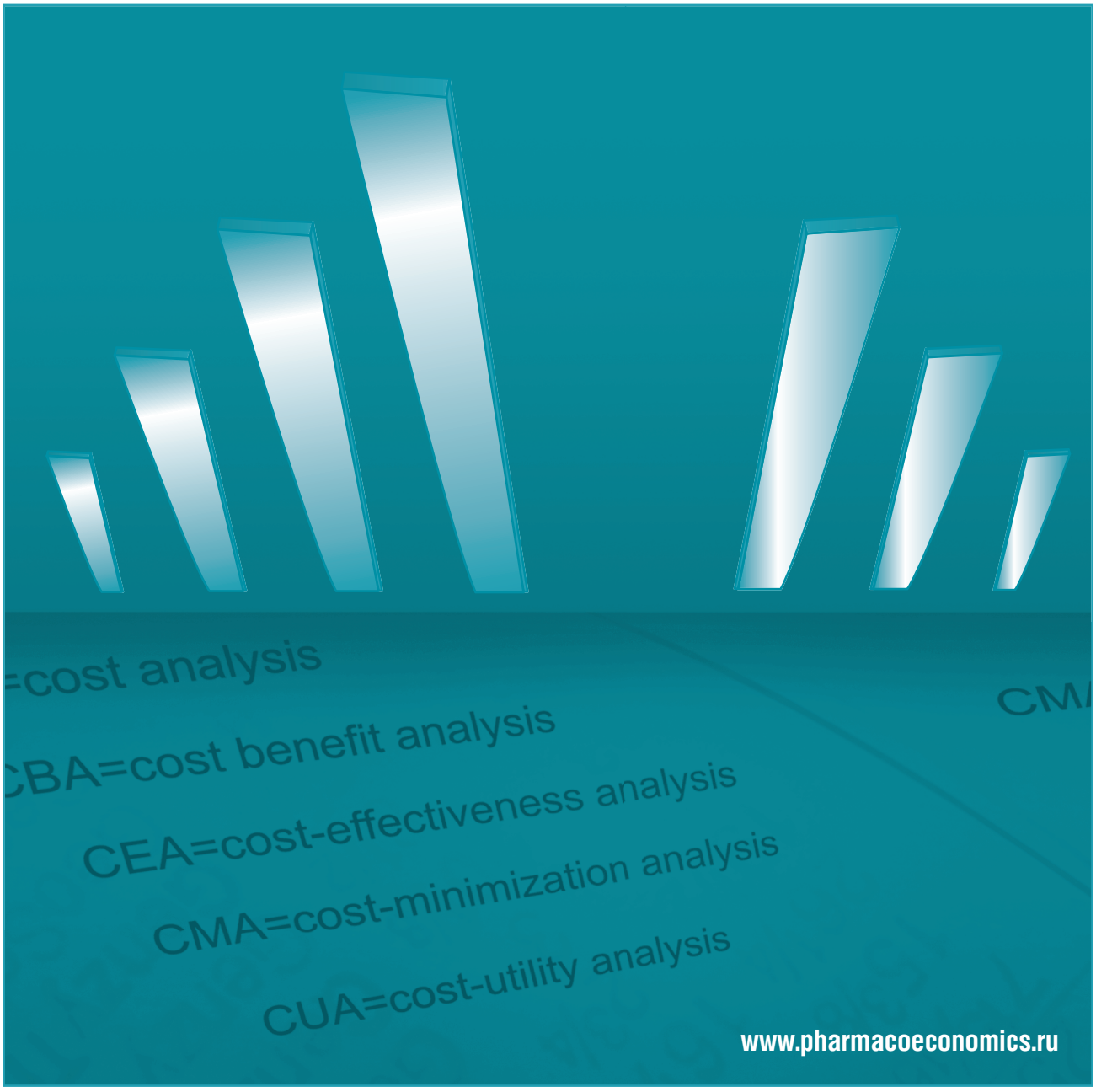
ISSN 2070-4909 (print)
ISSN 2070-4933 (online)

Фармакоэкономика

Современная фармакоэкономика и фармакоэпидемиология



FARMAKOEKONOMIKA

Modern Pharmacoeconomics and Pharmacoepidemiology

2021 Vol. 14 No. 4

№4

Том 14

2021

https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115

ISSN 2070-4909 (print)
ISSN 2070-4933 (online)

Машинное обучение на лабораторных данных для прогнозирования заболеваний

Гусев А.В.[1], Новицкий Р.Э.[1], Ившин А.А.[2], Алексеев А.А.[1]

[1] *Общество с ограниченной ответственностью «К-Скай» (наб. Варкауса, д. 17, пом. 62, Республика Карелия, Петрозаводск 185031, Россия)*

[2] *Федеральное государственное бюджетное образовательное учреждение высшего образования «Петрозаводский государственный университет» (пр-т Ленина, д. 33, Республика Карелия, Петрозаводск 185910, Россия)*

Для контактов: *Ившин Александр Анатольевич, e-mail: scipeople@mail.ru*

РЕЗЮМЕ

Цель: провести обзор отечественной и зарубежной литературы по проблеме применения методов машинного обучения в медицинских информационных системах (МИС) с использованием лабораторных данных, проанализировать точность и эффективность исследуемых технологий, их преимущества и недостатки, возможности внедрения в клиническую практику.

Материал и методы. Поиск литературы осуществляли в базах данных PubMed/MEDLINE за период с 2000 по 2020 гг. (по группам ключевых словосочетаний “machine learning”, “laboratory data”, “clinical events”, “prediction diseases”), КиберЛенинка («машинное обучение», «лабораторные данные», «клинические события», «прогнозирование заболеваний») и Papers With Code (“clinical events”, “prediction diseases”, “electronic health record”). После изучения полного текста 30 литературных источников, соответствующих критериям отбора, выбрано 19 статей, наиболее релевантных поставленной задаче.

Результаты. Выполнен анализ источников, описывающих применение технологий искусственного интеллекта для получения предиктивной аналитики с учетом доступных в МИС сведений о пациентах – демографических, анамнестических и лабораторных данных, данных инструментальных исследований, сведений об имеющихся и ранее перенесенных заболеваниях. Рассмотрены существующие способы прогнозирования неблагоприятных медицинских исходов с помощью методов машинного обучения, а также представлена информация о значимости используемых лабораторных данных для построения высокоточных предиктивных математических моделей.

Заключение. Внедрение алгоритмов машинного обучения в МИС представляется перспективным инструментом эффективного прогнозирования неблагоприятных медицинских событий для широкого применения в реальной клинической практике, что соответствует общемировой тенденции по развитию персонифицированной, основанной на расчете индивидуального риска медицины. Наблюдается рост активности исследований в области прогнозирования неинфекционных заболеваний с использованием технологий искусственного интеллекта.

КЛЮЧЕВЫЕ СЛОВА

Искусственный интеллект, прогнозирование, медицинские информационные системы, машинное обучение, нейронные сети, алгоритмы, факторы риска.

Статья поступила: 28.09.2021 г.; **в доработанном виде:** 02.11.2021 г.; **принята к печати:** 23.11.2021 г.; **опубликована онлайн:** 25.11.2021 г.

Конфликт интересов

Авторы заявляют об отсутствии необходимости раскрытия конфликта интересов в отношении данной публикации.

Финансирование

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Соглашения № 075-15-2021-665.

Вклад авторов

Все авторы сделали эквивалентный вклад в подготовку публикации.

Для цитирования

Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний. *ФАРМАКОЭКОНОМИКА. Современная фармакоэкономика и фармакоэпидемиология*. 2021; 14 (4): 581–592. https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115.

Machine learning based on laboratory data for disease prediction

Gusev A.V.[1], Novitskiy R.E.[1], Ivshin A.A.[2], Alekseev A.A.[1]

[1] K-SkAI LLC (17 Emb. Varkausa, premises 62, Republic of Karelia, Petrozavodsk 185031, Russia)

[2] Petrozavodsk State University (33 Lenin Ave., Republic of Karelia, Petrozavodsk 185910, Russia)

Corresponding author: *Aleksandr A. Ivshin, e-mail: scipeople@mail.ru*

SUMMARY

Objective: to review domestic and foreign literature on the issue of machine learning methods applied in medical information systems (MIS) using laboratory data, to analyze the accuracy and efficiency of the technologies under study, their advantages and disadvantages, the possibilities of implementation in clinical practice.

Material and methods. The literature search was performed in the PubMed/MEDLINE databases covering the period from 2000 to 2020 (using groups of keyphrases: "machine learning", "laboratory data", "clinical events", "prediction diseases"), CyberLeninka ("machine learning", "laboratory data", "clinical events", "prediction diseases" Russian keyphrases combinations) and Papers With Code ("clinical events", "prediction diseases", "electronic health record"). After reviewing the full text of 30 literature sources that met the selection criteria, the 19 most relevant articles were selected.

Results. An analysis of sources that describe the application of artificial intelligence techniques to obtain predictive analytics, taking into account information about patients, such as demographic, anamnestic, and laboratory data, the data of instrumental studies, information about existing and former diseases available in MIS, was performed. The existing ways of predicting adverse medical outcomes using machine learning methods were considered. Information about the significance of the used laboratory data for constructing high-precision predictive mathematical models is presented.

Conclusion. Implementation of machine learning algorithms in MIS seems to be a promising tool for effective prediction of adverse medical events for wide application in real clinical practice. It corresponds to the global trend in the development of personalized medicine based on the calculation of individual risk. There is an increase in the activity of research in the field of predicting noncommunicable diseases using artificial intelligence technologies.

KEYWORDS

Artificial intelligence, prediction, medical information systems, machine learning, neural networks, algorithms, risk factors.

Received: 28.09.2021; **in the revised form:** 02.11.2021; **accepted:** 23.11.2021; **published online:** 25.11.2021

Conflict of interests

The authors declare they have nothing to disclose regarding the conflict of interests with respect to this manuscript.

Funding

The research was carried out with the financial support of the Ministry of Science and Higher Education of the Russian Federation under the Agreement No. 075-15-2021-665.

Author's contrubution

The authors contributed equally to this article.

For citation

Gusev A.V., Novitskiy R.E., Ivshin A.A., Alekseev A.A. Machine learning based on laboratory data for disease prediction. *FARMAKOEKONOMIKA. Sovremennaya farmakoekonomika i farmakoepidemiologiya / FARMAKOEKONOMIKA. Modern Pharmacoeconomics and Pharmacoepidemiology*. 2021; 14 (4): 581–592 (in Russ.). https://doi.org/10.17749/2070-4909/farmakoekonomika.2021.115.

ВВЕДЕНИЕ / INTRODUCTION

В настоящее время идет активное внедрение методов машинного обучения (МО) в медицинские информационные системы (МИС). В первую очередь, это обусловлено необходимостью анализа большого объема информации о пациентах в режиме реального времени (сведений о динамическом изменении их состояния здоровья, базирующихся на результатах лабораторных и инструментальных методов исследования), а также прогнозирования обращения за амбулаторной помощью или госпитализации в течение заданного временного интервала.

Согласно докладу Всемирной организации здравоохранения (ВОЗ) была разработана глобальная система мониторинга и оценки факторов риска, профилактики и лечения неинфекционных заболеваний (НИЗ). Основные группы НИЗ представлены сердечно-сосудистыми заболеваниями (ССЗ), онкологическими патологиями, хроническими болезнями легких и сахарным диабетом. В 60% случаев данные заболевания являются причиной смерти человека. В рамках действующей стратегии ВОЗ планирует к 2025 г. снизить смертность от НИЗ на 25% [1].

Для решения задач прогнозирования заболеваний, в частности НИЗ, и особенностей их течения все чаще применяются такие методы МО, как искусственные нейронные сети [2]. Эффективные инструменты прогнозирования обращения за медицинской амбулаторной помощью и необходимости госпитализации позволяют своевременно и всесторонне оценить риск имеющихся болезней и новых заболеваний, что способствует выбору наиболее рациональной врачебной тактики в части профилактических мер или своевременному лечению болезни на ранних стадиях ее развития [3, 4]. Таким образом, снижается риск осложнений и преждевременной смерти.

Врачам-клиницистам важно иметь в своем арсенале диагностических средств максимально точный и эффективный инструмент прогнозирования, базирующийся на доступной информации о пациентах: демографических и анамнестических данных, сведениях о ранее перенесенных заболеваниях, результатах

Основные моменты

Что уже известно об этой теме?

- Согласно глобальной стратегии Всемирной организации здравоохранения (ВОЗ), к 2025 г. планируется на 25% снизить показатель смертности от неинфекционных заболеваний (НИЗ), к которым относятся сердечно-сосудистые, онкологические заболевания, сахарный диабет и хронические заболевания легких
- Поиски эффективных инструментов прогнозирования НИЗ вызваны стремлением выявить пациентов высокого риска с целью как можно раньше принять необходимые меры профилактики и лечения и таким образом снизить показатели смертности
- Активное внедрение методов машинного обучения в медицинские информационные системы обусловлено необходимостью анализа большого объема данных о пациентах в режиме реального времени

Что нового дает статья?

- Обзор направлен на информирование широкого круга специалистов в области медицины и цифровых технологий о достижениях в использовании алгоритмов машинного обучения в различных областях медицины
- Освещены возможности применения алгоритмов машинного обучения в различных областях медицины
- Проведен анализ алгоритмов, наиболее популярных и эффективных при работе с медицинскими информационными системами

Как это может повлиять на клиническую практику в обозримом будущем?

- Прогнозирование в цифровом формате откроет новые возможности для повышения точности расчета индивидуального риска НИЗ, что соответствует глобальной стратегии ВОЗ
- Выявление пациентов из группы высокого риска позволит рационально применять меры профилактики на амбулаторном этапе, определять показания для госпитализации, назначать необходимый объем терапии и своевременно распознавать развитие осложнений, что приведет к снижению демографических и экономических потерь

Highlights

What is already known about the subject?

- According to the global strategy of the World Health Organization (WHO), it is planned to reduce by 25% the mortality rate from noncommunicable diseases (NCDs) (cardiovascular, oncology, diabetes mellitus and chronic lung diseases) by 2025
- Search for effective tools to predict NCDs is caused by the desire to identify high risk patients in order to take the necessary preventive and treatment measures as early as possible and thus reduce complications and mortality
- The need to analyze a large volume of patient data in real-time determines the active implementation of machine learning methods in medical information systems

What are the new findings?

- The review was carried out to inform a wide range of specialists in medicine and digital technologies about the achievements in the use of machine learning algorithms in various fields of medicine
- The possibilities of using machine learning algorithms in various fields of medicine are highlighted
- The analysis of the algorithms that are most popular and effective when working with medical information systems was fulfiled

How might it impact the clinical practice in the foreseeable future?

- The digital format of prediction can improve the accuracy of calculating the personal risk of NIDs, which corresponds to the WHO global strategy
- The detection of patients from the high-risk group will ensure the rational use of preventive measures at the outpatient stage, the detection of indications for hospitalization, the appointment of the necessary therapy, which will reduce demographic and economic losses

лабораторных и инструментальных методов исследований. Перечисленные биомедицинские данные широко представлены в медицинских информационных системах. Ожидается, что алгоритмы МО, встроенные в МИС, позволят добиться высокой точности прогноза неблагоприятных медицинских событий. Вместе с тем внедрение предиктивного инструмента в медицинские информационные системы позволит врачам получить персонифицированное заключение о прогнозе возникновения заболевания или его осложнения непосредственно в процессе работы с МИС. В связи с этим представляется целесообразным рассмотреть существующие способы прогнозирования заболеваний и осложнений, оценки динамического изменения состояния здоровья, а также информацию о значимости используемых для прогноза лабораторных данных.

Цель – провести обзор отечественной и зарубежной литературы по проблеме применения методов машинного обучения в медицинских информационных системах с использованием лабораторных данных, проанализировать точность и эффективность исследуемых технологий, их преимущества и недостатки, возможности внедрения в клиническую практику.

МАТЕРИАЛ И МЕТОДЫ / MATERIAL AND METHODS

Поиск литературы осуществляли в следующих базах данных: PubMed/MEDLINE [5], КиберЛенинка [6], Papers With Code [7].

Критерии отбора литературных источников определялись, в первую очередь, релевантностью запросов по данной теме, наличием в научных статьях описания применения алгоритмов МО с целью выявления подозрений на заболевания, сравнения результатов работы алгоритмов с данными лабораторных исследований у пациентов. Схема дизайна обзора литературы представлена на **рисунке 1**.

В PubMed/MEDLINE выполнены запросы по группам ключевых словосочетаний: “machine learning”, “laboratory data”, “clinical events”, “prediction diseases” за период с 2000 по 2020 гг. По указанным запросам найдено 499 статей. После ознакомления с их аннотациями методом интеллектуального анализа отобрано 20 статей, соответствующих критериям отбора. Затем проведено изучение полного текста каждой статьи и выбор для обзора источников, наиболее релевантных задаче.

В КиберЛенинке (в разделе компьютерных и информационных наук) также были запрошены статьи по группам словосочетаний: «машинное обучение», «лабораторные данные», «клинические события», «прогнозирование заболеваний». Всего найдено 57 статей, после отбора в соответствии с вышеперечисленными критериями для изучения полного текста выбрано 4 источника.

В Papers With Code словосочетания использовались в качестве отдельных запросов: “clinical events”, “prediction diseases”, “electronic health record”. Это обусловлено тем, что в данной базе содержатся статьи только по теме машинного обучения и анализа данных и не возникает необходимости использовать фразу “machine learning” в качестве запроса. Также следует отметить, что в Papers With Code вместе с полнотекстовыми работами в открытом доступе опубликованы «исходники» программ по задачам, описанным в статьях. Найдено 30 статей, 6 из которых соответствовали критериям отбора и выбраны для изучения их полного текста и программных «исходников».

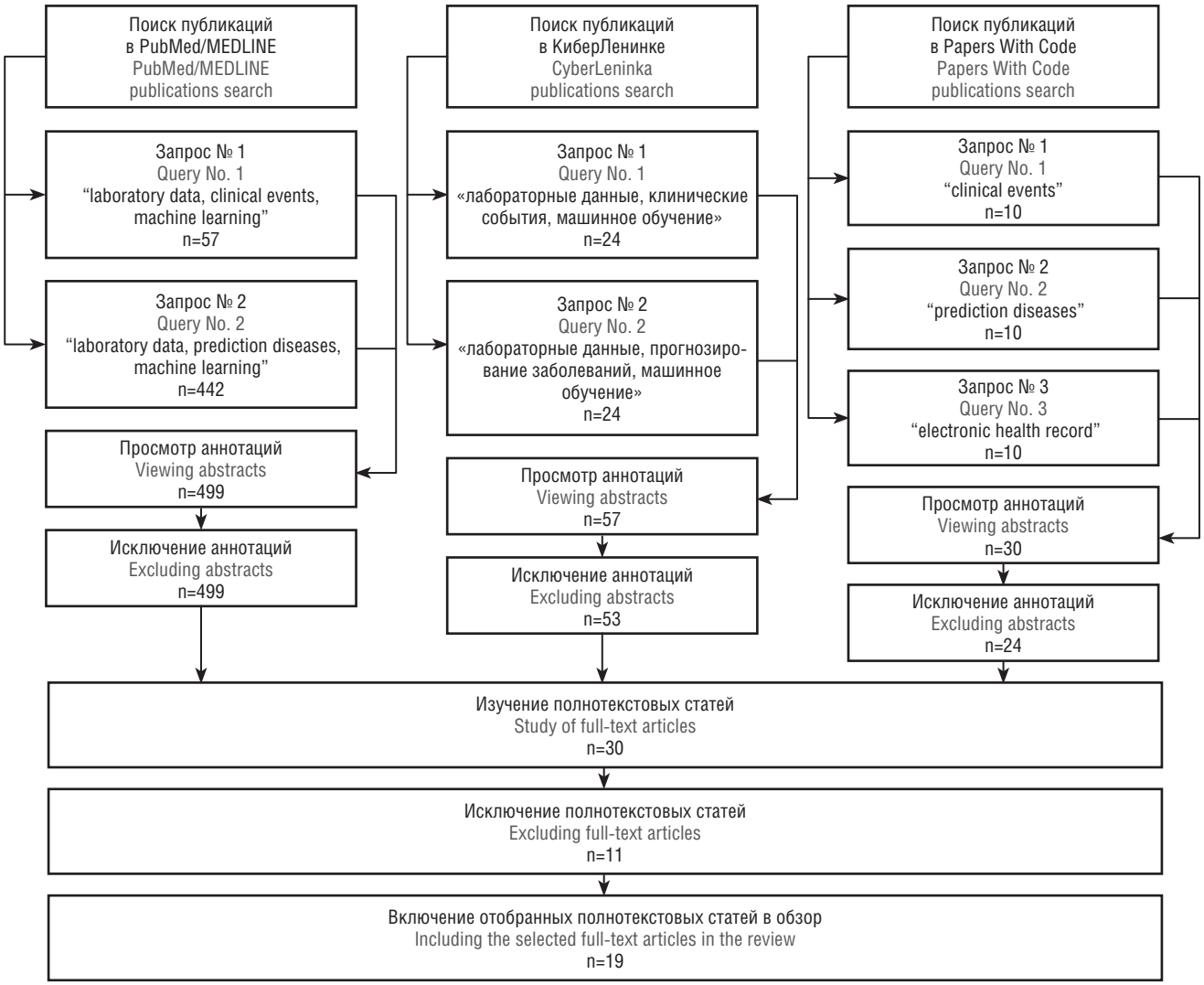


Рисунок 1. Схема дизайна обзора литературы по применению методов машинного обучения для выявления подозрений на заболевания

Figure 1. The design of literature review on the application of machine learning techniques to detect suspected diseases

После изучения полного текста 30 литературных источников, соответствующих критериям отбора, выбрано 19 статей, наиболее релевантных поставленной задаче для представления в данном литературном обзоре.

РЕЗУЛЬТАТЫ / RESULTS

Одной из первых работ, посвященных указанной тематике, является статья M.M. Churpek et al., опубликованная в феврале 2016 г. [8]. Авторы предлагают решение задач прогнозирования остановки сердца, перевода пациента в отделение интенсивной терапии (ОИТ) и смерти больного. Для этого использован набор данных по 269 999 пациентам, среди которых у 424 случился сердечный приступ, 13 188 были переведены в отделение интенсивной терапии, 2 840 умерли. Данные были как демографические, так и лабораторные: возраст, время с момента поступления в клинику, количество предыдущих нахождений в ОИТ, показатели жизненно важных функций (частота дыхания, частота сердцебиения, систолическое артериальное давление, диастолическое артериальное давление, пульсовое давление, температура) и регулярно собираемые результаты лабораторных исследований (анализы на уровни азота мочевины крови (АМК), лейкоцитов, глюкозы, тромбоцитов, гемоглобина, насыщения кислородом, креатинина, соотношения АМК и креатинина, кальция, бикарбоната, хлорида, калия, анионного интервала, натрия, щелочной фосфатазы, сывороточной глутаминовой оксалоуксусной трансаминазы, общего белка, общего билирубина, альбумина). Указанные данные были получены из электронных медицинских карт (англ. Electronic Health Record, EHR) (EPIC, Верона, Висконсин, США) в Чикагском университете и из хранилища электронных данных в клиниках NorthShore. Обучающая и тестовая выборки были сформированы в соотношении 60% и 40% соответственно. На сформированных наборах данных авторы сравнили работу следующих алгоритмов МО: логистическая регрессия (англ. Logistic Regression), деревья решений (англ. Decision Trees), метод опорных векторов (англ. Support Vector Machines, SVM), метод K ближайшего соседа (англ. K-Nearest Neighbors, KNN), нейронные сети (англ. Neural Networks), «случайный лес» (англ. Random Forest). Random Forest показал самую высокую точность прогнозирования: площадь под кривой (англ. area under curve, AUC) составила 0,80.

В феврале 2016 г. E. Choi et al. опубликовали статью по теме прогнозирования сердечной недостаточности [9]. Данные были получены от некоммерческой организации в области здравоохранения Sutter Palo Alto Medical Foundation (Sutter-PAMF, США).

Sutter-PAMF использует EHR более 10 лет. EHR включают демографические сведения, информацию о курении и употреблении алкоголя, клинические и лабораторные данные, коды Международной классификации болезней 9-го пересмотра, информацию о процедурах в кодах текущей процедурной терминологии (англ. Current Procedural Terminology, CPT) и рецепты лекарств. Были применены следующие методы МО: логистическая регрессия, SVM, KNN и искусственная нейронная сеть в виде многослойного персептрона с одним скрытым слоем (англ. Multi-Layer Perceptron, MLP). MLP дал самый высокий результат прогнозирования сердечной недостаточности: AUC=0,814.

Задачу прогнозирования ряда заболеваний (рака простаты, повышенного уровня специфического антигена простаты, рака груди, рака толстой кишки, дегенерации желтого пятна, сердечной недостаточности, болезней почек и печени – всего 25 видов заболеваний) решали авторы публикации N. Razavian et al. в августе 2016 г. [10]. Набор данных был представлен лабораторными сведениями 298 тыс. пациентов. Он содержал 18 признаков: уровни креатинина, АМК, калия, глюкозы, аланинаминотрансферазы (АЛТ), аспартатаминотрансферазы (АСТ), белка, альбумина, общего холестерина, триглицерида, холестерина в липопротеинах низкой плотности (ЛПНП), кальция, натрия, хлорида, диоксида углерода, АМК/креатинина, билирубина/глобулина. Данные были разбиты случайным образом на обучающую, валидационную и тестовую выборки: 100 тыс., 100 тыс. и 98 тыс. соответственно. Были применены метод долгой краткосрочной памяти (англ. Long Short-Term Memory, LSTM), рекуррентные нейронные сети (англ. Recurrent Neural Network, RNN), сверточные нейронные сети (англ. Convolutional Neural Networks, CNN), линейная регрессия и Random Forrest. CNN и Random Forrest показали самую высокую точность прогнозирования: AUC=0,774.

В декабре 2017 г. вышла статья М.В. Сахибгареевой и А.Ю. Заозерского по прогнозированию диагнозов заболеваний на основе искусственного интеллекта [11]. Авторы использовали выборку данных о 7918 случаях заболеваний по четырем нозологиям: D50 Железодефицитная анемия, E11 Инсулинонезависимый сахарный диабет, E74 Другие нарушения обмена углеводов, E78 Нарушения обмена липопротеидов и другие липидемии. Набор данных основан на результатах лабораторных тестов – анализов крови, мочи, цитологических исследований и т.д. Для решения задачи прогнозирования заболеваний авторы применяли следующие алгоритмы МО: деревья решений, искусственные нейронные сети, градиентный бустинг. Метод градиентного бустинга показал наилучший результат: AUC от 89 до 98%.

Алгоритмы МО также применяются при прогнозировании госпитализации и амбулаторного применения кортикостероидов у пациентов с воспалительными заболеваниями кишечника (ВЗК). На эту тему в декабре 2017 г. была опубликована статья A.K. Waljee et al. [12]. В исследовании использованы данные за период с 2002 по 2009 гг. по 20 368 пациентам с диагнозом ВЗК и 351 112 визитам к врачу. Набор данных включал следующие категории: 1) демографические – возраст, пол, раса; 2) лабораторные – уровни лейкоцитов, гемоглобина, гематокрита, средний объем эритроцитов, средняя концентрация гемоглобина в эритроците, уровень тромбоцитов, натрия, калия, глюкозы, АМК, креатинина сыворотки, кальция, бикарбоната, хлорида, альбумина, АСТ, АЛТ, общего белка, щелочной фосфатазы, билирубина; 3) данные лечения – прием лекарств: тиопурин, метотрексат, анти-TNF (англ. Tumor Necrosis Factor – фактор некроза опухоли) или комбинированная терапия; 4) дополнительные переменные: предыдущая госпитализация или назначение стероидов, предыдущие дозы кортикостероидов или госпитализации. Авторы построили модели с применением методов логистической регрессии и Random Forest для прогнозирования госпитализации и использования кортикостероидов при ВЗК в течение 6 мес. AUC для модели логистической регрессии составила 0,68, для модели Random Forrest – 0,85, для модели Random Forest с использованием данных о предыдущей госпитализации или приеме стероидов – 0,87.

В публикации C. Ye et al. за январь 2018 г. предложено решение задачи прогнозирования гипертензии у пациентов в течение 1 года [13], которая актуальна тем, что артериальная гипертензия имеет тяжелые и опасные для жизни последствия, такие как ССЗ, например инсульт. Данные для набора были получены из EHR больниц, федеральных медицинских центров и амбулаторных клиник штата Мэн (США). Всего в выборке содержалось 1,5 млн записей о пациентах с диагнозом гипертензии. Различные категории данных были извлечены из медицинских карт, включая демографические сведения, результаты лабораторных и рентгенографических тестов, основные и сопутствующие диагнозы и процедуры, рецепты лекарств, медицинские записи, а также ряд доступных социально-экономических показателей, полученные с веб-сайтов переписи населения США и Министерства сельского хозяйства США. Для дальнейшего анализа было выбрано 80 признаков, наиболее значимых в прогнозировании гипертензии. В качестве алгоритма МО был использован ансамблевый метод бустинга (англ. eXtreme Gradient Boosting, XGBoost), который показал следующие результаты: AUC=0,917 на обучающей выборке и AUC=0,870 на тестовой выборке.

В марте 2018 г. вышла статья L. Liu et al. по прогнозированию клинических событий, в частности смертельных исходов, на основе лабораторных и диагностических данных и применения лекарственных препаратов [14]. Данные были получены из базы со свободным доступом Medical Information Mart for Intensive Care III (MIMIC-III), содержащей обезличенные сведения, связанные с состоянием здоровья пациентов (более 40 тыс.), которые находились в ОИТ медицинского центра Beth Israel Deaconess (Бостон, Массачусетс, США), и с клиническими событиями (всего 18 192 события) за 2001–2012 гг. Выборка была разбита на обучающую, валидационную и тестовую – 70%, 10% и 20% соответственно. Был выбран нейросетевой алгоритм LSTM, который показал результат AUC=0,7987 на тестовой выборке.

J. Liu et al. в марте 2018 г. опубликовали научную работу по прогнозированию заболеваний (хроническая сердечная недостаточность, почечная недостаточность, инсульт) на основе записей в EHR [15]. Набор данных содержал клинические события у более 1 млн пациентов за период 2014–2017 гг. При помощи методов извлечения информации из текста были определены признаки по лабораторным данным: уровни АМК, креатинина, хлоридов, калия, натрия, двуокиси углерода, гемоглобина, гематокрита, глюкозы, АЛТ, АСТ, эритроцитов, щелочной фосфатазы, билирубина, тромбоцитов, кальция, лейкоцитов, липопротеины высокой плотности (ЛПВП), ЛПНП, альбумина. Выборка была разбита на обучающую, валидационную и тестовую – 70%, 10% и 20% соответственно. Были применены методы CNN и LSTM, последний показал наилучший результат на тестовой выборке: сердечная недостаточность – AUC=0,9, почечная недостаточность – AUC=0,833, инсульт – AUC=0,753.

В 2018 г. по теме прогнозирования компенсации и декомпенсации сахарного диабета у детей и подростков вышла публикация О.С. Кротовой и др. [16]. Для построения моделей прогнозирования была сформирована выборка данных, в которую вошли такие признаки, как рост, вес, температура тела, артериальное давление, частота сердечных сокращений, частота дыхания, длитель-

ность заболевания, показатели биохимического анализа крови. Применялись следующие алгоритмы: логистическая регрессия (точность прогноза 71%), деревья решений (71%), градиентный бустинг (69%).

J. Lin et al. в июле 2018 г. опубликовали работу о применении метода опорных векторов для прогнозирования неврологических ухудшений, в частности ишемического инсульта [17]. В этом исследовании был сформирован набор данных, содержащий информацию о 382 пациентах, госпитализированных с острым ишемическим инсультом. Он включал следующие параметры: возраст, пол, лабораторные данные (уровни натрия, калия, хлора в сыворотке, глюкозы и АМК) и факторы риска (гипертония, диабет, мерцательная аритмия, гиперлипидемия, курение, ишемическая болезнь сердца, гипергомоцистеинемия). Метод опорных векторов показал точность прогноза AUC=0,895.

В январе 2019 г. вышла статья G.P. Diller et al. по прогнозированию результатов лечения при врожденных пороках сердца у взрослых на основе данных одного специализированного центра о 10 019 пациентах [18]. Набор данных содержал ряд характеристик, в т.ч. результаты электрокардиографии и лабораторных исследований (натрийуретический пептид мозга, креатинин). Для классификации основного диагноза и тяжести заболевания применяли метод CNN, показавший точность 91–93% на тестовых выборках. Для прогнозирования результатов лечения использовали тот же алгоритм, точность составила 90,2% на тестовой выборке.

Y.W. Lin et al. в июле 2019 г. подготовили публикацию, в которой предложили решение задачи прогнозирования внеплановой повторной госпитализации пациента в течение 30 сут после выписки [19]. Набор данных был сформирован из записей 20 368 больных и 351 112 визитов к врачу, в качестве характеристик было использовано 59 признаков в т.ч. лабораторные данные (уровень глюкозы), диастолическое и систолическое артериальное давление. В числе прочих применяли следующие алгоритмы: наивный байесовский классификатор (англ. Complement Naive Bayes), Random Forest и SVM. Наилучший результат прогнозирования показал метод RNN с LSTM: AUC=0,791.

H.L. Wang et al. в августе 2019 г. провели исследование по прогнозированию клинических исходов (ишемического инсульта, аневризматического субарахноидального кровоизлияния и летального исхода) у пациентов с внутримозговым кровоизлиянием [20]. Набор данных содержал информацию о 333 больных; 22 характеристики включали в себя, в частности, лабораторные данные: уровни глюкозы в сыворотке крови, АСТ, АЛТ, АМК, креатинина, соотношения АМК/креатинин, гликозилированного гемоглобина, результаты общего анализа крови, уровни триглицеридов, общего холестерина, С-реактивного белка (СРБ), мочевой кислоты, а также протромбиновое время, активированное частичное тромбиновое время и гиперчувствительный тест на СРБ. В исследовании применяли 10-кратную матрицу кросс-валидации. Исходные данные были разделены на 10 подвыборок примерно одинакового размера. Одна из них использовалась в качестве набора проверочных данных для тестирования моделей, а остальные девять – как обучающие. Далее процесс перекрестной проверки был повторен 10 раз с одной из 10 подвыборок, используемых последовательно для каждой проверки. Затем 10 результатов каждой повторной проверки были усреднены для получения окончательной оценки. Среди 39 рассмотренных моделей Random Forrest обеспечил лучший прогноз результата: точность прогнозирования исхода 1-го месяца – AUC=0,899. Общая точность прогнозирования исхода через 6 мес – AUC=0,917.

J. Gordon и B. Lerner в октябре 2019 г. опубликовали статью о применении алгоритмов МО в задаче прогнозирования бокового амиотрофического склероза [21]. Авторы использовали записи 3772 пациентов из базы данных с открытым доступом ALS Clinical Trials (PRO-ACT). Набор характеризуется следующими видами данных: демографические, клинические, лабораторные (базофилы, эозинофилы, лимфоциты, моноциты, альбумин, щелочная фосфатаза, АЛТ, АСТ, бикарбонат, билирубин, АМК, кальций, хлорид, креатинкиназа, креатинин, глюкоза, гематокрит, гемоглобин, фосфор, тромбоциты, белок клетки, натрий, лейкоциты). В исследовании применяли 10-кратную матрицу кросс-валидации и следующие алгоритмы МО: деревья решений (алгоритм C5.0), а также Random Forrest, XGBoost, основанные на множественном построении деревьев решений. Random Forrest и XGBoost показали наилучшие результаты: 85%.

В октябре 2019 г. вышла статья H. Lai et al. о применении методов МО в прогнозировании сахарного диабета [22]. Данные, использованные в этом исследовании, были получены из источника Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [23]. Выборка содержала информацию о 13 309 пациентах, описанных следующими признаками: пол, возраст, индекс массы тела, уровни триглицеридов, сахара в крови натощак, систолическое артериальное давление, ЛПВП и ЛПНП. Выборка была разбита в пропорции 80/20 на обучающую и тестовую соответственно. Для построения модели прогнозирования применяли методы логистической регрессии и градиентного бустинга, последний из которых показал самый точный результат: AUC=0,85.

Исследование в области прогнозирования диабета и ССЗ было проведено в ноябре 2019 г. A. Dinh et al. [24]. Набор данных по 5 тыс. пациентов сформирован из источника National Health and Nutrition Examination Survey (NHANES) [25]. Он уникален тем, что сочетает в себе результаты опросников, медицинских осмотров и лабораторных исследований, проводимых в медицинских учреждениях. Данные обследования состояли из социально-экономических, демографических аспектов и вопросов, связанных с диетой и состоянием здоровья пациентов. Лабораторные данные включали медицинские, стоматологические, физические и физиологические измерения. При помощи методов отбора информативных признаков было выбрано 24 наиболее значимых для диагностики сахарного диабета и ССЗ. В исследовании применяли 10-кратную матрицу кросс-валидации и следующие алгоритмы МО: логистическая регрессия, SVM, методы ансамблей моделей – Random Forrest, XGBoost. Ансамблевая модель для ССЗ без учета лабораторных данных достигла 83,1% точности прогнозирования и 83,9% точности с лабораторными данными. В прогнозировании диабета модель XGBoost показала точность 86,2% без лабораторных данных и 95,7% с лабораторными данными. Для пациентов с преддиабетом модель ансамбля имела наивысший показатель 73,7% без лабораторных данных, а для лабораторных данных лучший результат продемонстрировал XGBoost: 84,4%.

W. Zhu и N. Razavian в декабре 2019 г. опубликовали результаты исследования по прогнозированию болезни Альцгеймера [26]. Авторы использовали записи EHR 1,64 млн пациентов с указанием лабораторных данных, проведенных процедур и демографических сведений. Выборка была разбита в соотношении 70/20/10 на обучающую, валидационную и тестовую соответственно. Были применены алгоритмы логистической регрессии, Random Forrest, MLP. Наилучший результат показала предлагаемая авторами сеть с механизмом внимания графа (англ. Graph Attention Network): AUC=0,8.

Статья I. Landi et al. (март 2020 г.) посвящена прогнозированию различных болезней на основе данных 1,6 млн пациентов [27].

Авторы обратили внимание на такие заболевания, как сахарный диабет 2-го типа, болезни Паркинсона, Альцгеймера, множественная миелома, рак простаты и груди, болезнь Крона, синдром дефицита внимания с гиперактивностью. В качестве признаков использовались данные о проводимых процедурах, применяемых медицинских препаратах, а также результаты лабораторных исследований: уровни лейкоцитов, глюкозы, гематокрита, эритроцитов, средняя концентрация гемоглобина в эритроците. Следует отметить, что в этой научной работе применялись алгоритмы МО без учителя, т.е. кластеризация больных по схожим признакам, когда составляется общая визуальная картина по пациентам и их возможным болезням. Авторы предлагают так называемую модель ConvAE (англ. Convolutional Auto-Encoder) – модель обучения, основанную на встраивании слов, сверточных нейронных сетях и автоэнкодерах. Результаты кластеризации представлены визуально в приложениях к тексту статьи.

В работе R. Weegar и K. Sundström (август 2020 г.) были использованы алгоритмы МО для прогнозирования рака шейки матки [28]. Исследованы записи EHR о 1723 пациентах. Набор данных характеризуется информацией о проведенных процедурах, лабораторных данных, а также признаках, извлеченных из EHR при помощи алгоритмов обработки текста (англ. Named Entities Extraction). Авторы применяли 10-кратную матрицу кросс-валидации и следующие алгоритмы МО: Random Forest, Complement Naive Bayes, Bernoulli Naive Bayes и SVM. Наиболее точный результат показал Random Forest: AUC=0,97.

Краткий обзор рассмотренных публикаций представлен в **таблице 1**.

ОБСУЖДЕНИЕ / DISCUSSION

Результаты последовательного интеллектуального анализа литературных источников демонстрируют высокую прогностическую ценность алгоритмов МО, внедренных в МИС. В большинстве исследований точность прогноза (AUC) составила 0,8 и более, что является высоким показателем работы предиктивных математических моделей. Так, например, в исследованиях сравнения методов МО для прогнозирования хронических неинфекционных заболеваний и неврологических расстройств AUC составила 0,9 [15, 18], клинических исходов у пациентов с первичным внутримозговым кровоизлиянием – 0,92 [21], сахарного диабета и ССЗ – 0,96 [25], рака шейки матки – 0,97 [29].

Таблица 1 (начало). Краткий обзор публикаций по теме применения методов машинного обучения с целью выявления подозрений на заболевания

Table 1 (beginning). Summary review of publications on applying machine learning methods to detect suspected diseases

Авторы Authors	**Год** Year	**Цель исследования** Purpose of study	**Набор данных** Data set	**Число признаков, n** Number of features, n	**Обучающая/ тестовая выборка, %** Training/test sample, %	**Наилучший метод МО** Best ML method	**Результат** Result
M.M. Churpek et al. [8]	2016	Сравнение методов МО и линейной регрессии для прогнозирования негативных исходов в клинических отделениях Comparison of ML and linear regression methods for predicting adverse outcomes in clinical wards	269 999 пациентов 269,999 patients	29	60/40	Random Forrest	AUC=0,8
E. Choi et al. [9]	2016	Сравнение методов МО для прогнозирования сердечной недостаточности Comparison of ML methods for predicting heart failure	265 336 пациентов, 555 609 уникальных клинических событий 265,336 patients, 555,609 unique clinical events	–	–	Neural Networks	AUC=0,814
N. Razavian et al. [10]	2016	Сравнение методов МО (нейронных сетей) и линейной регрессии для прогнозирования ряда заболеваний Comparison of ML (neural network) and linear regression methods for predicting a number of diseases	298 000 пациентов 298,000 patients	18	67/33	Random Forrest	AUC=0,774
М.В. Сахибгареева, А.Ю. Заозерский [11] / M.V. Sakhibgareeva, A.Yu. Zaozersky [11]	2017	Выбор и обоснование применения метода МО для прогнозирования нозологических диагнозов Selection and justification of ML methods for prediction of nosological diagnoses	7918 случаев заболеваний 7,918 cases	–	75/25	Gradient Boosting	AUC=0,89

Таблица 1 (продолжение). Краткий обзор публикаций по теме применения методов машинного обучения с целью выявления подозрений на заболевания

Table 1 (continuation). Summary review of publications on applying machine learning methods to detect suspected diseases

Авторы Authors	Год Year	Цель исследования Purpose of study	Набор данных Data set	Число признаков, n Number of features, n	Обучающая/тестовая выборка, % Training/test sample, %	Наилучший метод МО Best ML method	Результат Result
A.K. Waljee et al. [12]	2017	Сравнение методов МО для прогнозирования госпитализации и амбулаторного применения кортикостероидов у пациентов с воспалительным заболеванием кишечника Comparison of ML methods for predicting hospitalization and outpatient corticosteroid use in patients with inflammatory bowel disease	20 368 пациентов и 351 112 визитов к врачу 20,368 patients and 351,112 physician visits	32	70/30	Random Forrest	AUC=0,87
C. Ye et al. [13]	2018	Выбор и обоснование применения метода МО для прогнозирования гипертензии Selection and rationale for the use of an ML method for predicting hypertension	1 500 000 пациентов 1,500,000 patients	80	55/45	XGBoost	AUC=0,87
L. Liu et al. [14]	2018	Сравнение методов МО для прогнозирования клинических событий Comparison of ML methods for predicting clinical events	40 000 пациентов и 18 192 вида событий 40,000 patients and 18,192 types of events	9	80/20	LSTM	AUC=0,8
J. Liu et al. [15]	2018	Сравнение методов МО для прогнозирования хронических заболеваний Comparison of ML methods for predicting chronic diseases	1 000 000 пациентов 1,000,000 patients	50	80/20	LSTM	AUC=0,9
О.С. Кротова и др. [16] / O.S. Krotova et al. [16]	2018	Сравнение методов МО для прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков Comparison of ML methods for predicting stages of compensation and decompensation of diabetes mellitus in children and adolescents	–	–	–	Decision Trees, Logistic Regression	Prec.=0,71
J. Lin et al. [17]	2018	Выбор и обоснование применения метода МО для прогнозирования неврологического ухудшения Choice and rationale for the use of the ML method for predicting neurological deterioration	382 пациента 382 patients	18	80/20	Support Vectors Machine	AUC=0,9
G.P. Diller et al. [18]	2019	Выбор и обоснование применения метода МО для прогнозирования результатов лечения при врожденных пороках сердца у взрослых Choice and rationale for the use of the ML method to predict treatment outcomes in adult congenital heart disease	10 019 пациентов 10,019 patients	10–15	80/20	CNN	Prec.=0,86

Таблица 1 (окончание). Краткий обзор публикаций по теме применения методов машинного обучения с целью выявления подозрений на заболевания

Table 1 (end). Summary review of publications on applying machine learning methods to detect suspected diseases

Авторы Authors	Год Year	Цель исследования Purpose of study	Набор данных Data set	Число признаков, n Number of features, n	Обучающая/тестовая выборка, % Training/test sample, %	Наилучший метод МО Best ML method	Результат Result
Y.W. Lin et al. [19]	2019	Сравнение методов МО для прогнозирования внеплановой повторной госпитализации Comparison of ML methods for predicting unscheduled rehospitalization	40 000 пациентов и 60 000 записей о поступлении в ОИТ 40,000 patients and 60,000 OCU admission records	59	90/10	RNN, LSTM	AUC=0,79
H.L. Wang et al. [20]	2019	Сравнение методов МО для прогнозирования клинических исходов у пациентов с первичным внутримозговым кровоизлиянием Comparison of ML methods to predict clinical outcomes in patients with primary intracerebral hemorrhage	333 пациента 333 patients	22	90/10	Random Forrest	AUC=0,92
J. Gordon, B. Lerner [21]	2019	Сравнение методов МО для прогнозирования бокового амиотрофического склероза Comparison of ML methods for predicting amyotrophic lateral sclerosis	3772 пациента 3,772 patients	22	90/10	XGBoost, Random Forrest	Prec.=0,85
H. Lai et al. [22]	2019	Сравнение методов МО для прогнозирования сахарного диабета Comparison of ML methods for predicting diabetes mellitus	13 309 пациентов 13,309 patients	8	80/20	Gradient Boosting	AUC=0,85
A. Dinh et al. [24]	2019	Сравнение методов МО для прогнозирования диабета и сердечно-сосудистых заболеваний Comparison of ML methods for predicting diabetes and cardiovascular disease	5 000 пациентов 5,000 patients	24	80/20	XGBoost	AUC=0,96
W. Zhu, N. Razavian [26]	2019	Сравнение методов МО для прогнозирования болезни Альцгеймера Comparison of ML methods for predicting Alzheimer's disease	1 640 000 пациентов 1,640,000 patients	–	90/10	Graph Attention Network	AUC=0,8
I. Landi et al. [27]	2020	Сравнение методов МО для прогнозирования различных болезней Comparison of ML methods for predicting different diseases	1 600 000 пациентов 1,600,000 patients	–	50/50	ConvAE	–
R. Weegar, K. Sundström [28]	2020	Сравнение методов МО для прогнозирования рака шейки матки Comparison of ML methods for predicting cervical cancer	1723 пациента 1,723 patients	25	90/10	Random Forrest	AUC=0,97

Примечание. *МО – машинное обучение; ОИТ – отделение интенсивной терапии; Random Forrest – «случайный лес»; Neural Networks – нейронные сети; Gradient Boosting – градиентный бустинг; XGBoost (англ. eXtreme Gradient Boosting) – ансамблевый градиентный бустинг; LSTM (англ. Long Short-Term Memory) – долгая краткосрочная память; Decision Trees – деревья решений; Logistic Regression – логистическая регрессия; Support Vectors Machine – метод опорных векторов; CNN (англ. Convolutional Neural Networks) – сверточные нейронные сети; Graph Attention Network – сеть с механизмом внимания графа; ConvAE (англ. Convolutional Auto-Encoder) – сверточный автокодер; AUC (англ. area under curve) – площадь под кривой; Prec. (англ. precision) – точность.*

Note. *ML – machine learning; ICU – intensive care unit; XGBoost – eXtreme Gradient Boosting; LSTM – Long Short-Term Memory; CNN – Convolutional Neural Networks; ConvAE – Convolutional Auto-Encoder; AUC – area under curve; Prec. – precision.*

Среди используемых алгоритмов МО чаще всего предпочтение отдавалось ансамблевым методам, таким как Random Forest и XGBoost [8, 10, 13, 19, 25]. Это обусловлено высокой точностью прогнозирования, что является неоспоримым преимуществом. Однако следует обратить внимание на главные недостатки указанных методов – более длительное время обучения и относительную сложность разработки моделей. Также важно отметить, что в ряде приведенных публикаций используется разное число записей в обучающих и тестовых выборках.

Принимая во внимание результаты проведенного исследования, можно констатировать, что обученные на более объемных выборках модели показывают более точный результат, т.к. чаще всего десятков и сотен записей бывает недостаточно для обучения сложных моделей. Вероятно, для получения наилучших результатов работы прогнозных моделей потребуется определить оптимальное число записей и искомых признаков. Кроме того, задача оптимизации актуальна и для выбора параметров обучающих алгоритмов. Например, для искусственной нейронной сети необходимо найти оптимальное число скрытых слоев и нейронов в заданных слоях. В качестве решения задачи оптимизации предлагается рассмотреть генетические алгоритмы.

ЗАКЛЮЧЕНИЕ / CONCLUSION

Использование точных предиктивных моделей, основанных на алгоритмах машинного обучения, открывает новые возможности в прогнозировании неинфекционных заболеваний, их осложнений, вероятности госпитализации пациентов с целью своевременного назначения профилактических и лечебных мероприятий и разработки системы поддержки принятия врачебных решений.

Результаты литературного обзора подчеркивают высокую точность прогнозных математических моделей в таких областях медицины, как кардиология, неврология, генетика, гастроэнтерология, иммунология и гинекология. Наиболее распространенными и эффективными алгоритмами МО оказались Random Forest, нейронные сети и ансамблевые модели с градиентным усилением (XGBoost). Несмотря на более высокую сложность разработки указанных моделей, их применение оправдано повышением точности прогноза.

Внедрение алгоритмов МО в МИС представляется перспективным инструментом эффективного прогнозирования неблагоприятных медицинских событий для широкого применения в реальной клинической практике, что соответствует общемировой тенденции по развитию персонифицированной, основанной на расчете индивидуального риска медицины.

По итогам анализа научных публикаций можно сделать вывод о тенденции роста активности исследований в области прогнозирования неинфекционных заболеваний с использованием технологий искусственного интеллекта. С каждым годом число научных работ по данной проблеме неуклонно растет, что подтверждает актуальность дальнейшей разработки темы применения технологий МО в медицинских информационных системах.

ЛИТЕРАТУРА:

1. ВОЗ. Информационный бюллетень. Прогресс в борьбе с неинфекционными заболеваниями. Июнь 2017. *Социальные аспекты здоровья населения.* 2017; 4: 1–10.
2. Гаврилов Д.В., Серова Л.М., Корсаков И.Н. и др. Предсказание сердечно-сосудистых событий при помощи комплексной оценки факторов риска с использованием методов машинного обучения. *Врач.* 2020; 31 (5): 41–6. https://doi.org/10.29296/25877305-2020-05-08.
3. Гусев А.В., Гаврилов Д.В., Корсаков И.Н. и др. Перспективы использования методов машинного обучения для предсказания сердечно-сосудистых заболеваний. *Врач и информационные технологии.* 2019; 3: 41–7.
4. Федеральный справочник лабораторных исследований. Справочник лабораторных тестов. URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1080/version/3.28 (дата обращения 23.09.2021).
5. National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov/ (дата обращения 23.09.2021).
6. Научная электронная библиотека «КиберЛенинка». URL: https://cyberleninka.ru/ (дата обращения 23.09.2021).
7. Papers With Code, free resource with all data licensed under CC-BY-SA. URL: https://paperswithcode.com/ (дата обращения 23.09.2021).
8. Churpek M.M., Yuen T.C., Winslow C., et al. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med.* 2016; 44 (2): 368–74. https://doi.org/10.1097/CCM.0000000000001571.
9. Choi E., Schuetz A., Stewart W.F., Sun J. Medical concept representation learning from electronic health records and its application on heart failure prediction. 2016; arXiv: 1602.03686.
10. Razavian N., Marcus J., Sontag D. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. 2016; arXiv: 1608.00647.
11. Сахибгареева М.В., Заозерский А.Ю. Разработка системы прогнозирования диагнозов заболеваний на основе искусственного интеллекта. *Вестник Российского государственного медицинского университета.* 2017; 6: 42–6.
12. Waljee A.K., Lipson R., Wiitala W.L., et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis.* 2017; 24 (1): 45–53. https://doi.org/10.1093/ibd/izx007.
13. Ye C., Fu T., Hao S., et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res.* 2018; 20 (1): e22. https://doi.org/10.2196/jmir.9268.
14. Liu L., Shen J., Zhang M., et al. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. 2018; arXiv: 1803.04837.
15. Liu J., Zhang Z., Razavian N. Deep EHR: chronic disease prediction using medical notes. 2018; arXiv: 1808.04928.
16. Кротова О.С., Пиянзин А.И., Хворова Л.А., Жариков А.В. Некоторые математические подходы в построении моделей прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков. *Известия Алтайского государственного университета.* 2018; 4: 83–7. https://doi.org/10.14258/izvasu(2018)4-15.
17. Lin J., Jiang A., Ling M., et al. Prediction of neurologic deterioration based on support vector machine algorithms and serum osmolarity equations. *Brain Behav.* 2018; 8 (7): e01023. https://doi.org/10.1002/brb3.1023.
18. Diller G.P., Kempny A., Babu-Narayan S.V., et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J.* 2019; 40 (13): 1069–77. https://doi.org/10.1093/eurheartj/ehy915.
19. Lin Y.W., Zhou Y., Faghri F., et al. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural

networks with long short-term memory. *PloS One*. 2019; 14 (7): e0218942. https://doi.org/10.1371/journal.pone.0218942.
20. Wang H.L., Hsu W.Y., Lee M.H., et al. Automatic machine-learning-based outcome prediction in patients with primary intracerebral hemorrhage. *Front Neurol*. 2019; 10: 910. https://doi.org/10.3389/fneur.2019.00910.
21. Gordon J., Lerner B. Insights into amyotrophic lateral sclerosis from a machine learning perspective. *J Clin Med*. 2019; 8 (10): 1578. https://doi.org/10.3390/jcm8101578.
22. Lai H., Huang H., Keshavjee K., et al. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord*. 2019; 19 (1): 101. https://doi.org/10.1186/s12902-019-0436-6.
23. Canadian Primary Care Sentinel Surveillance Network (CPCSSN). URL: http://cpcssn.ca/ (дата обращения 23.09.2021).
24. Dinh A., Miertschin S., Young A., Mohanty S.D. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019; 19 (1): 211. https://doi.org/10.1186/s12911-019-0918-5.
25. National Center for Health Statistics. URL: https://www.cdc.gov/nchs/nhanes/ (дата обращения 23.09.2021).
26. Zhu W., Razavian N. Graph neural network on electronic health records for predicting Alzheimer's disease. 2019; arXiv: 1912.03761.
27. Landi I., Glicksberg B., Lee H., et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med*. 2020; 3: 96. https://doi.org/10.1038/s41746-020-0301-z.
28. Weegar R., Sundström K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PloS One*. 2020; 15 (8): e0237911. https://doi.org/10.1371/journal.pone.0237911.

REFERENCES:

1. WHO. Information Bulletin. Progress in the fight against non-communicable diseases. June 2017. *Social Aspects of Population Health*. 2017; 4: 1–10 (in Russ.).
2. Gavrilov D., Serova L., Korsakov I., et al. Cardiovascular diseases prediction by integrated risk factors assessment by means of machine learning. *Vrach*. 2020; 31 (5): 41–6 (in Russ.). https://doi.org/10.29296/25877305-2020-05-08.
3. Gusev A.V., Gavrilov D.V., Korsakov I.N., et al. Prospects for the use of machine learning methods for predicting cardiovascular disease. *Medical Doctor and IT*. 2019; 3: 41–7 (in Russ.).
4. Federal Guide of Laboratory Research. Guide of laboratory tests. Available at: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1080/version/3.28 (accessed 23.09.2021).
5. National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/ (accessed 23.09.2021).
6. Scientific electronic library "CyberLeninka". Available at: https://cyberleninka.ru/ (accessed 23.09.2021).
7. Papers With Code, free resource with all data licensed under CC-BY-SA. Available at: https://paperswithcode.com/ (accessed 23.09.2021).
8. Churpek M.M., Yuen T.C., Winslow C., et al. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016; 44 (2): 368–74. https://doi.org/10.1097/CCM.0000000000001571.
9. Choi E., Schuetz A., Stewart W.F., Sun J. Medical concept representation learning from electronic health records and its application on heart failure prediction. 2016; arXiv: 1602.03686.
10. Razavian N., Marcus J., Sontag D. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. 2016; arXiv: 1608.00647.
11. Sakhibgareeva M.V., Zaozersky A.Yu. Developing an artificial intelligence-based system for medical prediction. *Bulletin of Russian State Medical University*. 2017; 6: 42–6 (in Russ.).
12. Waljee A.K., Lipson R., Wiitala W.L., et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis*. 2017; 24 (1): 45–53. https://doi.org/10.1093/ibd/izx007.
13. Ye C., Fu T., Hao S., et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. *J Med Internet Res*. 2018; 20 (1): e22. https://doi.org/10.2196/jmir.9268.
14. Liu L., Shen J., Zhang M., et al. Learning the joint representation of heterogeneous temporal events for clinical endpoint prediction. 2018; arXiv: 1803.04837.
15. Liu J., Zhang Z., Razavian N. Deep EHR: chronic disease prediction using medical notes. 2018; arXiv: 1808.04928.
16. Krotova O.S., Piyanzin A.I., Khvorova L.A., Zharikov A.V. Some mathematical approaches to develop models for prediction of compensation and decompensation stages of diabetes mellitus among children and adolescents. *Izvestiya of Altai State University*. 2018; 4: 83–7 (in Russ.). https://doi.org/10.14258/izvasu(2018)4-15.
17. Lin J., Jiang A., Ling M., et al. Prediction of neurologic deterioration based on support vector machine algorithms and serum osmolarity equations. *Brain Behav*. 2018; 8 (7), e01023. https://doi.org/10.1002/brb3.1023.
18. Diller G.P., Kempny A., Babu-Narayan S.V., et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J*. 2019; 40 (13): 1069–77. https://doi.org/10.1093/eurheartj/ehy915.
19. Lin Y.W., Zhou Y., Faghri F., et al. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PloS One*. 2019; 14 (7): e0218942. https://doi.org/10.1371/journal.pone.0218942.
20. Wang H.L., Hsu W.Y., Lee M.H., et al. Automatic machine-learning-based outcome prediction in patients with primary intracerebral hemorrhage. *Front Neurol*. 2019; 10: 910. https://doi.org/10.3389/fneur.2019.00910.
21. Gordon J., Lerner B. Insights into amyotrophic lateral sclerosis from a machine learning perspective. *J Clin Med*. 2019; 8 (10): 1578. https://doi.org/10.3390/jcm8101578.
22. Lai H., Huang H., Keshavjee K., et al. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocr Disord*. 2019; 19 (1): 101. https://doi.org/10.1186/s12902-019-0436-6.
23. Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Available at: http://cpcssn.ca/ (accessed 23.09.2021).
24. Dinh A., Miertschin S., Young A., Mohanty S.D. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. 2019; 19 (1): 211. https://doi.org/10.1186/s12911-019-0918-5.
25. National Center for Health Statistics. Available at: https://www.cdc.gov/nchs/nhanes/ (accessed 23.09.2021).
26. Zhu W., Razavian N. Graph neural network on electronic health records for predicting Alzheimer's disease. 2019; arXiv: 1912.03761.
27. Landi I., Glicksberg B., Lee H., et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med*. 2020; 3: 96. https://doi.org/10.1038/s41746-020-0301-z.
28. Weegar R., Sundström K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PloS One*. 2020; 15 (8): e0237911. https://doi.org/10.1371/journal.pone.0237911.

Сведения об авторах

Гусев Александр Владимирович – к.т.н., директор по развитию бизнеса ООО «К-Скай» (Петрозаводск, Россия). ORCID: https://orcid.org/0000-0002-7380-8460; Scopus Author ID: 57222273391; РИНЦ SPIN-код: 9160-7024.

Новицкий Роман Эдвардович – генеральный директор ООО «К-Скай» (Петрозаводск, Россия). ORCID ID: https://orcid.org/0000-0002-2350-977X; Scopus Author ID: 57222272806; РИНЦ SPIN-код: 8309-1740.

Ившин Александр Анатольевич – к.м.н., заведующий кафедрой акушерства и гинекологии, дерматовенерологии ФГБОУ ВО «Петрозаводский государственный университет» (Петрозаводск, Россия). ORCID ID: https://orcid.org/0000-0001-7834-096X; Scopus Author ID: 57222275843; РИНЦ SPIN-код: 8196-6605. E-mail: scipeople@mail.ru.

Алексеев Александр Алексеевич – специалист ООО «К-Скай» (Петрозаводск, Россия).

About the authors

Aleksandr V. Gusev – PhD (Engineering), Business Development Director, K-SkAI LLC (Petrozavodsk, Russia). ORCID: https://orcid.org/0000-0002-7380-8460; Scopus Author ID: 57222273391; RSCI SPIN-код: 9160-7024.

Roman E. Novitskiy – Director General, K-SkAI LLC (Petrozavodsk, Russia). ORCID ID: https://orcid.org/0000-0002-2350-977X; Scopus Author ID: 57222272806; RSCI SPIN-code: 8309-1740.

Aleksandr A. Ivshin – MD, PhD, Chief of Chair of Obstetrics and Gynecology, Dermatovenerology, Petrozavodsk State University (Petrozavodsk, Russia). ORCID ID: https://orcid.org/0000-0001-7834-096X; Scopus Author ID: 57222275843; RSCI SPIN-code: 8196-6605. E-mail: scipeople@mail.ru.

Aleksandr A. Alekseev – Specialist, K-SkAI LLC (Petrozavodsk, Russia).