

Оценка вероятности тромбоэмболии легочной артерии при помощи модели машинного обучения

Гаврилов Д. В.¹, Андрейченко А. Е.¹, Ермак А. Д.¹, Кузнецова Т. Ю.², Гусев А. В.³

Цель. Разработать и валидировать модель машинного обучения, предназначенную для выявления подозрения на тромбоэмболию легочной артерии (ТЭЛА) по различным клиническим признакам из электронных медицинских карт (ЭМК) пациентов, обращающихся за амбулаторной и стационарной помощью.

Материал и методы. Данные 19730 пациентов из 7 регионов Российской Федерации были взяты для анализа. Период накопления данных ЭМК: с 21.03.2007 по 04.02.2022. В качестве диагностических признаков использованы жалобы, клинические признаки, результаты лабораторных исследований, сопутствующие заболевания. У 1379 пациентов была диагностирована ТЭЛА. Диагностика ТЭЛА была на основании кодов МКБ-10 в заключительном диагнозе случая лечения. Было применено 7 алгоритмов машинного обучения для выполнения задачи диагностики ТЭЛА: XGBoost, LightGBM, CatBoost, Logistic Regression, MLP Classifier, Random Forest Classifier, Gradient Boosting Classifier.

Результаты. Модель на основе алгоритма Gradient Boosting Classifier была выбрана для дальнейшей проспективной апробации: чувствительность 0,899 (95% доверительный интервал (ДИ): 0,864-0,932), специфичность 0,875 (95% ДИ: 0,863-0,86), площадь под ROC-кривой 0,952 (95% ДИ: 0,938-0,964). Наибольшую значимость для предсказания имели признаки: кашель, дыхательные нарушения, креатинин крови, температура тела, общая слабость, частота сердечных сокращений, частота дыхания, отеки, антигипертензивная терапия, сатурация и возраст.

Заключение. Обученная модель рассчитана для использования при первичном обращении за медицинской помощью пациентов с жалобами и подозрением на ТЭЛА вне зависимости от вида помощи.

Ключевые слова: тромбоэмболия легочной артерии, электронные медицинские карты, машинное обучение, система поддержки клинических решений.

Отношения и деятельность: нет.

¹К-Скай, Петрозаводск; ²ФГБОУ ВО ПетрГУ, Петрозаводск; ³ФГБУ Центральный научно-исследовательский институт организации и информатизации здравоохранения Минздрава России, Москва, Россия.

Гаврилов Д. В.* — руководитель медицинского направления, ORCID: 0000-0002-8745-857X, Андрейченко А. Е. — руководитель направления искусственного интеллекта, ORCID: 0000-0001-6359-0763, Ермак А. Д. — аналитик данных направления искусственного интеллекта, ORCID: 0000-0002-0513-8557, Кузнецова Т. Ю. — д.м.н., доцент, зав. кафедрой факультетской терапии, фтизиатрии, инфекционных болезней и эпидемиологии, ORCID: 0000-0002-6654-1382, Гусев А. В. — к.т.н., эксперт, ORCID: 0000-0002-7380-8460.

*Автор, ответственный за переписку (Corresponding author): dgavrilov@webiomed.ru

АД — артериальное давление, БД — база данных, ДИ — доверительный интервал, ЗНО — злокачественное новообразование, КТ — компьютерная томография, ЛА — легочная артерия, МКБ-10 — международная классификация болезней 10 пересмотра, НД — набор данных, ТЭЛА — тромбоэмболия легочной артерии, ЭКГ — электрокардиография, ЭМК — электронные медицинские карты.

Рукопись получена 03.11.2023

Рецензия получена 16.01.2024

Принята к публикации 27.03.2024



Для цитирования: Гаврилов Д. В., Андрейченко А. Е., Ермак А. Д., Кузнецова Т. Ю., Гусев А. В. Оценка вероятности тромбоэмболии легочной артерии при помощи модели машинного обучения. *Российский кардиологический журнал*. 2024;29(4):5679. doi: 10.15829/1560-4071-2024-5679. EDN QHZHPQ

Assessment of pulmonary embolism probability using a machine learning model

Gavrilov D. V.¹, Andreichenko A. E.¹, Ermak A. D.¹, Kuznetsova T. Yu.², Gusev A. V.³

Aim. To develop and validate a machine learning model designed to identify suspected pulmonary embolism (PE) based on various clinical features from electronic health records (EHRs) of out- and inpatients.

Material and methods. Data from 19730 patients from 7 Russian regions were taken for analysis. EHR data were analyzed for the period from March 21, 2007 to February 4, 2022. Complaints, clinical and laboratory data, and concomitant diseases were used as diagnostic signs. PE was diagnosed in 1379 patients. Diagnosis of PE was based on ICD-10 codes. Seven machine learning algorithms were applied to diagnose pulmonary embolism: XGBoost, LightGBM, CatBoost, Logistic Regression, MLP Classifier, Random Forest Classifier, Gradient Boosting Classifier.

Results. The Gradient Boosting Classifier-based model was selected for further prospective testing with the sensitivity of 0,899 (95% confidence interval (CI), 0,864-0,932), specificity of 0,875 (95% CI, 0,863-0,86), area under the ROC curve of 0,952 (95% CI, 0,938-0,964). The following signs had the greatest prediction value: cough, respiratory disorders, blood creatinine, body temperature, general weakness, heart rate, respiratory rate, edema, antihypertensive therapy, saturation and age.

Conclusion. The model is designed for the initial encounter of patients with complaints and suspected PE, regardless of the type of care.

Keywords: pulmonary embolism, electronic health records, machine learning, clinical decision support system.

Relationships and Activities: none.

¹K-Sky, Petrozavodsk; ²Petrozavodsk State University, Petrozavodsk; ³Central Research Institute for Health Organization and Informatics, Moscow, Russia.

Gavrilov D. V.* ORCID: 0000-0002-8745-857X, Andreichenko A. E. ORCID: 0000-0001-6359-0763, Ermak A. D. ORCID: 0000-0002-0513-8557, Kuznetsova T. Yu. ORCID: 0000-0002-6654-1382, Gusev A. V. ORCID: 0000-0002-7380-8460.

*Corresponding author: dgavrilov@webiomed.ru

Received: 03.11.2023 **Revision Received:** 16.01.2024 **Accepted:** 27.03.2024

For citation: Gavrilov D. V., Andreichenko A. E., Ermak A. D., Kuznetsova T. Yu., Gusev A. V. Assessment of pulmonary embolism probability using a machine learning model. *Russian Journal of Cardiology*. 2024;29(4):5679. doi: 10.15829/1560-4071-2024-5679. EDN QHZHPQ

Ключевые моменты

- Диагностика тромбоэмболии легочной артерии (ТЭЛА) представляет собой трудности из-за неспецифичности клинических проявлений, совпадающих с многими терапевтическими заболеваниями.
- Объединение клинических проявлений ТЭЛА с использованием технологии машинного обучения позволяет анализировать большое количество клинически значимых параметров и улучшать предсказательную точность.
- Обученная модель включает в себя ранее используемые диагностические параметры и новые признаки, встречающиеся в электронных медицинских картах, с целью поддержки врачебных решений для диагностики ТЭЛА.

Тромбоэмболия легочной артерии (ТЭЛА) представляет собой неотложное состояние, имеющее потенциально высокий риск смерти. Распространенность ТЭЛА среди острых состояний уступает только инфаркту миокарда и инсульту, занимая третье место в мире [1]. Несмотря на современные алгоритмы диагностики и подходы к терапии, смертность от этого заболевания остается высокой и составляет 30% при не леченной и 8% при леченной ТЭЛА [2]. При этом имеется проблема своевременной диагностики этого состояния: из проанализи-

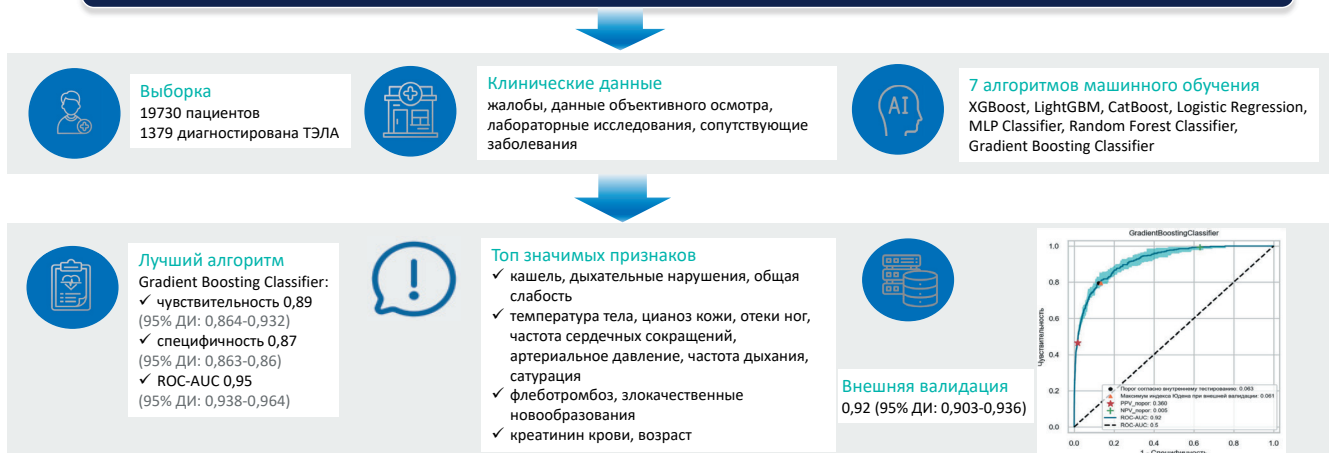
Key messages

- Diagnosis of pulmonary embolism (PE) is difficult due to nonspecific clinical manifestations, which coincide with many therapeutic diseases.
- Integrating clinical manifestations of pulmonary embolism using machine learning allows the analysis of a large number of clinically relevant parameters and improves predictive accuracy.
- The trained model incorporates previously used diagnostic parameters and new features found in electronic health records to support decision making in suspected PE.

рованных >370 тыс. смертей от ТЭЛА 34% умерли внезапно до начала терапии, у 59% ТЭЛА как причина летального исхода была диагностирована лишь посмертно [3].

Сложность распознавания ТЭЛА обусловлена неспецифичностью клинических проявлений, которые в значительной мере совпадают с другими смежными заболеваниями: острым коронарным синдромом, сердечной недостаточностью, пневмонией, обструктивной болезнью легких [4]. Для решения этой проблемы клинические признаки были объединены в алгоритмы, построенные с использованием логистической регрессии и позволяющие оценить вероятность ТЭЛА на основании балльной оценки каждого признака. Разработанные алгоритмы Geneva

Данные пациентов (N=2,9 млн) 7 регионов РФ с амбулаторными и стационарными обращениями с 01.01.2001 до 31.12.2022 по терапевтическим заболеваниям со схожей с ТЭЛА клинической картиной



Поддержка принятия врачебного решения при подозрении на ТЭЛА у терапевтических больных при обращении за медицинской помощью.
Использование большого числа клинических предикторов.
Применимость для амбулаторных и стационарных пациентов.

Сокращения: ДИ — доверительный интервал, ТЭЛА — тромбоэмболия легочной артерии.

Таблица 1

Клинические признаки,
отобранные для обучения модели

№	Признак	Характеристика признака
1	Возраст пациента	лет
2	Пол	мужской/женский
3	Рост	см
4	Масса тела	кг
5	Температура тела	° С
6	Систолическое артериальное давление	мм рт.ст.
7	Диастолическое артериальное давление	мм рт.ст.
8	Кровохарканье	
9	Кашель	
10	Набухание вен шеи	
11	Цианоз кожи	
12	Нарушение сознания	
13	Дыхательные нарушения*	
14	Боль в грудной клетке	
15	Тромбоз глубоких вен ног**	код МКБ-10 I80.0-I80.3
16	Злокачественные новообразования**	код МКБ-10 C00-C97
17	Сатурация	%
18	Табакокурение	
19	Частота дыхания	в мин
20	Частота сердечных сокращений	в мин
21	Отеки	
22	Антигипертензивная терапия	
23	Креатинин крови	мкмоль/л

Примечание: * — признак "дыхательные нарушения" сформирован из признаков "одышка", "одышка в покое", "одышка при физической нагрузке", "инспираторная одышка", "экспираторная одышка"; ** — признаки сформированы на основании данных о заболеваниях пациента, которые были диагностированы до даты обращения за помощью по рассматриваемому случаю лечения.

[5] и Wells [6] были валидированы и рекомендованы к практическому использованию [7], в то время как точность данных алгоритмов не признавалась высокой. Например, для шкалы Wells площадь под ROC-кривой 0,72, а валидация на проспективных данных пациентов в стратифицированных группах с низкой, средней и высокой вероятностью показала распространенность ТЭЛА 2%, 15% и 43% [8].

Важным аспектом ТЭЛА является то, что возросшая настороженность и увеличивающаяся доступность диагностических инструментальных методов, таких как компьютерная томография (КТ) легочных артерий (ЛА), привели к тому, что клиницисты сегодня подозревают ТЭЛА и инициируют соответствующее обследование чаще, чем ранее: подтверждение ТЭЛА среди пациентов, проходящих диагностическое обследование, в недавних исследованиях составило всего 5% против показанной в начале 1980-х годов распространенности, составившей ~50% [9]. Имеющийся тренд свидетельствует об увеличении финансовых затрат при подозрении на ТЭЛА и рисков осложнений диагностических процедур. Таким образом, крайне важно, чтобы современные диагностические алгоритмы для ТЭЛА были способны безопасно исключать ТЭЛА в популяциях пациентов с низкой претестовой вероятностью заболевания.

Методы машинного обучения позволяют создавать модели, способные включать в анализ неограниченное количество клинических признаков, определять их значимость для рассчитываемого события, повышать точность оценки, а работа таких моделей с данными электронных медицинских карт (ЭМК) более точно поддерживает клинические решения [10]. Эти особенности являются потенциалом как для улучшения мониторинга пациентов с высоким риском ТЭЛА и для ее более ранней диагностики, так и для безопасного исключения этого заболевания.

Целями данного исследования стали разработка и валидация модели машинного обучения для выявления подозрения на ТЭЛА по клиническим признакам из ЭМК.

Материал и методы

Источник данных

Источником информации для создания модели послужила база данных (БД) платформы прогнозной аналитики "Webiomed", содержащая деперсонифицированные формализованные данные ЭМК пациентов реальной клинической практики. Данные были собраны ретроспективно и деидентифицированы в соответствии с законами Российской Федерации о персональных данных и, следовательно, не требовали одобрения этического комитета и добровольного согласия пациентов. На обработку взяты данные 2,9 млн пациентов, связанные с ними обращения в медицинские организации амбулаторного и ста-

ционарного звена в период с 01.01.2001 до 31.12.2022, находящиеся в следующих регионах Российской Федерации: Кировская, Челябинская, Волгоградская области, Ямало-Ненецкий автономный округ, Республика Карелия, Пермский и Алтайский край. Данные о пациентах включают в себя: демографические, лабораторные и инструментальные данные, симптомы, лекарственные препараты и диагнозы.

При описании построения модели были использованы рекомендации TRIPOD (<https://bmcmmedicine.biomedcentral.com/articles/10.1186/s12916-014-0241-z/tables/1>).

Когорта пациентов для формирования набора данных

В БД были использованы зарегистрированные случаи лечения. Под случаем лечения понимаются записи ЭМК, соответствующие следующим условиям: даты начала и завершения, кодирование случая при помощи кодов МКБ-10, медицинские документы, соответствующие периоду начала и завершения случая.

Критерии первичного отбора пациентов из БД:

- возраст пациента на дату начала случая старше 14 лет;



Рис. 1. Схема отбора данных и дизайн исследования.

Сокращения: ТЭЛА — тромбоэмболия легочной артерии, ХОБЛ — хроническая обструктивная болезнь легких.

- амбулаторный или стационарный случаи лечения;
- имеющиеся заболевания, со схожей клинической картиной ТЭЛА, отобранные по заключительным диагнозам случая лечения согласно кодам и подкодам МКБ-10: пневмония (J12-J16, J18); острый бронхит (J20-J22); хронический бронхит (J41, J42); хроническая обструктивная болезнь легких (J43, J44); бронхиальная астма (J45); сердечная недостаточность (I50); ТЭЛА (I26).

Собранные случаи лечения были разделены на две группы: с ТЭЛА и без ТЭЛА (пневмония, острый бронхит, хронический бронхит; хроническая обструктивная болезнь, бронхиальная астма, сердечная недостаточность).

Подготовка данных

В набор данных (НД) были включены случаи лечения, в которых присутствовали значения признаков "возраст" и "пол"; если у уникального пациента было несколько последовательных либо одновременных случаев, то оставляли один с наибольшей заполненностью значений признаков. Для всех отобранных пациентов из БД были выгружены значения признаков на даты случая лечения и за 4 дня до даты начала случая (табл. 1). Признаки отбирались согласно знаниям о клинических проявлениях описанных выше заболеваний, присутствию их в ЭМК. Такие признаки, как иммобилизация, болезненная пальпация вен ног, хирургическая операция, отсутствовали в ЭМК и не были взяты для анализа. Если значений у признаков было несколько в этот период, то бралось самое позднее. Если значения количественных признаков оказывались вне пределов допустимых

значений, то данные значения признаков удалялись из итогового НД.

Для бинарных признаков пропущенные значения были заполнены значениями "нет". Для методов машинного обучения Logistic Regression, MLP Classifier, Gradient Boosting Classifier и Random Forest Classifier пропуски в значениях количественных признаков были заполнены фиксированным значением: -10000. Для методов машинного обучения XGBoost, CatBoost и LightGBM пропуски не заполнялись.

Общая схема подготовки данных и дизайн исследования представлены на рисунке 1.

Описание процесса разработки и оценки/валидации модели

На сформированном НД на языке программирования Python были обучены модели машинного обучения, диагностирующие ТЭЛА исходя из значений 23 признаков пациентов. Мы сравнили 7 различных методов машинного обучения к созданию моделей на основе табличных данных: Logistic Regression (логистическая регрессия), MLP Classifier (многослойный перцептон), Random Forest Classifier (алгоритм случайного леса), XGBoost, LightGBM, CatBoost, Gradient Boosting Classifier (алгоритмы градиентного бустинга).

Для методов XGBoost, LightGBM, CatBoost использовались одноименные библиотеки, для всех остальных методов библиотека Scikit-learn.

Отдельный регион был выбран для внешней валидации моделей (рис. 1) согласно рекомендациям по валидации многофакторных предиктивных моделей TRIPOD. Оставшийся НД был разделен случайным образом в соотношении 80% — выборка для обучения, 20% — для тестирования.

Таблица 2

Состав выборки для обучения, тестирования и внешней валидации

Характеристики	Обучающая выборка	Тестовая выборка	Внешняя валидация
Уникальные пациенты	10831	2708	6191
Пациенты без целевого события (класс 0)	10040	2510	5801
Пациенты с целевым событием (класс 1)	791	198	390

Для оценки точности моделей использовался ROC-анализ и основные метрики точности при пороге срабатывания, определенном для максимального значения индекса Юдена для чувствительности и специфичности. Использовались метрики точности: чувствительность (Recall), специфичность, прогностическая ценность положительного класса (Precision), прогностическая ценность отрицательного класса, F1 (положительный класс), F1 (отрицательный класс), площадь под ROC-кривой, точность.

Выбор наилучшей модели проводился по максимальному значению площади под ROC-кривой и с учетом минимального изменения оптимального порога срабатывания между тестовой выборкой и выборкой внешней валидации. Доверительные интервалы (ДИ) для метрик точности и площади под характеристической кривой (95%) были рассчитаны методом

Бутстрэп с числом итераций 1000 с помощью функции `sklearn.utils.resample`. Значимость признаков для выбранной модели определялась путем расчета относительного влияния каждой переменной на результат работы модели с помощью метода SHAP.

Статистический анализ. Статистический анализ и построение моделей машинного обучения выполняли на языке программирования Python, версия 3.9. Нормальность распределения количественных переменных оценивалась при помощи критерия Шапиро-Уилка, их данные отражены в виде медианы с указанием межквартильного размаха [Q25%; Q75%], 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 Пирсона, значение $p < 0,05$ считалось статистически значимым.

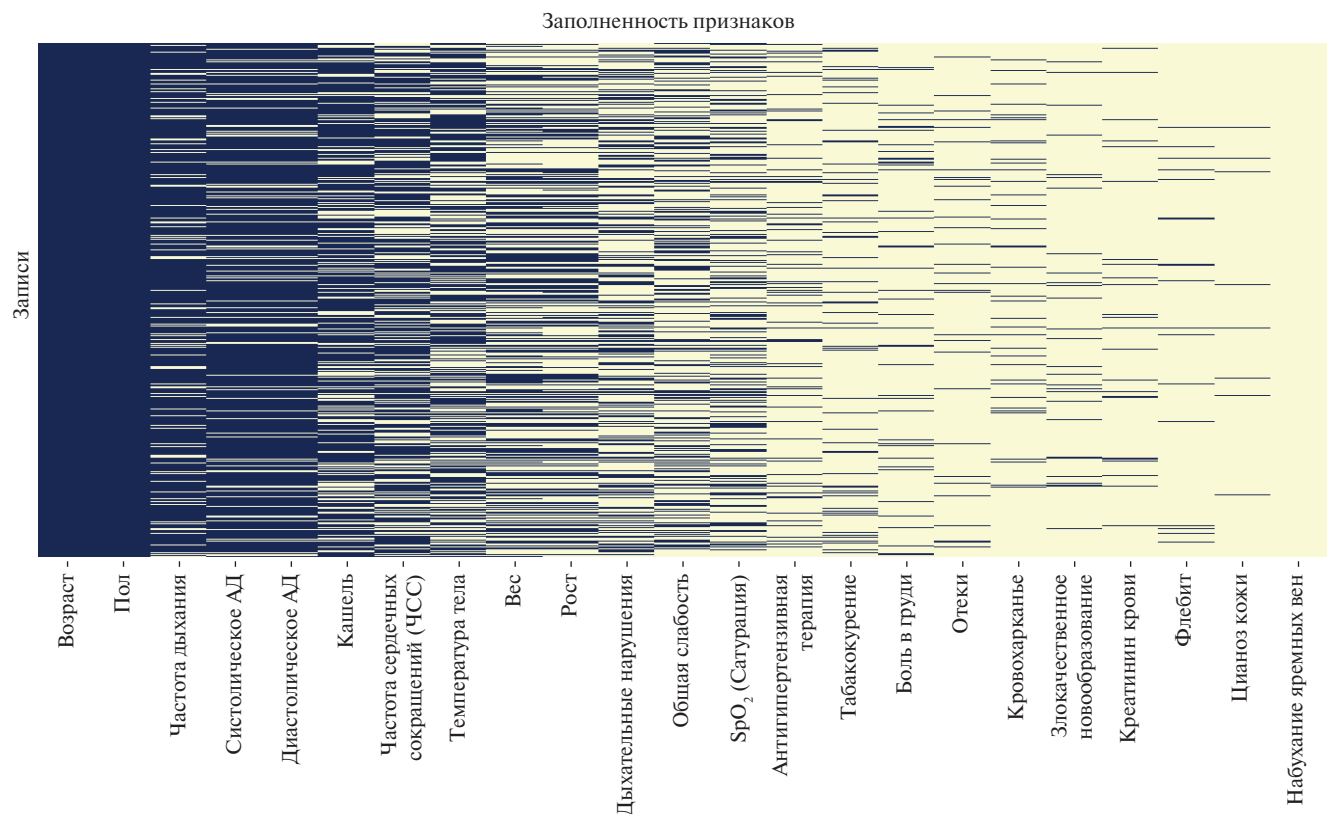


Рис. 2. Исходная заполненность значений признаков НД.

Примечание: синий цвет — значение признака заполнено, желтый — не заполнено. Цветное изображение доступно в электронной версии журнала.

Сокращения: АД — артериальное давление, SpO₂ — сатурация крови кислородом.

Таблица 3

Распространенность признаков и их значений в НД

Признак	Параметр	Класс 0 (N=18351)	Класс 1 (N=1379)	Вся выборка (N=19730)	P value
SpO ₂	Min-Мак значения	(54,0; 100,0)	(24,0; 100,0)	(24,0; 100,0)	p<0,001***
	Медиана (25, 75 процентиля)	97,0 (95,0, 98,0)	95,0 (90,0, 97,0)	97,0 (95,0, 98,0)	p<0,001**
	Число заполненных записей	4631 (25,0%)	864 (63,0%)	5495 (28,0%)	
Антигипертензивная терапия	Отрицательное значение	16405 (89,4%)	930 (67,4%)	17335 (87,9%)	p<0,001*
	Положительное значение	1946 (10,6%)	449 (32,6%)	2395 (12,1%)	
Боль в груди	Отрицательное значение	16886 (92,0%)	1064 (77,2%)	17950 (91,0%)	p<0,001*
	Положительное значение	1465 (8,0%)	315 (22,8%)	1780 (9,0%)	
Вес	Min-Мак значения	(40,0; 196,0)	(42,0; 200,0)	(40,0; 200,0)	p<0,001***
	Медиана (25, 75 процентиля)	76,0 (65,0, 88,0)	80,0 (70,0, 95,0)	76,0 (65,0, 89,0)	p<0,001**
	Число заполненных записей	7102 (39,0%)	500 (36,0%)	7602 (39,0%)	
Возраст	Min-Мак значения	(14,0; 104,0)	(17,0; 97,0)	(14,0; 104,0)	p<0,001***
	Медиана (25, 75 процентиля)	55,0 (38,0, 64,0)	63,0 (52,0, 73,0)	55,0 (39,0, 65,0)	p<0,001**
	Число заполненных записей	18351 (100,0%)	1379 (100,0%)	19730 (100,0%)	
Диастолическое АД	Min-Мак значения	(20,0; 140,0)	(20,0; 130,0)	(20,0; 140,0)	p<0,001***
	Медиана (25, 75 процентиля)	80,0 (75,0, 80,0)	80,0 (70,0, 83,5)	80,0 (70,0, 80,0)	p<0,001**
	Число заполненных записей	14779 (81,0%)	1219 (88,0%)	15998 (81,0%)	
Дыхательные нарушения	Отрицательное значение	11983 (65,3%)	377 (27,3%)	12360 (62,6%)	p<0,001*
	Положительное значение	6368 (34,7%)	1002 (72,7%)	7370 (37,4%)	
Злокачественное новообразование	Отрицательное значение	17300 (94,3%)	1238 (89,8%)	18538 (94,0%)	p<0,001*
	Положительное значение	1051 (5,7%)	141 (10,2%)	1192 (6,0%)	
Кашель	Отрицательное значение	4861 (26,5%)	960 (69,6%)	5821 (29,5%)	p<0,001*
	Положительное значение	13490 (73,5%)	419 (30,4%)	13909 (70,5%)	
Креатинин крови	Min-Мак значения	(10,0; 500,0)	(10,0; 817,0)	(10,0; 817,0)	p<0,001***
	Медиана (25, 75 процентиля)	57,5 (28,0, 82,0)	91,0 (76,0, 113,0)	78,0 (45,0, 100,0)	p<0,001**
	Число заполненных записей	396 (2,0%)	416 (30,0%)	812 (4,0%)	
Кровохарканье	Отрицательное значение	17334 (94,5%)	1165 (84,5%)	18499 (93,8%)	p<0,001*
	Положительное значение	1017 (5,5%)	214 (15,5%)	1231 (6,2%)	
Набухание яремных вен	Отрицательное значение	18351 (100,0%)	1378 (99,9%)	19729 (100,0%)	
	Положительное значение	0 (0,0%)	1 (0,1%)	1 (0,0%)	
Общая слабость	Отрицательное значение	12275 (66,9%)	719 (52,1%)	12994 (65,9%)	p<0,001*
	Положительное значение	6076 (33,1%)	660 (47,9%)	6736 (34,1%)	
Отеки	Отрицательное значение	17482 (95,3%)	812 (58,9%)	18294 (92,7%)	p<0,001*
	Положительное значение	869 (4,7%)	567 (41,1%)	1436 (7,3%)	
Пол	Женщина	11064 (60,3%)	744 (54,0%)	11808 (59,8%)	p<0,001*
	Мужчина	7287 (39,7%)	635 (46,0%)	7922 (40,2%)	
Рост	Min-Мак значения	(120,0; 200,0)	(125,0; 200,0)	(120,0; 200,0)	p<0,001***
	Медиана (25, 75 процентиля)	164,0 (158,0, 171,0)	166,0 (160,0, 174,0)	164,0 (158,0, 171,0)	p<0,001**
	Число заполненных записей	7365 (40,0%)	522 (38,0%)	7887 (40,0%)	
Систолическое АД	Min-Мак значения	(43,0; 240,0)	(40,0; 230,0)	(40,0; 240,0)	0,029
	Медиана (25, 75 процентиля)	120,0 (120,0, 130,0)	120,0 (110,0, 140,0)	120,0 (120,0, 130,0)	p<0,001**
	Число заполненных записей	14781 (81,0%)	1224 (89,0%)	16005 (81,0%)	
Табакокурение	Отрицательное значение	16305 (88,9%)	1199 (86,9%)	17504 (88,7%)	0,035
	Положительное значение	2046 (11,1%)	180 (13,1%)	2226 (11,3%)	
Температура тела	Min-Мак значения	(34,0; 41,4)	(35,0; 39,5)	(34,0; 41,4)	p<0,001***
	Медиана (25, 75 процентиля)	36,8 (36,6, 37,5)	36,6 (36,5, 37,4)	36,8 (36,6, 37,5)	p<0,001**
	Число заполненных записей	10706 (58,0%)	537 (39,0%)	11243 (57,0%)	
Флебит	Отрицательное значение	18098 (98,6%)	1103 (80,0%)	19201 (97,3%)	p<0,001*
	Положительное значение	253 (1,4%)	276 (20,0%)	529 (2,7%)	
Цианоз кожи	Отрицательное значение	18257 (99,5%)	1174 (85,1%)	19431 (98,5%)	p<0,001*
	Положительное значение	94 (0,5%)	205 (14,9%)	299 (1,5%)	
Частота дыхания	Min-Мак значения	(8,0; 65,0)	(10,0; 60,0)	(8,0; 65,0)	p<0,001***
	Медиана (25, 75 процентиля)	17,0 (16,0, 18,0)	18,0 (17,0, 20,8)	17,0 (16,0, 18,0)	p<0,001**
	Число заполненных записей	15245 (83,0%)	1130 (82,0%)	16375 (83,0%)	

Таблица 3. Продолжение

Признак	Параметр	Класс 0 (N=18351)	Класс 1 (N=1379)	Вся выборка (N=19730)	P value
Частота сердечных сокращений	Min-Max значения	(30,0; 182,0)	(30,0; 181,0)	(30,0; 182,0)	p<0,001***
	Медиана (25, 75 процентиля)	76,0 (70,0, 81,0)	87,0 (74,0, 100,0)	76,0 (70,0, 83,0)	p<0,001**
	Число заполненных записей	11045 (60,0%)	1043 (76,0%)	12088 (61,0%)	

Примечание: * — критерий согласия хи-квадрат (H0: распределение значений признаков не зависит от класса); ** — тест Шапиро-Уилка для всей выборки (H0: распределение значений признаков нормальное); *** — U-критерий Манна-Уитни (H0: распределение значений признаков не зависит от класса).

Сокращения: АД — артериальное давление, SpO₂ — сатурация крови кислородом.

Таблица 4

Метрики диагностической точности моделей на тестовой выборке

Модель	Площадь под характеристической кривой/ ROC-AUC	Доля правильных ответов (Accuracy)	Полнота (Recall)	Специфичность (Spec)	Прогностическая ценность положительного класса (Prec)	Прогностическая ценность отрицательного класса	F1 (положительный класс)	F1 (отрицательный класс)
XGBoost	0,945 [0,931; 0,958]	0,888 [0,878; 0,898]	0,864 [0,824; 0,902]	0,890 [0,881; 0,900]	0,383 [0,343; 0,419]	0,988 [0,984; 0,992]	0,531 [0,489; 0,567]	0,937 [0,931; 0,942]
LightGBM	0,954 [0,940; 0,967]	0,922 [0,914; 0,930]	0,848 [0,804; 0,886]	0,928 [0,921; 0,936]	0,483 [0,440; 0,526]	0,987 [0,983; 0,991]	0,615 [0,574; 0,655]	0,957 [0,952; 0,961]
CatBoost	0,952 [0,938; 0,964]	0,892 [0,882; 0,902]	0,884 [0,847; 0,92]	0,892 [0,883; 0,902]	0,393 [0,359; 0,430]	0,99 [0,986; 0,993]	0,544 [0,509; 0,580]	0,939 [0,933; 0,944]
Logistic Regression	0,934 [0,915; 0,949]	0,889 [0,879; 0,898]	0,833 [0,786; 0,876]	0,894 [0,883; 0,903]	0,382 [0,343; 0,420]	0,986 [0,981; 0,990]	0,524 [0,482; 0,563]	0,937 [0,932; 0,943]
MLP Classifier	0,860 [0,834; 0,886]	0,807 [0,794; 0,820]	0,763 [0,717; 0,814]	0,810 [0,797; 0,823]	0,241 [0,214; 0,268]	0,977 [0,972; 0,983]	0,366 [0,332; 0,400]	0,886 [0,878; 0,894]
Random Forest Classifier	0,945 [0,931; 0,957]	0,909 [0,899; 0,918]	0,828 [0,785; 0,872]	0,915 [0,906; 0,924]	0,435 [0,394; 0,477]	0,985 [0,981; 0,989]	0,570 [0,533; 0,610]	0,949 [0,943; 0,954]
Gradient Boosting Classifier	0,952 [0,938; 0,964]	0,877 [0,865; 0,887]	0,899 [0,864; 0,932]	0,875 [0,863; 0,885]	0,362 [0,323; 0,398]	0,991 [0,988; 0,994]	0,516 [0,474; 0,552]	0,929 [0,922; 0,935]

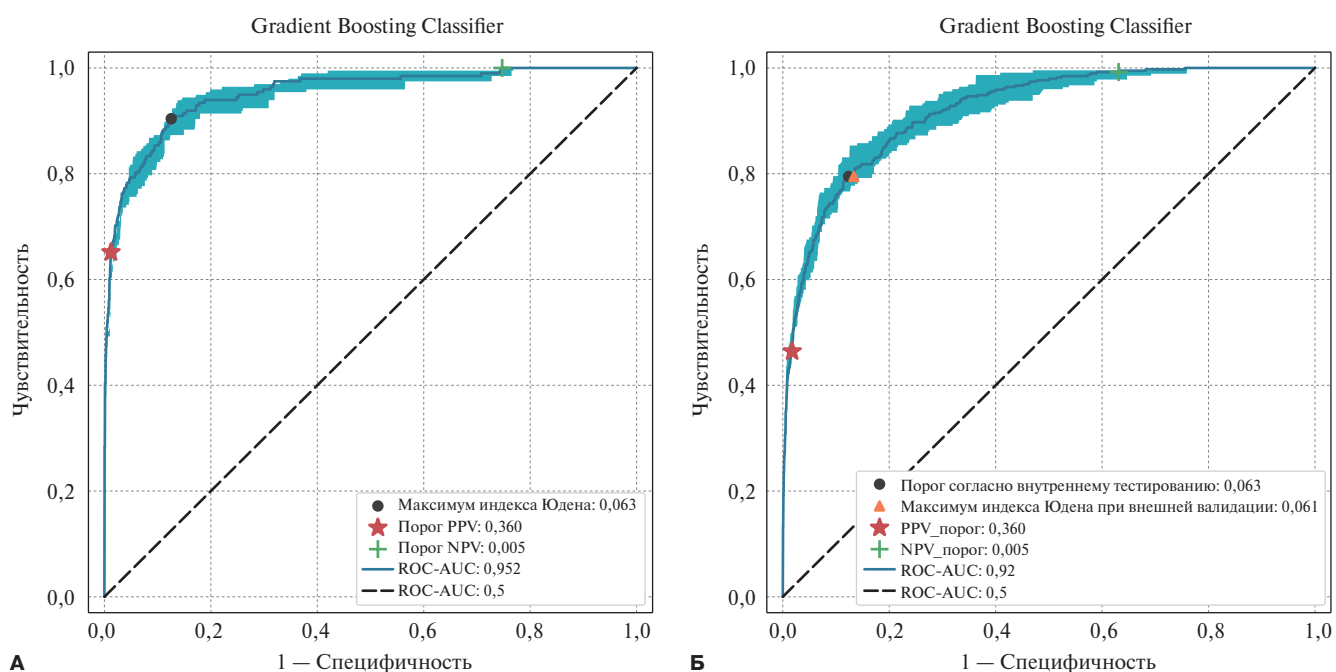


Рис. 3. ROC-кривая модели Gradient Boosting Classifier по результатам внутреннего тестирования (А) и внешней валидации (Б).

Таблица 5
Метрики диагностической точности моделей на выборке для внешней валидации

Модель	Тип порога	Значение порога	Площадь под кривой/ROC-AUC	Доля правильных ответов (Accuracy)	Полнота (Recall)	Специфичность (Spec)	Прогностическая ценность положительного класса (Prec)	Прогностическая ценность отрицательного класса	F1 (положительный класс)	F1 (отрицательный класс)
XGBoost	Порог согласно внутреннему тестированию	0,712	0,919 [0,902; 0,935]	0,891 [0,881; 0,901]	0,733 [0,677; 0,788]	0,902 [0,892; 0,911]	0,334 [0,295; 0,372]	0,981 [0,975; 0,985]	0,459 [0,414; 0,501]	0,939 [0,934; 0,945]
	Максимум индекса Юдена при внешней валидации	0,583	0,919 [0,902; 0,935]	0,84 [0,829; 0,851]	0,841 [0,795; 0,885]	0,840 [0,829; 0,852]	0,261 [0,233; 0,29]	0,987 [0,983; 0,991]	0,399 [0,362; 0,433]	0,908 [0,9; 0,915]
LightGBM	Порог согласно внутреннему тестированию	0,106	0,927 [0,911; 0,942]	0,915 [0,906; 0,924]	0,690 [0,628; 0,749]	0,930 [0,922; 0,938]	0,400 [0,352; 0,446]	0,978 [0,973; 0,983]	0,506 [0,459; 0,551]	0,954 [0,949; 0,958]
	Максимум индекса Юдена при внешней валидации	0,042	0,927 [0,911; 0,941]	0,838 [0,827; 0,849]	0,864 [0,822; 0,907]	0,836 [0,824; 0,847]	0,261 [0,233; 0,291]	0,989 [0,986; 0,993]	0,401 [0,365; 0,437]	0,906 [0,899; 0,913]
CatBoost	Порог согласно внутреннему тестированию	0,071	0,919 [0,900; 0,935]	0,876 [0,866; 0,885]	0,769 [0,719; 0,817]	0,883 [0,873; 0,893]	0,307 [0,270; 0,344]	0,983 [0,979; 0,987]	0,439 [0,396; 0,479]	0,93 [0,924; 0,936]
	Максимум индекса Юдена при внешней валидации	0,049	0,919 [0,900; 0,936]	0,831 [0,82; 0,844]	0,851 [0,808; 0,894]	0,83 [0,818; 0,842]	0,252 [0,225; 0,281]	0,988 [0,984; 0,992]	0,389 [0,354; 0,424]	0,902 [0,895; 0,910]
Logistic Regression	Порог согласно внутреннему тестированию	0,535	0,922 [0,903; 0,939]	0,905 [0,895; 0,913]	0,731 [0,677; 0,781]	0,917 [0,907; 0,925]	0,371 [0,329; 0,41]	0,981 [0,976; 0,985]	0,492 [0,446; 0,533]	0,948 [0,942; 0,952]
	Максимум индекса Юдена при внешней валидации	0,311	0,922 [0,904; 0,939]	0,83 [0,818; 0,841]	0,879 [0,838; 0,920]	0,826 [0,814; 0,838]	0,254 [0,225; 0,281]	0,99 [0,987; 0,994]	0,394 [0,357; 0,428]	0,901 [0,893; 0,908]
MLP Classifier	Порог согласно внутреннему тестированию	0,002	0,856 [0,825; 0,885]	0,833 [0,822; 0,845]	0,741 [0,685; 0,796]	0,839 [0,828; 0,850]	0,237 [0,207; 0,268]	0,98 [0,975; 0,984]	0,359 [0,321; 0,398]	0,904 [0,897; 0,911]
	Максимум индекса Юдена при внешней валидации	0,005	0,856 [0,824; 0,885]	0,872 [0,862; 0,883]	0,713 [0,657; 0,767]	0,883 [0,873; 0,893]	0,291 [0,257; 0,326]	0,979 [0,974; 0,983]	0,413 [0,375; 0,455]	0,928 [0,922; 0,934]
Random Forest Classifier	Порог согласно внутреннему тестированию	0,555	0,909 [0,89; 0,926]	0,896 [0,887; 0,906]	0,677 [0,614; 0,734]	0,911 [0,902; 0,920]	0,338 [0,296; 0,379]	0,977 [0,971; 0,982]	0,451 [0,405; 0,494]	0,943 [0,937; 0,948]
	Максимум индекса Юдена при внешней валидации	0,453	0,909 [0,89; 0,925]	0,854 [0,842; 0,865]	0,813 [0,761; 0,859]	0,857 [0,844; 0,868]	0,276 [0,242; 0,308]	0,986 [0,981; 0,99]	0,412 [0,369; 0,449]	0,917 [0,909; 0,923]
Gradient Boosting Classifier	Порог согласно внутреннему тестированию	0,063	0,920 [0,904; 0,936]	0,871 [0,861; 0,882]	0,795 [0,745; 0,843]	0,877 [0,867; 0,887]	0,302 [0,268; 0,339]	0,985 [0,98; 0,988]	0,438 [0,398; 0,480]	0,927 [0,921; 0,934]
	Максимум индекса Юдена при внешней валидации	0,061	0,920 [0,903; 0,936]	0,862 [0,851; 0,872]	0,810 [0,762; 0,856]	0,865 [0,854; 0,876]	0,288 [0,256; 0,320]	0,985 [0,981; 0,989]	0,424 [0,386; 0,463]	0,921 [0,915; 0,928]

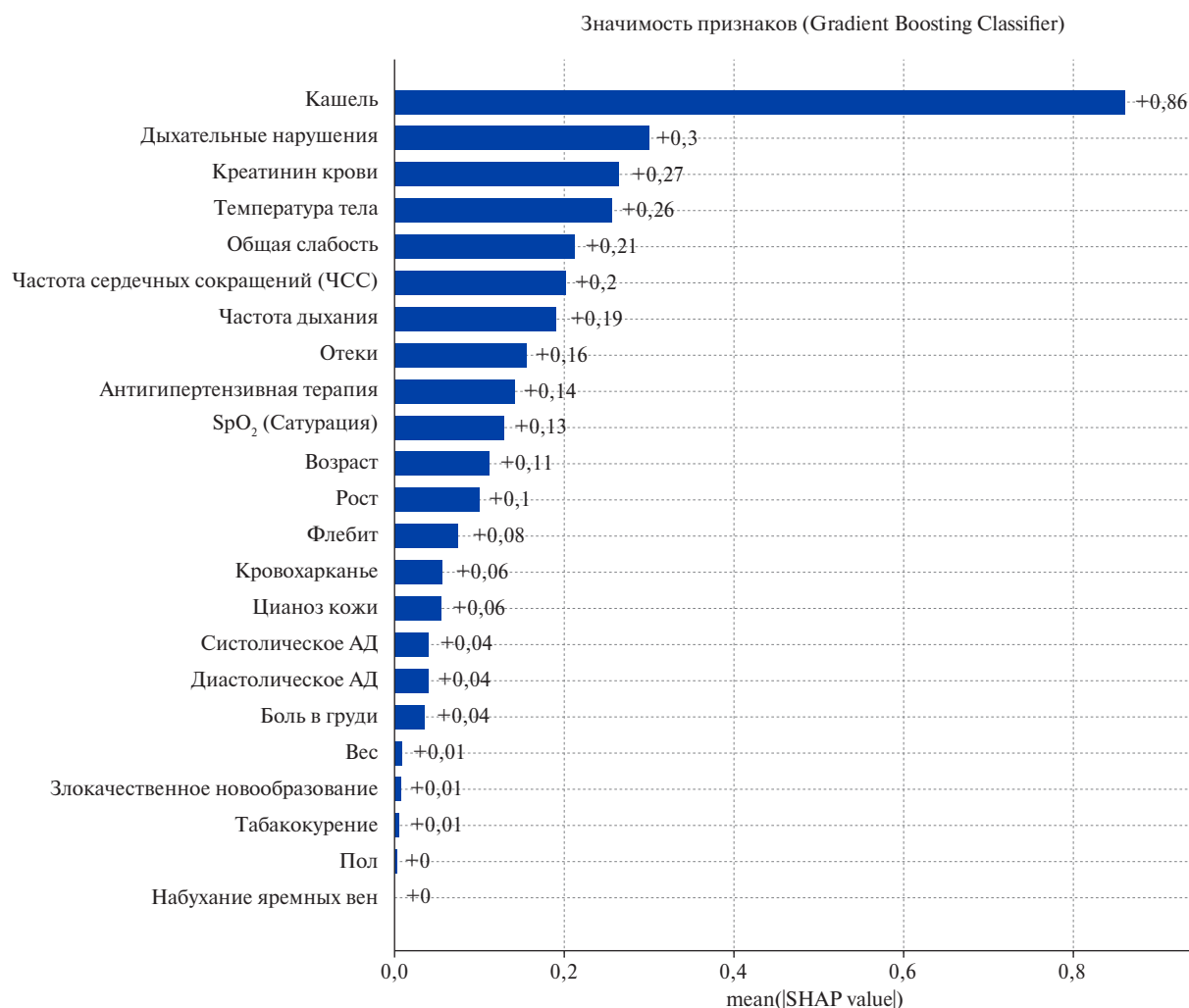


Рис. 4. Значимость признаков у модели Gradient Boosting Classifier.

Сокращения: АД — артериальное давление, SpO₂ — сатурация крови кислородом.

Результаты

Данные пациентов, вошедших в итоговый набор, были получены в период с марта 2007 по февраль 2022гг. Всего модели обучались на данных 19730 пациентов. В обучающей выборке пациентов с целевым событием было 7,3% (791 из 10831), в тестовой — 7,3%, валидационной — 6,3%. Состав НД с указанием разделения на обучающую, тестовую выборки и выборку для внешней валидации приведен в таблице 2. В среднем, пациенты перенесшие ТЭЛА были старше, женского пола, чаще у них встречались кровохарканье, симптомы дыхательных нарушений, цианоз кожи и отеки, диагностированы злокачественные новообразования (ЗНО) и тромбофлебит, имели повышенный креатинин крови, частоту сердечных сокращений. Полное описание признаков НД для обучения приведено в таблице 3. Такие признаки, как симптомы, диагнозы, являющиеся булевыми, половозрастные характеристики, являющи-

еся обязательными, имеют полную заполненность в НД. Количественные признаки, такие как измеряемые параметры при осмотре и лабораторные, имеют различную распространенность, меньшую чем 100%. Заполненность значений признаков итогового НД представлена на рисунке 2.

Были применены 7 алгоритмов машинного обучения для построения моделей. Основные метрики диагностической точности вместе с 95% ДИ для всех исследованных моделей приведены в таблицах 4 (тестовая выборка) и 5 (внешняя валидация). Из всех исследованных моделей наименьшую точность на одном и том же НД показала модель MLP Classifier. Остальные модели, а именно LightGBM, Random Forest Classifier и Logistic Regression, CatBoost, XGBoost и Gradient Boosting Classifier показали сравнимые метрики, и значения площади под ROC-кривой пересекались в пределах ДИ. Диапазон значений площади под ROC-кривой у этих моделей

на тестовой выборке составил (0,934-0,954), а на выборке внешней валидации (0,909-0,927).

Наименьшую разницу в порогах срабатывания (0,002) на тестовой выборке и при внешней валидации показала модель на основе алгоритма Gradient Boosting Classifier, поэтому данная модель была выбрана наиболее подходящей для проспективной апробации.

ROC-кривая модели Gradient Boosting Classifier представлена на рисунке 3. Выход модели представляет собой число от 0 до 1 — вероятность наличия заболевания ТЭЛА у пациента. Порог классификации для модели 0,063. Если выход модели ниже отметки 0,063, то вероятность наличия ТЭЛА у пациента низкая, если выше — вероятность наличия ТЭЛА у пациента высокая. Порог классификации был подобран с помощью максимума индекса Юдена. Значимость признаков для модели, построенной с помощью Gradient Boosting Classifier, приведена на рисунке 4.

Обсуждение

В проведенном исследовании нами продемонстрирована способность моделей машинного обучения выявлять пациентов высокого риска возникновения ТЭЛА. Хотя для этой задачи было использовано 7 различных подходов, алгоритм градиентного бустинга работал наиболее точно, что, возможно, связано с лучшей обработкой этим алгоритмом пропущенных данных, которые часто встречались в используемом НД. Вполне возможно, что другие нейронные сети могут быть более приспособлены к этой задаче, чем многослойная модель персептрона, используемая в настоящем исследовании.

При планировании данной модели изначально предполагалась поддержка клинического решения для выявления ТЭЛА на этапе до назначения дополнительных методов исследования: Д-димера, КТ ЛА. Для решения этой задачи исследованы признаки в момент начала амбулаторного или стационарного обращения, включающие в себя жалобы, данные объективного осмотра, анамнеза, ранее имеющиеся заболевания, лабораторные данные. По результатам проведенного машинного обучения определены признаки, которые имеют наибольшую значимость для предсказания ТЭЛА: кашель, дыхательные нарушения, креатинин крови, температура тела, общая слабость, отеки ног, частота сердечных сокращений, частота дыхания. Все эти признаки согласуются с распространенными клиническими проявлениями ТЭЛА [7]. К признакам с меньшей значимостью относятся сатурация, возраст, тромбоз, флебит, кровохарканье, цианоз кожи, систолическое и диастолическое артериальное давление (АД), ЗНО. Систолическое АД и ЗНО традиционно считаются значимыми факторами и проявлениями ТЭЛА, но на наших данных не было получено сильное их

влияние на целевое событие. Мы можем объяснить низкую роль систолического АД особенностями введения данных в ЭМК, когда возможны ошибки введения информации о давлении или тиражирование шаблона врачебного осмотра без коррекции значений параметров. Несильное влияние ЗНО объясняется их незначительной частотой в НД для обучения (5,7%) и отсутствием анализа по виду рака, что имеет значение для риска тромбоза. В нашем исследовании использован НД для обучения со схожими по клинической картине заболеваниями, такими как пневмония, острый бронхит, хронический бронхит; хроническая обструктивная болезнь, бронхиальная астма, сердечная недостаточность с той целью, что модель более точно определит связь смежных симптомов с ТЭЛА. Это позволяет использовать модель на раннем этапе диагностического поиска ТЭЛА: сбор жалоб, анамнеза, объективных данных.

Опубликовано немного моделей машинного обучения и исследований, связанных с использованием данных ЭМК для диагностики ТЭЛА. В исследовании Logan R, et al. [11] была разработана модель предсказания ТЭЛА на данных ЭМК госпитализированных пациентов, и аналогично нашим данным алгоритм градиентного бустинга показал свое превосходство (площадь под ROC-кривой 0,85). Особенности этого исследования были: сбор данных, ограниченный первыми 3 ч пребывания в стационаре, использование данных одного центра, анализ анамнестических, расширенных лабораторных и ограниченных объективных данных; верификация ТЭЛА на основании кодов МКБ-10 и приема антикоагулянтов или проведения тромболитической терапии. В нашем исследовании ТЭЛА была верифицирована только при помощи кодов МКБ-10, что объясняется ограниченными возможностями БД. Коды МКБ-10 для верификации ТЭЛА имеют высокую чувствительность (~90%) [12], но риск ложноположительных результатов при таком подходе все же имеется, а использование дополнительных критериев верификации диагноза ТЭЛА позволяет уменьшить этот риск. Особенностью исследования van Es N, et al. [13] было то, что использовались данные 16 проспективных исследований с 28 тыс. пациентов для построения модели индивидуальной вероятности ТЭЛА. По результатам обучения были взяты 9 предикторов: возраст, пол, предыдущие венозные тромбозы, хирургические вмешательства/иммобилизация, кровохарканье, ЗНО, нахождение в стационаре, уровень Д-димера. Обученная модель показала точность по статистическому показателю 0,87. Авторы сделали вывод, что модель можно использовать в качестве альтернативы традиционным алгоритмам для принятия решения о необходимости визуализации ЛА. В исследовании Somani SS, et al. [14] были разработаны 3 модели для диагностики ТЭЛА: по данным электрокардиогра-

фии (ЭКГ), ЭМК и гибридная, объединяющая данные 1 и 2 модели. Гибридная модель (0,81 — площадь под ROC-кривой) превзошла по точности как модели ЭКГ (0,59) и ЭМК (0,65), так и шкалы, и Geneva (AUROC 0,50–0,58). В исследовании Banerjee I, et al. [15] была разработана модель машинного обучения PERFORM, которая учитывает множество факторов риска ЭМК у пациентов с подозрением на ТЭЛА. Модель преобразует необработанные данные ЭМК во временные векторы признаков и моделирует решение для отбора и направления на КТ ЛА. Для обучения и валидации было использовано 3397 КТ-исследований по поводу ТЭЛА у 3214 пациентов. Модели прошли внешнюю валидацию на 240 пациентах. Наилучшая точность модели по площади под ROC-кривой 0,90 (95% ДИ: 0,87–0,91) на внутренних данных и 0,71 (95% ДИ: 0,69–0,72) на внешних.

Разработанная нами модель имеет ряд ограничений. Данные врачебных записей в ЭМК могут иметь ошибки интерпретации и ввода клинических данных, прежде всего жалоб. Их влияние на прогноз сейчас не известно и будет проверяться при эксплуатации модели. Не были учтены некоторые значимые в патогенезе ТЭЛА признаки (иммобилизация, операции) из-за отсутствия их в ЭМК или невозможности извлечения. Для извлечения признаков из медицинских текстов были использованы NLP-модели, которые имеют собственную точность работы, колеблющуюся от 0,80 до 0,99, что может влиять на правильность извлечения признаков и дальнейшую их обработку. При обучении модели не использовались ЭКГ данные, ко-

торые могут увеличить результативность работы данной модели. Поэтому имеются перспективы для повышения точности обученной модели — дообучение на новых данных, использование дополнительных предикторов. Перспективным с клинической точки зрения является сравнение предсказательной способности разработанной модели и шкал, применяемых для выявления подозрения на ТЭЛА, являющихся стандартом клинической практики.

Заключение

Была разработана модель для оценки вероятности ТЭЛА на российских данных ЭМК с использованием технологии машинного обучения. Точность модели по метрике площадь под ROC-кривой 0,952 (95% ДИ: 0,938–0,964) согласно внутреннему тестированию и 0,920 (95% ДИ: 0,903–0,936) на внешней валидации. Обученная модель рассчитана для использования при первичном обращении за медицинской помощью пациентов с жалобами и подозрением на ТЭЛА вне зависимости от вида помощи. Использование комбинации моделей машинного обучения (анализ данных ЭМК, рентгеновских и электрокардиографических изображений) позволит смоделировать полноценный диагностический процесс ТЭЛА или выдвинуть альтернативную диагностическую гипотезу.

Отношения и деятельность: все авторы заявляют об отсутствии потенциального конфликта интересов, требующего раскрытия в данной статье.

Литература/References

1. Raskob GE, Angchaisuksiri P, Blanco AN, et al. Thrombosis: a major contributor to global disease burden. *Arterioscler Thromb Vasc Biol.* 2014;34:2363–71. doi:10.1161/ATV.114.304488.
2. Beckman MG, Hooper WC, Critchley SE, Ortel TL. Venous thromboembolism: a public health concern. *Am J Prev Med.* 2010;38:495–501. doi:10.1016/j.amepre.2009.12.017.
3. Cohen AT, Agnelli G, Anderson FA, et al. VTE Impact Assessment Group in Europe (VITAE). Venous thromboembolism (VTE) in Europe. The number of VTE events and associated morbidity and mortality. *Thromb Haemost.* 2007;98:756–64.
4. Pollack CV, Schreiber D, Goldhaber SZ, et al. Clinical characteristics, management, and outcomes of patients diagnosed with acute pulmonary embolism in the emergency department: initial report of EMPEROR (Multicenter Emergency Medicine Pulmonary Embolism in the Real World Registry). *J Am Coll Cardiol.* 2011;57:700–6. doi:10.1016/j.jacc.2010.05.071.
5. Grégoire Le Gal, Marc Righini, Pierre-Marie Roy, et al. Prediction of pulmonary embolism in the emergency department: the revised Geneva score. *Ann Intern Med.* 2006;144(3):165–71. doi:10.7326/0003-4819-144-3-200602070-00004.
6. Wells PS, Anderson DR, Rodger M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. *Thromb Haemost.* 2000;83(3):416–20.
7. Konstantinides SV. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). *Russian Journal of Cardiology.* 2020;25(8):3848. (In Russ.) Konstantinides SV. Рекомендации ESC по диагностике и лечению острой легочной эмболии, разработанные в сотрудничестве с Европейским респираторным обществом (ERS), 2019. *Российский кардиологический журнал.* 2020;25(8):3848. doi:10.15829/1560-4071-2020-3848.
8. Wolf SJ, McCubbin TR, Feldhaus KM, et al. Prospective validation of Wells Criteria in the evaluation of patients with suspected pulmonary embolism. *Ann Emerg Med.* 2004;44(5):503–10. doi:10.1016/j.annemergmed.2004.04.002.
9. Righini M, Robert-Ebadi H, Le Gal G. Diagnosis of acute pulmonary embolism. *J Thromb Haemost.* 2017;15:1251–61. doi:10.1111/jth.13694.
10. Adlung L, Cohen Y, Mor U, Elinav E. Machine learning in clinical decision making. *Med* 2021;2(6):642–65. doi:10.1016/j.medj.2021.04.006.
11. Ryan L, Maharjan J, Mataraso S, et al. Predicting pulmonary embolism among hospitalized patients with machine learning algorithms *Pulmonary Circulation.* 2022;12:e12013. doi:10.1002/pul2.12013.
12. Burles K, Innes G, Senior K, et al. Limitations of pulmonary embolism ICD-10 codes in emergency department administrative data: let the buyer beware. *BMC Med Res Methodol.* 2017;17:89. doi:10.1186/s12874-017-0361-1.
13. van Es N, Takada T, Kraaijspoel N, et al. Diagnostic management of acute pulmonary embolism: a prediction model based on a patient data meta-analysis. *European Heart Journal.* 2023;44(32):3073–81. doi:10.1093/eurheartj.
14. Somani SS, Honarvar H, Narula S, et al. Development of a machine learning model using electrocardiogram signals to improve acute pulmonary embolism screening. *European Heart Journal — Digital Health.* 2022;3(1):56–66. doi:10.1093/ehjdh/ztab101.
15. Banerjee I, Sofela M, Yang J, et al. Development and Performance of the Pulmonary Embolism Result Forecast Model (PERFORM) for Computed Tomography Clinical Decision Support. *JAMA Netw Open.* 2019;2(8):e198719. doi:10.1001/jamanetworkopen.2019.8719.