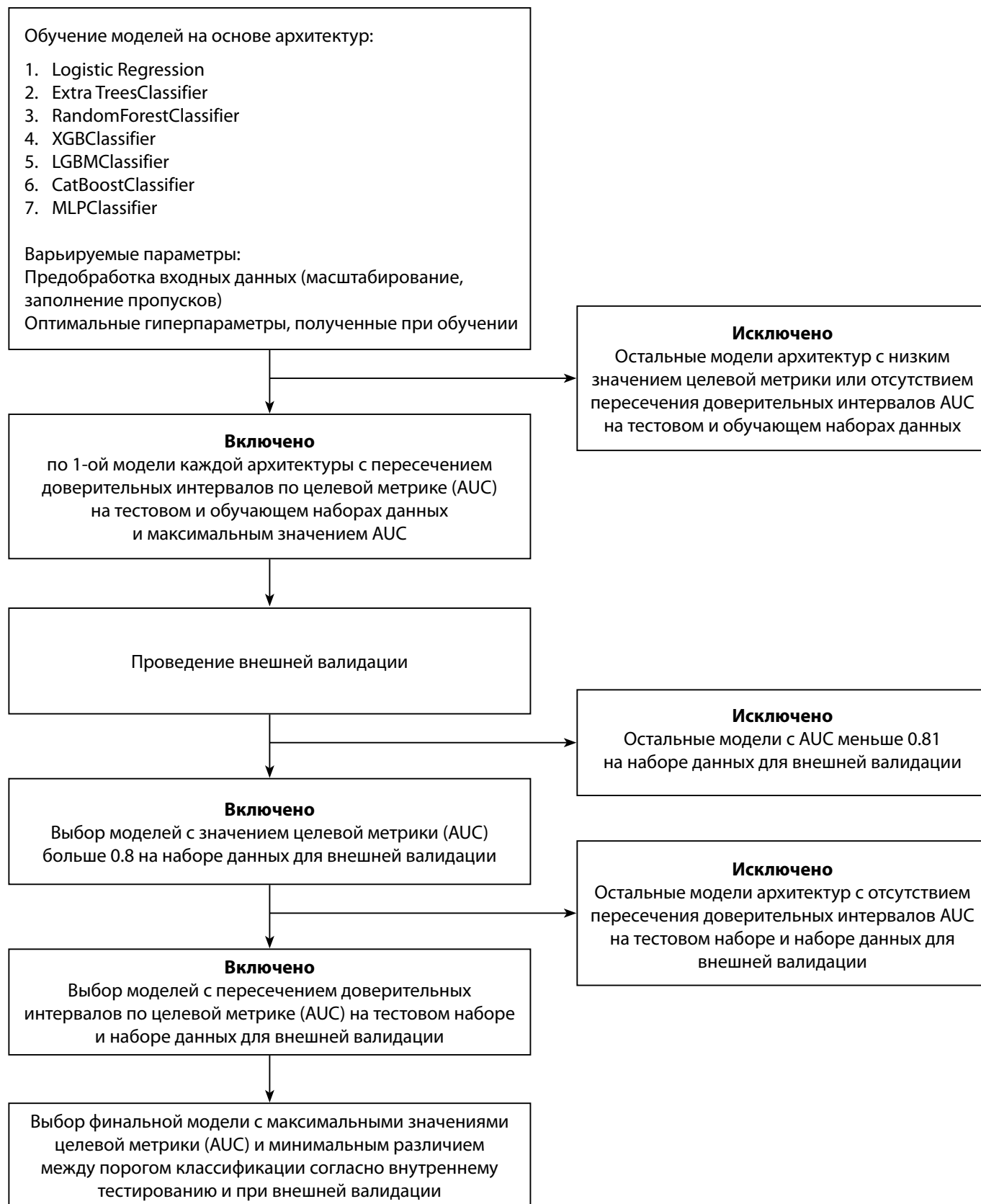


Рисунок 3. Алгоритм выбора итоговой модели.

РЕЗУЛЬТАТЫ

Описательная статистика. По результатам всех этапов отбора пациентов и их записей в качестве факторов прогнозирования для работы моделей было отобрано 33 предиктора (прил. 1, выделены жирным шрифтом). Заполненность выбранных признаков в НД до разделения

его на выборки для разработки и внешней валидации отражена на рис. 4А, распределение случаев (записей в НД) по количеству заполненных признаков с извлеченными значениями из базы данных Webiomed представлено на рис. 4Б. В сформированном наборе наилучшей заполненностью обладали половозрастная характеристика и антропометрические показатели пациентов. Среди

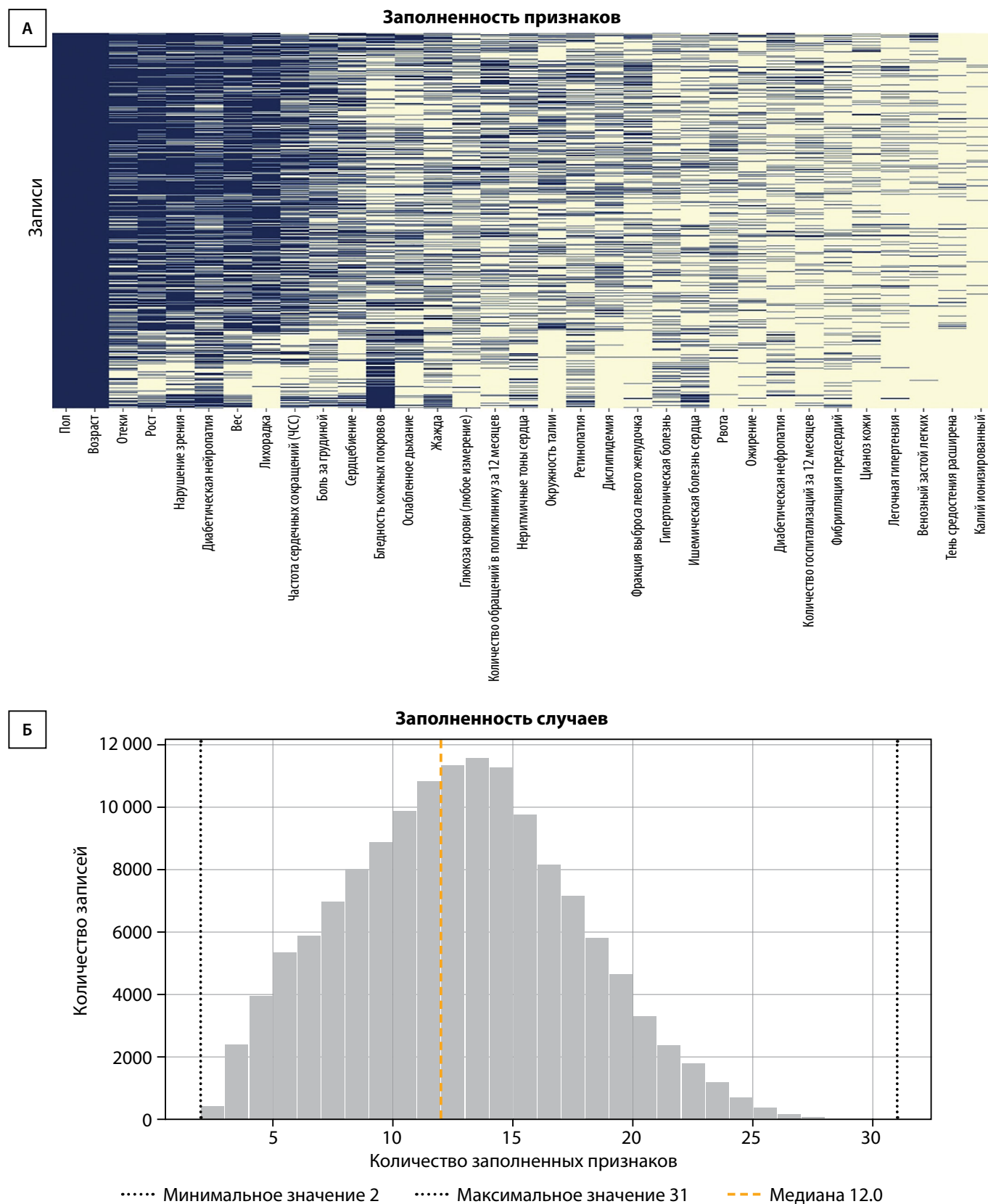


Рисунок 4. Заполненность значений признаков (А) и распределение количества заполненных признаков в случаях лечения/записей (Б) в итоговом наборе данных.

Таблица 2. Описательная статистика признаков набора данных

Признаки	[min, max]	Класс 0 (N=123328, 86,2%)	Класс 1 (N=19678, 13,8%)	P value	Набор для внешней валидации (N=27135)	Набор для обучения и внутреннего тестирования (N=143 006)	P value
Анамнестические							
ГБ, n (%)		27 437 (22,2)	5223 (26,5)	<0,0001	15380 (56,7)	32660 (22,8)	<0,0001
Диабетическая нейропатия, n (%)		34 427 (27,9)	10 311 (52,4)	<0,0001	7807 (28,8)	44 738 (31,3)	<0,0001
Диабетическая нефропатия, n (%)		15 231 (12,3)	5946 (30,2)	<0,0001	2554 (9,4)	21 177 (14,8)	<0,0001
Дислипидемия, n (%)		35 050 (28,4)	3677 (18,7)	<0,0001	2685 (9,9)	38 727 (27,1)	<0,0001
ИБС, n (%)		24 969 (20,2)	6350 (32,3)	<0,0001	5147 (19,0)	31 319 (21,9)	<0,0001
Количество предыдущих госпитализаций		0 (0; 1,0)	1,0 (0; 1,0)	<0,0001	0 (0; 1,0)	1,0 (0; 1,0)	
Количество обращений в поликлинику		1,0 (0; 2,0)	1,0 (0; 2,0)	<0,0001	1,0 (0; 2,0)	1,0 (0; 2,0)	
Легочная гипертензия, n (%)		12 980 (10,5)	3486 (17,7)	<0,0001	9039 (33,3)	16 466 (11,5)	<0,0001
Ожирение, n (%)		19 256 (15,6)	2306 (11,7)	<0,0001	3967 (14,6)	21 562 (15,1)	
Ретинопатия, n (%)		30 883 (25,0)	7951 (40,4)	<0,0001	5716 (21,1)	38 834 (27,2)	<0,0001
ФП, n (%)		14 635 (11,9)	4606 (23,4)	<0,0001	1708 (6,3)	19 241 (13,5)	<0,0001
Конституциональные							
Вес, кг	[20, 200]	80,0 [69,0; 93,0]	81,0 [70,0; 93,0]	0,0001	78,0 [68,6; 90,0]	80,0 [70,0; 93,0]	
Возраст, лет	[18, 100]	63,0 [53,0; 70,0]	65,0 [57,0; 72,0]	<0,0001	65,0 [56,0; 74,0]	63,0 [54,0; 70,0]	
Окружность талии, см		95,0 [86,0; 105]	96,0 [84,0; 105]		95,0 [87,0; 105]	95,0 [86,0; 105]	
Пол мужской, n (%)		35 089 (28,5)	7486 (38,0)	<0,0001	7623 (28,1)	42 575 (29,8)	
Рост, см	[100, 250]	163 [157; 170]	164 [158; 171]	<0,0001	162 [156; 169]	163 [157; 170]	
Клинические							
Бледность кожных покровов, n (%)		46 913 (38,0)	10 425 (53,0)	<0,0001	757 (2,8)	57 338 (40,1)	<0,0001
Боль за грудиной, n (%)		56 196 (45,6)	10 192 (51,8)	<0,0001	10 141 (37,4)	42 409 (29,7)	<0,0001
Жажда, n (%)		34 792 (28,2)	9564 (48,6)	<0,0001	10 899 (40,2)	44 356 (31,0)	<0,0001
Лихорадка, n (%)		84 164 (68,2)	11 580 (58,8)	<0,0001	22 775 (83,9)	95 744 (67,0)	<0,0001
Нарушение зрения, n (%)		88 893 (72,1)	13 565 (68,9)	<0,0001	22 224 (81,9)	102 458 (71,6)	<0,0001
Неритмичные тоны сердца, n (%)		34 317 (27,8)	7875 (40,0)	<0,0001	10 739 (39,6)	42 192 (29,5)	<0,0001
Ослабленное дыхание, n (%)		40 805 (33,1)	6579 (33,4)		1008 (3,7)	47 384 (33,1)	<0,0001
Отеки, n (%)		89 964 (73,0)	14 974 (76,1)	<0,0001	22 762 (83,9)	99 150 (69,3)	<0,0001
Рвота, n (%)		23 292 (18,9)	4360 (22,2)	<0,0001	11 277 (41,6)	27 652 (19,3)	<0,0001
Сердцебиение, n (%)		48 900 (39,7)	9669 (49,1)	<0,0001	16 749 (61,7)	58 569 (41,0)	<0,0001
Цианоз кожи, n (%)		14 389 (11,7)	3905 (19,8)	<0,0001	2249 (8,3)	18 294 (12,8)	<0,0001
Инструментальные							
Венозный застой легких, n (%)		8832 (7,2)	2796 (14,2)	<0,0001	2415 (8,9)	11 628 (8,1)	<0,0001
Тень средостения расширена, n (%)		8364 (6,8)	880 (4,5)	<0,0001	0 (0)	9244 (6,5)	
ФВЛЖ, %	[0, 100]	61,5 [57,0; 66,0]	60,0 [53,0; 65,2]	<0,0001	65,0 [61,0; 68,0]	61,0 [56,0; 66,0]	
ЧСС, уд. в мин	[25, 300]	72,0 [69,0; 78,0]	74,0 [70,0; 78,0]	<0,0001	72,0 [68,0; 76,5]	72,0 [69,0; 78,0]	
Лабораторные							
Глюкоза крови, ммоль/л	[1, 50]	6,8 [5,6; 8,9]	8,4 [6,3; 11,9]	<0,0001	7,0 [6,0; 8,8]	7,0 [5,7; 9,2]	
Калий крови, ммоль/л	[1, 10]	4,4 [4,1; 4,8]	4,5 [4,1; 4,8]		4,5 [4,1; 4,8]	4,4 [4,1; 4,8]	

Примечание. ГБ — гипертоническая болезнь; ИБС — ишемическая болезнь сердца; ФП — фибрилляция предсердий; ФВЛЖ — фракция выброса левого желудочка; ЧСС — частота сердечных сокращений. Значение $p < 0,05$ считалось статистически значимым.

наиболее часто встречающихся жалоб отмечены отеки различной локализации, нарушение зрения, боль за грудиной и сердцебиение. Самыми распространенными осложнениями основного заболевания и сопутствующими заболеваниями среди исследуемых пациентов были диабетическая нейропатия, ретинопатия и ишемическая болезнь сердца (ИБС). Медианное количество признаков с заполненными значениями в записях НД равнялось 12, при этом для наиболее заполненного случая из базы данных Webiomed были извлечены значения 31 признака из 33 использованных для прогнозирования.

По итогам формирования НД для обучения и внутреннего тестирования (т.е. разработки) в группу с наличием целевого события были включены 19 678 (14%) записей, в течение 12 месяцев после которых у пациентов была зафиксирована госпитализация по поводу СД или его осложнений. Оставшиеся 123 324 (86%) записи, в течение года после которых не было зафиксировано случаев лечения в круглосуточном стационаре по поводу СД при наличии любой информации о пациенте после 12 месяцев, отнесли к группе без целевого события (класс 0). При этом в результате работы алгоритма разделения случаев на группы с наличием или отсутствием целевого события 160 817 записей были полностью исключены из набора ввиду невозможности их отнесения к отдельной группе из-за отсутствия информации после законченного случая. Статистические показатели величины и частоты отобранных признаков в НД для обучения и тестирования, в зависимости от наличия целевого события, имели статистически значимые отличия и представлены в табл. 2.

В НД для внешней валидации в общей сложности вошло 27 135 записей пациентов из республики Карелии, среди которых было отмечено следующее разделение записей по классам: класс 1 — 959 (3,5%), класс 0 — 26 176 (96,5%). При сравнении распределения и частоты встреча-

емости отобранных нами предикторов, между двумя наборами были обнаружены статистически значимые различия в частоте встречаемости гипертонической болезни (ГБ), диабетической нейропатии, нефропатии и ретинопатии, ИБС, фибрилляции предсердий (ФП), легочной гипертензии (ЛГ), дислипидемии и ряда симптомов и жалоб пациентов (боль за грудиной, жажда, нарушение зрения, отеки, сердцебиение и т.д.). Общая характеристика использовавшихся НД, а также их сравнение представлены в табл. 2.

Коррекция выбросов и заполнение пропущенных значений. Заполнение пропущенных значений в численных параметрах выполняли с помощью фиксированного значения «-10000» [18], в бинарных — с помощью нулевых значений (отрицаний наличия признака). При обработке количественных признаков значения, выходящие за установленные медицинским экспертом на основании реальной клинической практики границы, также заменялись на значение «-10000». Использованные границы значений количественных признаков для удаления выбросов представлены в табл. 2. Также перед моделированием численные параметры трансформировали с помощью гистограммной нормализации [19].

Производительность моделей. Наиболее высокую дискриминативную способность и устойчивость к внешним данным в задаче прогнозирования госпитализаций пациентов с СД по поводу основного заболевания и его осложнений, при использовании отобранных в конечном счете 33 признаков, показала модель LightGBM со значением целевой метрики AUC 0,818 (95% ДИ 0,802–0,834) при внутреннем тестировании и 0,802 (95% ДИ 0,773–0,832) при внешней валидации. Значения AUC, полученные на наборах для внешней валидации и внутреннего тестирования для всех отобранных архитектур, представлены на рис. 5.

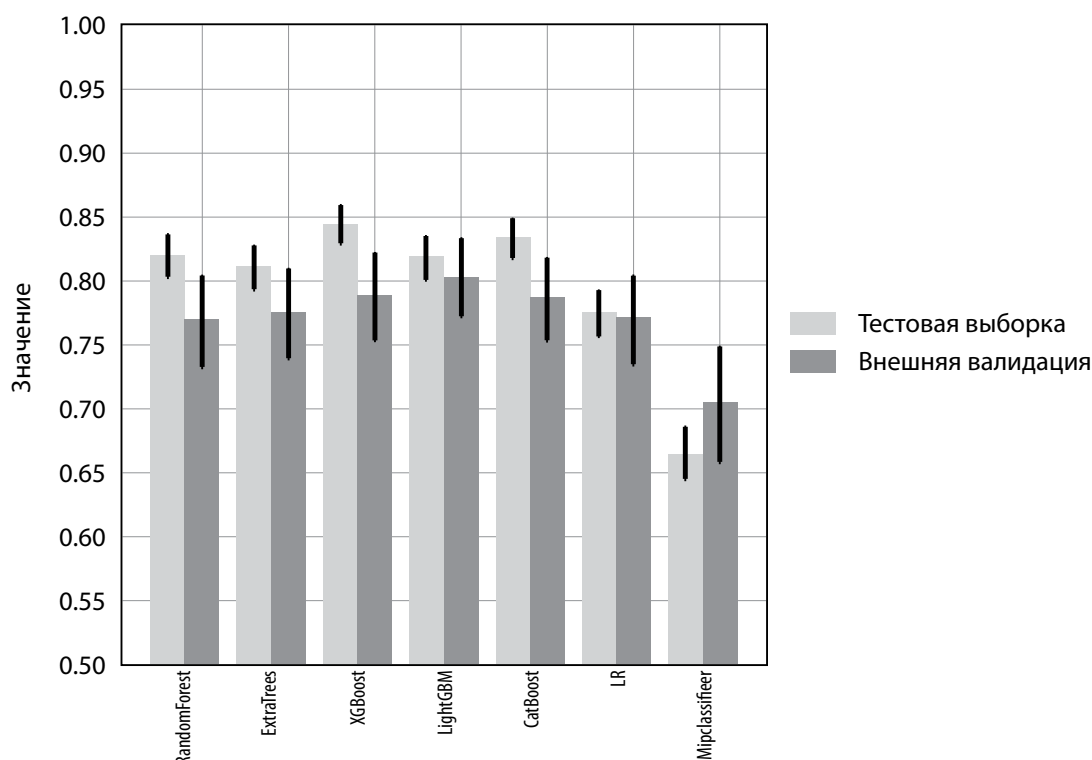


Рисунок 5. Прогнозирование госпитализаций у пациентов с сахарным диабетом. AUC на наборах для внутреннего тестирования и внешней валидации для отобранных моделей всех исследованных методов машинного обучения. Черные вертикальные линии указывают пределы 95% доверительных интервалов.

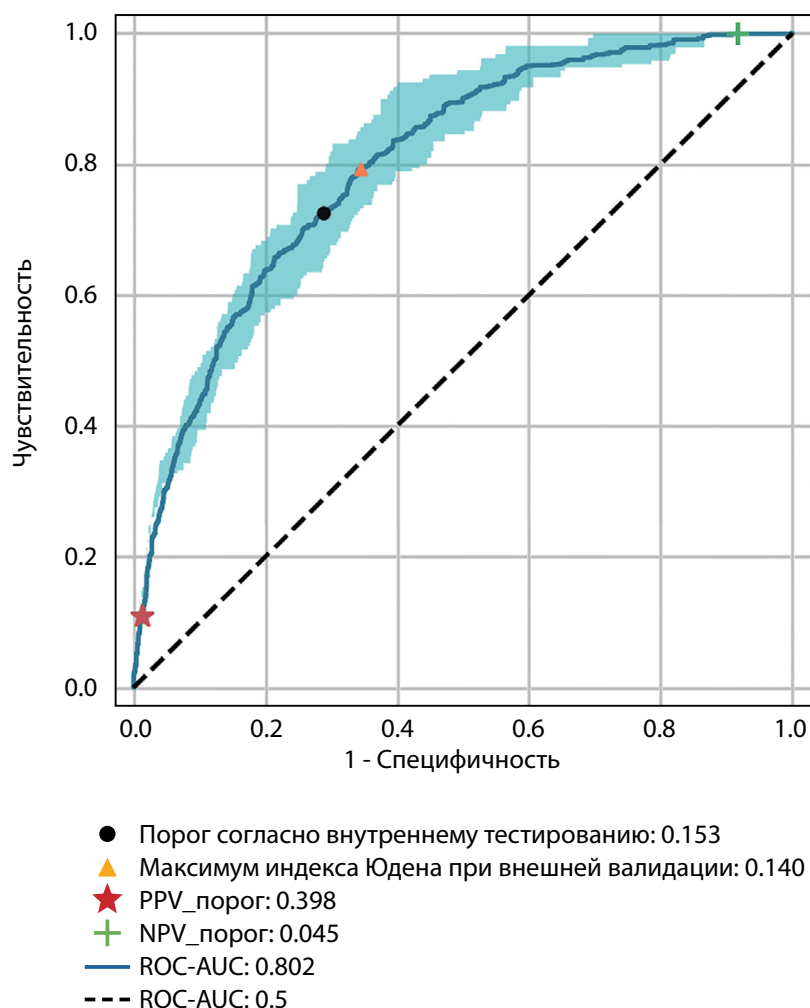


Рисунок 6. ROC-кривая с 95% ДИ итоговой модели LightGBM, полученная по результатам прогнозирования на наборе данных для внешней валидации. PPV_порог — целевое ПЦПР. NPV_порог — целевое ПЦОР.

Стоит отметить, что среди всех использовавшихся архитектур только модель LightGBM успешно прошла внешнюю валидацию и соответствовала всем критериям разработанного нами алгоритма отбора (рис. 3). Другие модели на основе градиентного бустинга (XGBoost и Catboost), несмотря на высокое значение целевой метрики на внутреннем тестировании (AUC 0,844 и 0,833 соответственно), показали значительную деградацию при использовании внешних данных (AUC 0,789 и 0,786 соответственно). В то же время при использовании деревьев решений отмечалось пересечение ДИ значений целевой метрики качества, полученных при внутреннем и внешнем тестировании, однако значения AUC были ниже целевого (0,81). Прогностическая ценность модели на основе нейронных сетей также оказалась крайне низкой на использовавшихся нами данных, AUC в задаче предсказания целевого события при внешней валидации составила 0,705 (95% ДИ 0,661–0,75).

На основании предложенного в недавних работах подхода с формированием трех групп риска [20] на наборе данных для внутреннего тестирования нами были дополнительно рассчитаны два порога бинаризации в зависимости от целевых прогностических ценностей отрицательного (0,99) и положительного (0,65) классов.

ROC-кривая финальной модели LightGBM при проведении внешней валидации с указанием трех порогов активности отражена на рис. 6.

Для внешних данных точность нашей модели с порогом классификации 0,045, при котором было достигнуто целевое значение ПЦОР на внутреннем тестировании (0,99), составила 0,117 (95% ДИ 0,108–0,126), чувствительность — 0,999 (95% ДИ 0,994–1), специфичность — 0,085 (95% ДИ 0,077–0,093). Метрики качества при использовании второго порога (0,398) с ожидаемой ПЦПР при тестировании, равном 0,65: точность — 0,957 (95% ДИ 0,951–0,963), чувствительность — 0,109 (95% ДИ 0,067–0,159), специфичность — 0,988 (95% ДИ 0,985–0,991). Полная характеристика эффективности итоговой модели после применения упомянутых порогов классификации, а также максимума индекса Юдена, рассчитанного на тестовом наборе, представлена в табл. 3 и 4, для внутреннего тестирования и внешней валидации соответственно. Абсолютные различия между порогами классификации для моделей на тестовом НД и максимумами индекса Юдена при внешней валидации указаны в табл. 5.

Значимость 10 наиболее важных признаков, определенных согласно рассчитанным векторам Шепли как ключевые в принятии решения выбранной модели,

Таблица 3. Метрики качества итоговой модели при проведении внутреннего тестирования с использованием различных порогов бинаризации

Порог	Значение	AUC	Точность	Чувств.	Специф.	ПЦПР	ПЦОР	F1 (полож.)	F1 (отриц.)	LR+	LR-
Максимум индекса Юдена	0,153	0,818 (0,802–0,834)	0,747 (0,736–0,759)	0,731 (0,699–0,765)	0,75 (0,737–0,763)	0,318 (0,297–0,34)	0,946 (0,939–0,953)	0,443 (0,42–0,468)	0,837 (0,828–0,846)	2,925 (2,745–3,134)	0,359 (0,314–0,402)
Максимум ПЦПР	0,398	0,818 (0,8–0,833)	0,876 (0,866–0,885)	0,205 (0,175–0,235)	0,983 (0,979–0,986)	0,653 (0,59–0,714)	0,886 (0,876–0,895)	0,312 (0,271–0,351)	0,932 (0,926–0,937)	11,775 (8,976–15,631)	0,809 (0,778–0,841)
Максимум ПЦОР	0,045	0,818 (0,803–0,834)	0,286 (0,274–0,299)	0,989 (0,981–0,996)	0,174 (0,163–0,186)	0,16 (0,15–0,171)	0,99 (0,982–0,996)	0,276 (0,26–0,292)	0,296 (0,28–0,313)	1,197 (1,179–1,217)	0,064 (0,024–0,113)

Примечание. В квадратных скобках указаны 95% доверительные интервалы метрик. ПЦПР — прогностическая ценность положительного результата. ПЦОР — прогностическая ценность отрицательного результата. F1 — F-мера. LR+(-) — отношение правдоподобия положительного (отрицательного) результатов.

Таблица 4. Метрики качества итоговой модели при проведении внешней валидации с использованием различных порогов бинаризации

Порог	Значение	AUC	Точность	Чувств.	Специф.	ПЦПР	ПЦОР	F1 (полож.)	F1 (отриц.)	LR+	LR-
Максимум индекса Юдена	0,153	0,802 (0,77–0,834)	0,712 (0,7–0,724)	0,726 (0,659–0,797)	0,711 (0,698–0,724)	0,084 (0,071–0,098)	0,986 (0,982–0,99)	0,151 (0,13–0,173)	0,826 (0,818–0,835)	2,513 (2,278–2,772)	0,386 (0,285–0,48)
Целевое ПЦПР	0,398	0,802 (0,773–0,832)	0,957 (0,951–0,963)	0,109 (0,067–0,159)	0,988 (0,985–0,991)	0,249 (0,155–0,356)	0,968 (0,963–0,973)	0,152 (0,095–0,215)	0,978 (0,975–0,981)	9,041 (5,119–14,555)	0,901 (0,852–0,945)
Целевое ПЦОР	0,045	0,802 (0,771–0,833)	0,117 (0,108–0,126)	0,999 (0,994–1)	0,085 (0,077–0,093)	0,038 (0,033–0,044)	1,0 (0,997–1)	0,074 (0,064–0,084)	0,156 (0,143–0,17)	1,091 (1,08–1,101)	0,012 (0–0,076)

Примечание. В квадратных скобках указаны 95% доверительные интервалы метрик. ПЦПР — прогностическая ценность положительного результата, ПЦОР — прогностическая ценность отрицательного результата.

Таблица 5. Пороги бинарной классификации, использовавшейся при разработке

Порог согласно внутреннему тестированию			Максимум индекса Юдена при внешней валидации		Абсолютная разница порогов	
RandomForest	0,167		0,152		0,015	
ExtraTrees	0,173		0,15		0,023	
XGBoost	0,822		0,857		0,035	
LightGBM	0,153		0,14		0,013	
CatBoost	0,155		0,204		0,049	
LR	0,496		0,462		0,034	
Mlpclassifier	0,162		0		0,162	

показана на рис. 7. Важнейшее значение для итоговой модели имеет факт наличия у пациента на момент прогноза установленных диагнозов диабетической нефропатии, нейропатии, ФП и дислипидемии, жалоб на жажду и упоминание лихорадки. Из конституциональных и антропометрических данных большую значимость имели мужской пол, возраст и окружность талии пациента. Средняя абсолютная ошибка калибровочной кривой равнялась 6,21%, ее график представлен на рис. 8.

ОБСУЖДЕНИЕ

СД относится к категории заболеваний с неуклонными темпами роста распространенности и с большим общественным бременем за счет ранней инвалидиза-

ции и высокой смертности. Общая численность людей, страдающих СД и состоящих на диспансерном учете в Российской Федерации, составила 4 962 762 человека (3,31% населения) по состоянию на 01.01.2023 г. [21]. Основной причиной смерти пациентов с СД продолжают оставаться хроническая сердечная недостаточность, ИБС, инфаркт миокарда и нарушения мозгового кровообращения [3], а экстренные госпитализации по поводу этих осложнений в настоящий момент приводят к значительной нагрузке на стационарное звено. Таким образом, прогнозирование риска госпитализации пациентов с СД в круглосуточный стационар и своевременная коррекция тактики их ведения особенно важны как для здравоохранения в общем, так и для профилактики развития осложнений у отдельного пациента.

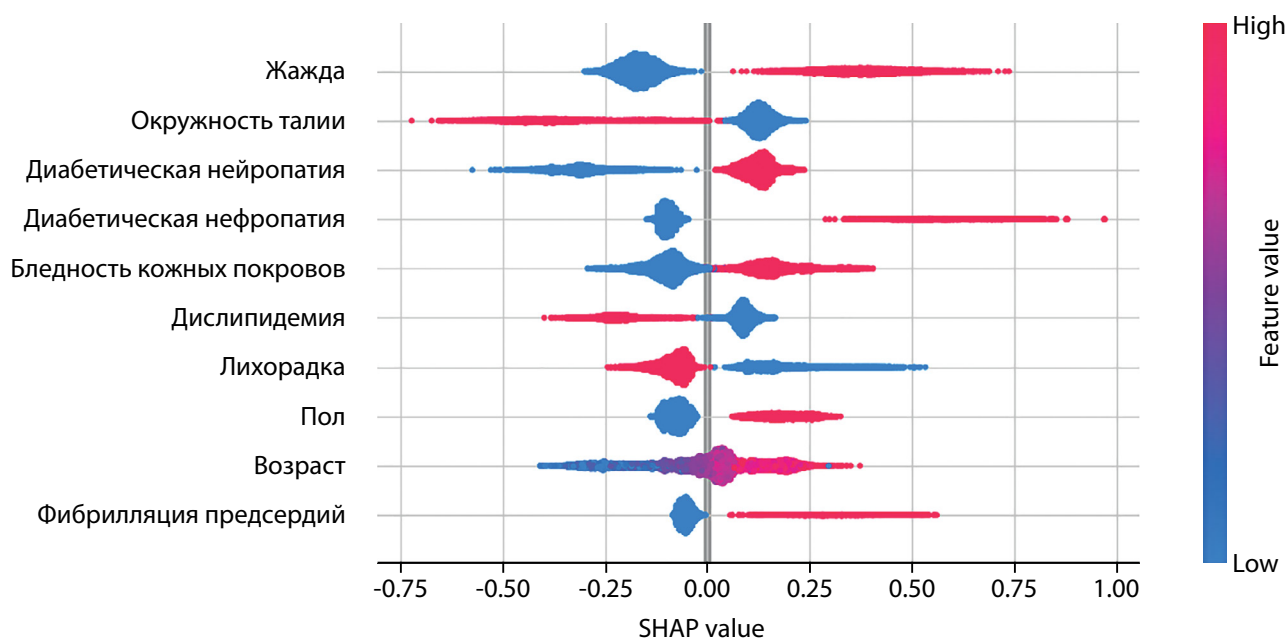


Рисунок 7. Топ-10 значимых признаков лучшей модели LightGBM.

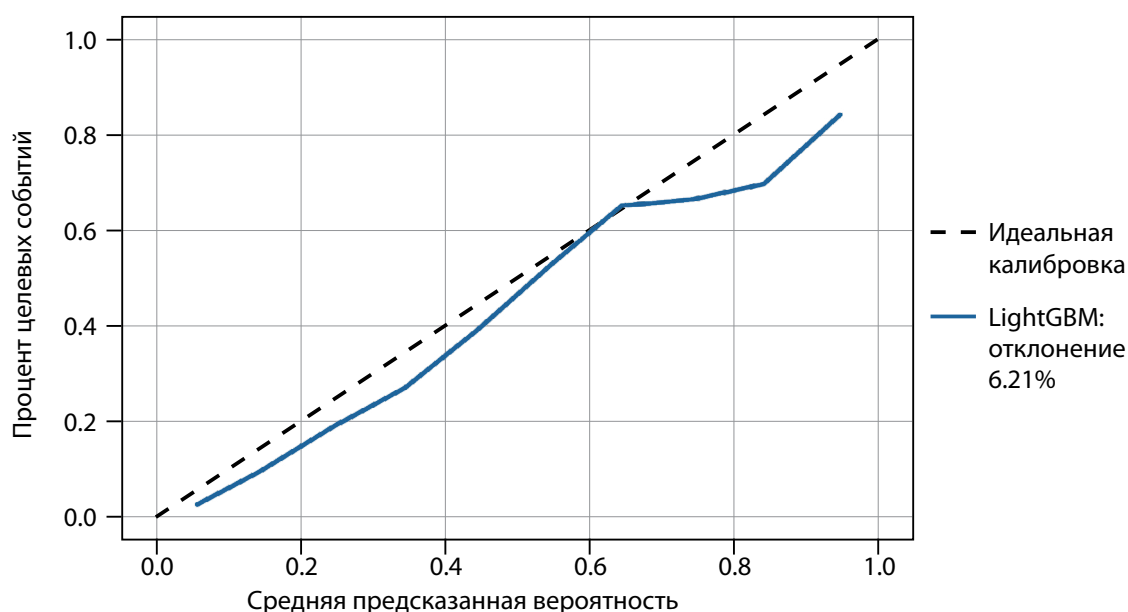


Рисунок 8. Калибровочная кривая вероятностей лучшей модели LightGBM.

Сопоставление с другими публикациями

Полученный нами результат по уровню целевой метрики превосходит аналогичные разработки зарубежных авторов. В упомянутых ранее работах, в отличие от нашего исследования, для создания моделей прогнозирования госпитализаций использовали записи пациентов с СД2. Лишь в публикации Naohui Lu и соавт. [10] критерием включения пациента в исследование был факт наличия СД вне зависимости от типа. Уклон большинства современных работ обращен в сторону предсказания риска повторных госпитализаций пациентов после выписки из стационара, в то время как в нашем исследовании в качестве случаев прогноза в НД входили как госпитализации, так и случаи амбулаторной медицинской помощи и диспансерные наблюдения пациентов. Из отличительных особенностей выбора предикторов в этих исследованиях хотелось бы отметить использование таких важных факторов, как длительность основного заболевания, факт использования различных гипогликемических препаратов, а также уделяние большего внимания лабораторным показателям.

Клиническая значимость результатов

При использовании двух рассчитанных порогов нами было предложено разделение записей на три группы риска в зависимости от выхода модели. При бинаризации на уровне первого порога (0,045), определенного с учетом ожидаемой прогностической ценности отрицательного класса в 0,99, финальная модель обладала высокой чувствительностью, что при отрицательном результате позволяет достоверно выделить пациентов с низким риском госпитализации, не требующих изменения тактики ведения. Использование второго порога (0,398) характеризуется высоким значением специфичности и прогностической ценности положительного класса. Прогнозирование наличия целевого события при использовании этого порога крайне важно, так как с большой надежностью предсказывает госпитализацию у анализируемого пациента, а значит указывает на необходимость более тщательного наблюдения за пациентом. Однако использование двух порогов неизбежно ведет к появлению «серой зоны» при интерпретации результатов модели, когда пациенты не могут быть отнесены ни к высокому, ни к низкому риску госпитализации. Это означает, что для данных пациентов результаты работы модели не имеют предиктивной ценности. Совокупность упомянутых фактов говорит о наличии у разработанной нами модели большого потенциала в качестве дополнительного инструмента для мониторинга за пациентами с СД. С учетом продемонстрированной устойчивости модели в рамках внешней валидации она может быть допущена к практическому использованию в рамках проспективных исследований.

Ограничения исследования и направления дальнейших исследований

Важным ограничением исследования является отсутствие во входящих признаках лекарственной терапии СД. Факты изменения подходов к лекарственной терапии за период извлечения данных в 21 год, невысокая точность извлечения препаратов и их доз, невозможность оценки приверженности терапии пациентами на реальных клинических данных не позволили включить лекарственную терапию в обучение моделей. Для дальнейше-

го повышения метрик точности необходимо увеличение размера обучающей выборки, в частности за счет записей с наличием целевого события [22]. Учитывая растущую потребность в достоверном выделении как группы высокого риска, так и группы пациентов с крайне низким риском госпитализации, необходимы новые разработки алгоритма определения оптимальных значений целевых ПЦОР и ПЦПР при расчете порогов. Отдельного внимания для будущих исследований заслуживает интерпретация модели на случаях прогноза, для которых результат работы оказался в интервале между двумя предлагаемыми нами порогом. Использование перевода англоязычного термина «неопределенного риска» не передает какого-либо клинического смысла, а значит, в настоящее время неприменим и требует дальнейшего исследования по улучшению классификации. Использование в качестве признаков схем гипогликемических препаратов и оценка приверженности к ним могут значительно повысить ценность модели. Определенным потенциалом также обладает создание ансамблей моделей на основе деления пациентов по типу СД, осложнений и коморбидных состояний.

ЗАКЛЮЧЕНИЕ

Методы МО являются эффективными инструментами в разработке моделей для прогнозирования исходов таких многофакторных заболеваний, как СД. При этом в качестве факторов используются рутинные клинико-лабораторные параметры, которые можно легко отслеживать и контролировать при рутинном визите пациента. Метрики полученной итоговой модели превосходили опубликованные ранее аналоги. Среди использовавшихся алгоритмов наилучшие метрики качества показали модели на основе градиентного бустинга, превзойдя результаты деревьев решений, логистической регрессии и нейронных сетей. Результаты внешней валидации показали относительную устойчивость модели к новым данным из другого региона, что в совокупности с показателями качества отражает возможность ее использования в реальной клинической практике. Использование подхода с выделением трех групп риска значительно повышает ценность использования модели за счет выделения групп пациентов с действительно высоким или крайне низким риском госпитализации.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Источники финансирования. Исследование выполнено за счет собственных средств компании.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с содержанием настоящей статьи.

Участие авторов. Андрейченко А.Е. — дизайн исследования, написание статьи; Ермак А.Д. — анализ данных, моделирование, написание статьи; Гаврилов Д.В. — интерпретация результатов, редакция статьи; Новицкий Р.Э. — получение данных, редакция статьи; Гусев А.В. — концепция исследования, редакция статьи.

Все авторы одобрили финальную версию статьи перед публикацией, выразили согласие нести ответственность за все аспекты работы, подразумевающую надлежащее изучение и решение вопросов, связанных с точностью или добросовестностью любой части работы.

ПРИЛОЖЕНИЕ 1.

**Список предикторов, использованных в разработке первоначальных версий моделей
(в алфавитном порядке).****Жирным выделены предикторы, использующиеся в итоговых моделях.**

1. Аланинаминотрансфераза
2. Алкогольная зависимость
3. Аспаратаминотрансфераза
4. Атеросклероз
5. Бледность кожных покровов
6. Боль в животе
7. Боль за грудиной
8. Боль, связанная с дыханием
9. Венозный застой легких
10. Вес
11. Вздутие живота
12. Возраст
13. Гемоглобин гликированный
14. Гемоглобин крови
15. Гепатомегалия
16. Гипертоническая болезнь
17. Гипертрофия левого желудочка
18. Глюкоза крови (любое измерение)
19. Диабетическая нейропатия
20. Диабетическая нефропатия
21. Диарея
22. Диастолическое АД
23. Дислипидемия
24. Жажда
25. Жар
26. Железодефицитная анемия
27. Жесткое дыхание
28. Запоры
29. Изменение фазы реполяризации
30. Ишемическая болезнь сердца
31. Калий ионизированный
32. Кожный зуд
33. Количество госпитализаций за 12 месяцев
34. Количество обращений в поликлинику по лечебно-диагностическим поводам за 12 месяцев
35. Креатинин крови
36. Крупноочаговые изменения
37. Легочная гипертензия
38. Лейкоциты крови
39. Лихорадка
40. Международное нормализованное соотношение
41. Мокрота
42. Нарушение зрения

43. Нарушение локальной сократимости миокарда ЛЖ

44. Нарушения мочеиспускания

45. Нейтрофилы

46. Неритмичные тоны сердца

47. Неустойчивый стул

48. Носовые кровотечения

49. Общая слабость

50. Общий билирубин

51. Общий холестерин

52. Одышка

53. Ожирение

54. Окружность талии

55. Ослабленное дыхание

56. Отек легких

57. Отеки

58. Отягощенная наследственность по инфаркту миокарда в возрасте до 60 лет у ближайших родственников

59. Плохая переносимость физических нагрузок

60. Пол

61. Полная блокада левой ножки пучка Гиса

62. Почечная недостаточность

63. Ретинопатия

64. Рост

65. Сердечная недостаточность

66. Сердцебиение

67. Систолическое АД

68. Суправентрикулярная тахикардия

69. Суставные боли

70. Сухость кожи

71. Табакокурение

72. Температура тела

73. Тень средостения расширена

74. Тошнота

75. Тромбоциты

76. Усиление легочного рисунка

77. Утомляемость

78. Фибрилляция предсердий

79. Фракция выброса левого желудочка

80. Хрипы в легких

81. Хронический бронхит

82. Цереброваскулярные заболевания

83. Цианоз кожи

84. Частота дыхания

85. Частота сердечных сокращений

86. Эритроциты

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

1. Алгоритмы специализированной медицинской помощи больным сахарным диабетом. Под редакцией И.И. Дедова, М.В. Шестаковой, А.Ю. Майорова. 10-й выпуск. *Сахарный диабет*. 2021. — Т. 24. — 15. — С. 1-148. [Dedov II, Shestakova M.V., Mayorov A.Yu. 10th edition. *Diabetes mellitus*. 2022;24(15):1-148. (In Russ.)] doi: <https://doi.org/10.14341/DM12802>
2. World Health Organization, International Diabetes Federation. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. Report of a WHO/IDF consultation. Geneva; 2006
3. Дедов И.И., Викулова О.К., Железнякова А.В., Исаков М.А. Эпидемиологические характеристики сахарного диабета в Российской Федерации: клинко-статистический анализ по данным регистра сахарного диабета на 01.01.2021 // *Сахарный диабет*. 2021. — Т. 24. — №3. — С. 204-221. [Dedov II, Shestakova M.V., Vikulova O.K., Zheleznyakova A.V., Isakov M.A. Epidemiological characteristics of diabetes mellitus in the Russian Federation: clinical and statistical analysis according to the Federal diabetes register data of 01.01.2021. *Diabetes Mellit*. 2021;24(3):204-221. (In Russ.)] doi: <https://doi.org/10.14341/DM12759>
4. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*. 2019; 157:107843
5. Jeong IS, Kang CM. Time to Diagnosis and Treatment of Diabetes Mellitus among Korean Adults with Hyperglycemia: Using a Community-Based Cohort Study. *International Journal of Environmental Research and Public Health*. 2022; 19(19):12090. doi: <https://doi.org/10.3390/ijerph191912090>
6. Bommer C, Heeseemann E, Sagalova V, et al. The global economic burden of diabetes in adults aged 20-79 years: a cost-of-illness study. *Lancet Diabetes Endocrinol* 2017;5(6):423-30. doi: [https://doi.org/10.1016/S2213-8587\(17\)30097-9](https://doi.org/10.1016/S2213-8587(17)30097-9)
7. Caughey GE, Pratt NL, Barratt JD, Shakib S, Kemp-Casey AR, Roughead EE. Understanding 30-day re-admission after hospitalisation of older patients for diabetes: identifying those at greatest risk. *Med J Aust*. 2017;206(4):170-175. doi: <https://doi.org/10.5694/mja16.00671>
8. Li T-C, Li C-I, Liu C-S, et al. Development and validation of prediction models for the risks of diabetes-related hospitalization and in-hospital mortality in patients with type 2 diabetes. *Metabolism*. 2018;85:38-47. doi: <https://doi.org/10.1016/j.metabol.2018.02.003>
9. Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. Predicting diabetes-related hospitalizations based on electronic health records. *Stat Methods Med Res*. 2019; 28: 3667-82
10. Lu H, Uddin S. Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. *Information*. 2022; 13(9):436
11. Hai AA, et al. Deep Learning vs Traditional Models for Predicting Hospital Readmission among Patients with Diabetes
12. Bhuvan MS, Kumar A, Zafar A, Kishore V. Identifying Diabetic Patients with High Risk of Readmission. Feb. 2016, [Online]
13. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-73
14. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. 2020; 2(1): 56-67. doi: <https://doi.org/10.1038/s42256-019-0138-9>
15. Strilets V, Bakumenko N, Donets V, et al. Machine Learning Methods in Medicine Diagnostics Problem, 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer: proceedings, 2020, Vol. II: pp. 89-101
16. Van Calster B, McLernon D J, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019; 17(1): 230. doi: <https://doi.org/10.1186/s12916-019-1466-7>
17. Zoubir AM, Iskander DR. Bootstrap Methods and Applications. In *IEEE Signal Processing Magazine*. 2007;24(4):10-19. doi: <https://doi.org/10.1109/MSP.2007.4286560>
18. Ding, Yufeng and Jeffrey S. Simonoff. An Investigation of Missing Data Methods for Classification Trees. *Econometrics: Data Collection & Data Estimation Methodology eJournal* (2006)
19. Awais M, Shamshad F, Bae S. (2019). Towards an Adversarially Robust Normalization Approach. *ArXiv*, abs/2006.11007
20. Fischer BG, Evans AT 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.
21. Дедов И.И., Шестакова М.В., Викулова О.К., и др. Сахарный диабет в Российской Федерации: динамика эпидемиологических показателей по данным Федерального регистра сахарного диабета за период 2010–2022 гг. *Сахарный диабет*. 2023. — Т. 26. — №2. — С. 104-123. [Dedov II, Shestakova M.V., Vikulova O.K., et al. Diabetes mellitus in the Russian Federation: dynamics of epidemiological indicators according to the Federal Register of Diabetes Mellitus for the period 2010–2022. *Diabetes mellitus*. 2023;26(2):104-123. (In Russ.)] doi: <https://doi.org/10.14341/DM13035>
22. Riley RD, Debray TPA., Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*. 2021; 40(19): 4230-4251

ИНФОРМАЦИЯ ОБ АВТОРАХ [AUTHORS INFO]

***Андрейченко Анна Евгеньевна**, к.ф.-м.н. [Anna E. Andreychenko, PhD in Physics and Mathematics]; адрес: Россия, 185910 Петрозаводск, набережная Варкауса, д. 17 [address: 17 Varkaus Embankment, 185901 Petrozavodsk, Russia]; ORCID: <https://orcid.org/0000-0001-6359-0763>; e-mail: aandreychenko@webiomed.ru

Ермак Андрей Дмитриевич [Andrey D. Ermak, Data analyst, Artificial Intelligence Department, K-SkAI]; ORCID: <https://orcid.org/0000-0002-0513-8557>; e-mail: aermak@webiomed.ru

Гаврилов Денис Владимирович [Denis V. Gavrilov]; ORCID: <https://orcid.org/0000-0002-8745-857X>; e-mail: dgavrilov@webiomed.ru

Новицкий Роман Эдвардович [Roman E. Novitskiy]; ORCID: <https://orcid.org/0000-0002-2350-977X>; e-mail: roman@webiomed.ru

Гусев Александр Владимирович, к.т.н., ст.н.с. [Alexander V. Gusev, PhD in Engineering, Senior Researcher]; ORCID: <https://orcid.org/0000-0002-7380-8460>; e-mail: agusev@webiomed.ai

ЦИТИРОВАТЬ:

Андрейченко А.Е., Ермак А.Д., Гаврилов Д.В., Новицкий Р.Э., Гусев А.В. Разработка и валидация моделей машинного обучения, прогнозирующих риск госпитализации пациентов с сахарным диабетом в течение последующих 12 месяцев // *Сахарный диабет*. — 2024. — Т. 27. — №2. — С. 142-157. doi: <https://doi.org/10.14341/DM13065>

TO CITE THIS ARTICLE:

Andreychenko AE, Ermak AD, Gavrilov DV, Novitskiy RE, Gusev AV. Development and validation of machine learning models to predict unplanned hospitalizations of patients with diabetes within the next 12 months. *Diabetes Mellitus*. 2024;27(2):142-157. doi: <https://doi.org/10.14341/DM13065>