DOI: https://doi.org/10.17816/DD678373 EDN: QSANCA

Применение больших языковых моделей в лучевой диагностике: обзор предметного поля

Ю.А. Васильев[1], Р.В. Решетников[1], О.Г. Нанова[1], А.В. Владзимирский[1], К.М. Арзамасов[1], О.В. Омелянская[1], М.Р. Коденко[1], Р.А. Ерижоков[1], А.П. Памова[1], С.Р. Сераджи[1], И.А. Блохин[1], А.П. Гончар[1,2], П.Б. Гележе[1], Д.А. Ахмедзянова[1], Ю.Ф. Шумская[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий, Москва, Россия;
[2] Городская клиническая больница им. С.С. Юдина, Москва, Россия

АННОТАЦИЯ

Обоснование. Современные большие языковые модели обладают потенциалом использования в лучевой диагностике для решения широкого спектра рутинных задач.

Цель исследования. Провести обзор предметного поля применения больших языковых моделей в лучевой диагностике с анализом возможных сценариев их использования и оценкой качества методологии соответствующих исследований.

Методы. Провели два варианта поиска — первичный (PubMed и eLibrary), ориентированный на выявление полнотекстовых публикаций с максимально проработанной методологией, и дополнительный (PubMed), направленный на широкий охват сценариев применения больших языковых моделей в лучевой диагностике за период 2023–2025 гг. Извлекали библиометрические данные, формулировку исследовательской задачи, сценарий применения больших языковых моделей, нозологический профиль, ключевые методологические параметры, а также количественные и качественные показатели диагностической эффективности как моделей, так и участвующих специалистов, включая их число и опыт. Качество исследований оценивали с использованием модифицированного опросника QUADAS-CAD.

Результаты. При первичном поиске для анализа отобрано 9 публикаций, при дополнительном — 216. Найдено 9 основных сценариев применения больших языковых моделей в лучевой диагностике. Наиболее распространёнными из них было переформулирование рентгенологических заключений с целью повышения их доступности восприятия пациентами. Преимущественно использовали модели GPT-4 и BERT, а также GPT-3.5, Llama 2, Med42, GPT-4V и Gemini Pro. Большая языковая модель GPT-4 продемонстрировала высокую точность при диагностике опухолей головного мозга (73,0%), миокардитов (83,0%), а также в случае принятия решений о проведении инвазивной процедуры при остром коронарном синдроме (86,0%). В свою очередь, она продемонстрировала низкую диагностическую точность в отношении патологий нервной системы различной этиологии (50,0%) и заболеваний опорно-двигательной системы (43,0%). Модель BERT показала высокую диагностическую точность в задачах детекции лёгочных узелков (99,0%) и признаков внутричерепного кровоизлияния (чувствительность и специфичность — 97,0 и 90,0% соответственно), а также при классификации заключений (точность 84,3%).

Большинство работ (88,9%) содержат вероятность систематической ошибки. Основные причины этого: маленький объём и несбалансированность выборок, пересечение обучающих и тестовых наборов данных, недостаточно аккуратная подготовка и описание референсных стандартов.

Заключение. Показатели диагностической точности больших языковых моделей сильно варьируют между разными исследованиями. Для их внедрения в клиническую практику необходимо проведение стандартизированных и методологически качественных исследований, включающих увеличение объёма и сбалансированности выборок, оптимизацию структуры и объёма наборов данных, формирование неперекрывающихся обучающих и тестовых выборок, тщательную подготовку и описание референсных стандартов, а также накопление эмпирических данных по отдельным задачам лучевой диагностики.

Ключевые слова: искусственный интеллект; большие языковые модели; лучевая диагностика; рентгенологический протокол; систематический обзор.

Как цитировать:
Васильев Ю.А., Решетников Р.В., Нанова О.Г., Владзимирский А.В., Арзамасов К.М., Омелянская О.В., Коденко М.Р., Ерижоков Р.А., Памова А.П., Сераджи С.Р., Блохин И.А., Гончар А.П., Гележе П.Б., Ахмедзянова Д.А., Ю.Ф. Шумская Ю.Ф. Применение больших языковых моделей в лучевой диагностике: обзор предметного поля // Digital Diagnostics. 2025. Т. 6, № 2. С. 268–285. DOI: 10.17816/DD678373 EDN: QSANCA

Рукопись получена: 06.05.2025 **Рукопись одобрена:** 12.06.2025 **Опубликована online:** 17.06.2025

ЭКО • ВЕКТОР

Статья доступна по лицензии CC BY-NC-ND 4.0 International
© Эко-Вектор, 2025

DOI: https://doi.org/10.17816/DD678373 EDN: QSANCA

Application of Large Language Models in Radiological Diagnostics: A Scoping Review

Yuriy A. Vasilev[1], Roman V. Reshetnikov[1], Olga G. Nanova[1], Anton V. Vladzymyrskyy[1], Kirill M. Arzamasov[1], Olga V. Omelyanskaya[1], Maria R. Kodenko[1], Rustam A. Erizhokov[1], Anastasia P. Pamova[1], Seal R. Seradzhi[1], Ivan A. Blokhin[1], Anna P. Gonchar[1,2], Pavel B. Gelezhe[1], Dina A. Akhmedzyanova[1], Yuliya F. Shumskaya[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] Moscow City Hospital named after S.S. Yudin, Moscow, Russia

ABSTRACT

BACKGROUND: Modern large language models show potential for application in radiological diagnostics across a wide range of routine tasks.

AIM: The work aimed to conduct a scoping review of the application of large language models in radiological diagnostics by analyzing possible use-case scenarios and assessing the methodological quality of relevant studies.

METHODS: Two search strategies were employed: a primary search (PubMed and eLibrary) targeting full-text publications with well-developed methodology, and a supplementary search (PubMed) aimed at broader coverage of large language model use cases in radiological diagnostics during 2023–2025. Extracted data included bibliometric characteristics, study objectives, use-case scenarios of large language models, nosological profiles, key methodological parameters, and both quantitative and qualitative indicators of diagnostic performance—for both the models and the specialists involved, including their number and experience. The quality was assessed using the modified QUADAS-CAD questionnaire.

RESULTS: The primary search yielded 9 studies for analysis; the supplementary search yielded 216. A total of 9 major use-case scenarios for large language models in radiology were identified. The most common among them was the rephrasing of radiology reports in order to improve their accessibility for patient understanding. Models predominantly used were GPT-4 and BERT, along with GPT-3.5, Llama 2, Med42, GPT-4V, and Gemini Pro. The large language model GPT-4 demonstrated high diagnostic accuracy in identifying brain tumors (73.0%), myocarditis (83.0%), and in making decisions on invasive procedures for acute coronary syndrome (86.0%). In turn, it demonstrated low diagnostic accuracy for nervous system disorders of various etiologies (50.0%) and for musculoskeletal diseases (43.0%). The BERT model exhibited high diagnostic accuracy in detecting pulmonary nodules (99.0%) and signs of intracranial hemorrhage (sensitivity and specificity: 97.0% and 90.0%, respectively), as well as in report classification (accuracy: 84.3%).

Most articles (88.9%) carried a high risk of bias. The main reasons for this included small and imbalanced sample sizes, overlap between training and test datasets, and insufficiently precise preparation and description of reference standards.

CONCLUSION: The diagnostic performance of large language models varies significantly across articles. Their clinical implementation requires standardized, methodologically robust research, including larger and more balanced samples, optimization of the structure and volume of datasets, separation of training and testing samples, thorough preparation and description of reference standards, as well as the accumulation of empirical data for specific radiological tasks.

Keywords: artificial intelligence; large language models; radiological diagnostics; radiology report; systematic review.

To cite this article:
Vasilev YuA, Reshetnikov RV, Nanova OG, Vladzymyrskyy AV, Arzamasov KM, Omelyanskaya OV, Kodenko MR, Erizhokov RA, Pamova AP, Seradzhi SR, Blokhin IA, Gonchar AP, Gelezhe PB, Akhmedzyanova DA, Shumskaya YuF. Application of Large Language Models in Radiological Diagnostics: A Scoping Review. *Digital Diagnostics.* 2025;6(2):268–285. DOI: 10.17816/DD678373 EDN: QSANCA

Submitted: 06.05.2025 **Accepted:** 12.06.2025 **Published online:** 17.06.2025

ECO•VECTOR

Article can be used under the CC BY-NC-ND 4.0 International License
© Eco-Vector, 2025

DOI: https://doi.org/10.17816/DD678373 EDN: QSANCA

大语言模型在放射诊断中的应用：范围综述

Yuriy A. Vasilev[1], Roman V. Reshetnikov[1], Olga G. Nanova[1], Anton V. Vladzymyrskyy[1], Kirill M. Arzamasov[1], Olga V. Omelyanskaya[1], Maria R. Kodenko[1], Rustam A. Erizhokov[1], Anastasia P. Pamova[1], Seal R. Seradzhi[1], Ivan A. Blokhin[1], Anna P. Gonchar[1,2], Pavel B. Gelezhe[1], Dina A. Akhmedzyanova[1], Yuliya F. Shumskaya[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] Moscow City Hospital named after S.S. Yudin, Moscow, Russia

摘要

论证。现代大语言模型具备在放射诊断中应用于解决广泛常规任务的潜力。

目的：综述大语言模型在放射诊断中的应用范围，分析其使用场景，并评估相关研究的方法学质量。

方法。开展两轮文献检索：初步检索（PubMed和eLibrary）聚焦于具备详实方法学的全文研究，补充检索（PubMed）旨在广泛覆盖2023–2025年间大语言模型在放射诊断中应用的各种情境。提取了书目信息、研究任务的表述、大语言模型的应用场景、疾病谱、关键方法学参数，以及诊断效能的定量与定性指标，涵盖模型本身及参与专家，包括其人数与经验。采用改良版QUADAS-CAD问卷对研究质量进行评估。

结果。初步检索纳入9项研究，补充检索纳入216项。共识别出在放射诊断中应用大语言模型的9种主要场景。其中最常见的是为提升患者理解而对放射学报告进行改写。最常使用的模型包括GPT-4和BERT，以及GPT-3.5、Llama 2、Med42、GPT-4V和Gemini Pro。大语言模型GPT-4在脑肿瘤（73.0%）、心肌炎（83.0%）以及急性冠状动脉综合征中介入治疗决策（86.0%）方面表现出较高的诊断准确性。但在诊断不同病因的神经系统疾病（50.0%）和肌肉骨骼疾病（43.0%）方面准确性较低。BERT模型在肺结节检测（99.0%）和颅内出血征象识别（灵敏度97.0%、特异度90.0%）方面表现优异，在放射学报告分类中准确率为84.3%。

大多数研究（88.9%）存在系统性偏倚的可能。其主要原因包括：样本量小且分布不均、训练集与测试集重叠、参考标准准备和描述不够严谨。

结论。大语言模型的诊断准确性在不同研究间差异显著。其进入临床实践前，亟需开展标准化且方法学严谨的研究，包括扩大并平衡样本量、优化数据集结构与规模、明确划分训练集与测试集、严谨制定和描述参考标准，并针对特定放射诊断任务积累实证数据。

关键词：人工智能；大语言模型；放射诊断；放射学报告；系统综述。

引用本文：
Vasilev YuA, Reshetnikov RV, Nanova OG, Vladzymyrskyy AV, Arzamasov KM, Omelyanskaya OV, Kodenko MR, Erizhokov RA, Pamova AP, Seradzhi SR, Blokhin IA, Gonchar AP, Gelezhe PB, Akhmedzyanova DA, Shumskaya YuF. 大语言模型在放射诊断中的应用：范围综述. *Digital Diagnostics.* 2025;6(2):268–285. DOI: 10.17816/DD678373 EDN: QSANCA

收到：06.05.2025 **接受：**12.06.2025 **发布日期：**17.06.2025

ECO•VECTOR
Article can be used under the CC BY-NC-ND 4.0 International License
© Eco-Vector, 2025

ОБОСНОВАНИЕ

Большие языковые модели (БЯМ) — это модели искусственного интеллекта (ИИ), основанные на технологиях глубокого обучения, способные обрабатывать естественный язык, генерировать и интерпретировать тексты. Несмотря на то что первые языковые модели созданы ещё несколько десятилетий назад, значительный прогресс в этой области достигнут только в 2018 году с появлением архитектуры трансформеров, которая обеспечила возможность обучения моделей на больших объёмах текстовых данных и стала основой построения современных БЯМ [1–3].

Применение БЯМ в лучевой диагностике активно исследуют в последние годы. Современные модели демонстрируют высокий уровень подготовки, позволяющий им успешно проходить экзаменационные тесты по лучевой диагностике, сопоставимо со специалистами или, в отдельных случаях, превосходя их по точности ответов [1, 2]. БЯМ могут выполнять полезные для врачей-рентгенологов функции: генерировать структурированные рентгенологические заключения, читать неструктурированные протоколы и осуществлять диагностику с их помощью, а также извлекать информацию о клинически значимых изменениях [3, 4].

Однако степень практической применимости БЯМ в лучевой диагностике, их диагностическая точность и воспроизводимость результатов в разных клинических задачах остаются недостаточно изученными.

ЦЕЛЬ

Провести обзор предметного поля применения БЯМ в лучевой диагностике с анализом возможных сценариев их использования и оценкой качества методологии соответствующих исследований.

МЕТОДЫ

Настоящее исследование выполнено в соответствии с рекомендациями PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) по проведению обзоров предметного поля [5].

Стратегия поиска

Первичный поиск исследовательских работ [6] осуществляли с использованием двух поисковых систем PubMed и eLibrary. Временной интервал: 2023–2025 гг.

Поисковый запрос в PubMed выглядел следующим образом: (((Large language models) OR (LLM)) OR (Natural language processing)) AND (Radiology).

В разделе фильтров выбраны следующие опции: Free Full Text, Full Text, Classical Article, Clinical Study, Comparative Study, Controlled Clinical Trial, Multicenter Study с целью отбора наиболее убедительных доказательств.

Поисковый запрос в eLibrary осуществляли с помощью ключевых слов: «большие языковые модели», «рентгенология».

Кроме того, проводили ручной поиск по спискам литературы в уже отобранных публикациях и двух обзорах [3, 4].

Поисковая стратегия включала два этапа.

- Сначала проанализировали названия и аннотации всех публикаций, найденных по заданным поисковым запросам. Для дальнейшего анализа отобрали статьи, соответствующие целям настоящего исследования. Исключили протоколы и планы исследований, работы, посвящённые применению БЯМ в экзаменационных задачах по лучевой диагностике, а также публикации, не содержащие данных о диагностических показателях моделей.
- На втором этапе проанализировали полные тексты и их доступность из отобранного пула работ и составили выборку для основного анализа.

Дополнительный поиск проводили в поисковой системе PubMed без ограничений по наличию доступного текста с использованием следующего поискового запроса: (((Large language models) OR (LLM)) OR (Natural language processing)) AND (Radiology) NOT review[pt] NOT letter[pt] NOT editorial[pt] NOT comment[pt].

Временной интервал: 2023–2025 гг. Дата последнего проведённого поиска: 26.05.2025.

В этом случае в анализ включали все публикации, имеющие английское резюме. Рассматривали исследовательские статьи в рецензируемых научных журналах. Критерии исключения:

- статьи, не являющиеся оригинальными исследованиями;
- статьи, не посвящённые тематике генеративного ИИ;
- статьи, не рассматривающие вопросы практического применения БЯМ в лучевой диагностике, включая публикации, посвящённые определению их возможностей при ответах на тестовые вопросы по медицинским тематикам.

Отбор работ проводили восемь экспертов. Финальный список включённых публикаций дополнительно оценивали два эксперта. В качестве экспертов выступали научные сотрудники с опытом работы в медицинской информатике более 10 лет.

Извлечение информации и оценка качества статьи

Из полных текстов отобранных статей извлекали следующие характеристики:

- библиометрические данные — имя первого автора, название публикации, год выхода, DOI (Digital Object Identifier — цифровой идентификатор объекта), наименование журнала, его импакт-фактор, страну проведения исследования;
- формулировку исследовательской задачи, тип изучаемой патологии, направление и основные характеристики исследований (объём выборки, дизайн исследования, наличие внешней валидации, а также использованный тип БЯМ);

- значения диагностической точности моделей (чувствительность, специфичность, точность);
- сравнение диагностической эффективности ИИ и медицинских специалистов;
- количество врачей и уровень их квалификации;
- наличие конфликта интересов либо его отсутствие.

Оценку качества отобранных публикаций провели с помощью модифицированного опросника QUADAS-CAD (Quality Assessment of Diagnostic Accuracy Studies Computer-Aided Detection), разработанного для исследований с использованием ИИ [7].

Для результатов дополнительного поиска извлекали следующие данные:

- библиометрические данные;
- цель исследования и применяемый методологический подход;
- модальность лучевой диагностики;
- исследуемая патология;
- тип использованной БЯМ;
- характеристики исследуемой популяции;
- дизайн исследования;
- количество участвующих врачей и их опыт;
- тип использованного референс-стандарта;
- тип данных, с которыми работала БЯМ;
- общий вывод исследования.

Оценку качества статей для результатов дополнительного поиска не проводили.

РЕЗУЛЬТАТЫ

Результаты первичного поиска

Результаты первичного систематического поиска литературы продемонстрированы на рис. 1.

Базовые характеристики статей

Базовые характеристики отобранных публикаций и соответствующих исследований представлены в табл. 1–3.

Диагностику на основе рентгенологического протокола проводили в трёх исследованиях [8–10], кроме того, в одном из них дополнительно использовали изображения [8]. Одно из исследований выполнено в области нейрорентгенологии (диагностика опухолей головного мозга) [9], другое — диагностики заболеваний опорно-двигательной системы [8], и ещё одно посвящено диагностике миокардитов [10].

Детекцию находок в протоколах проводили в двух работах [11, 12]. Из них одна работа выполнена в области диагностики злокачественных новообразований лёгкого (детекция лёгочных узелков) [11], а вторая посвящена выявлению признаков внутричерепного кровоизлияния [12].

Эффективность ответа БЯМ на клинические вопросы, включающих как текст, так и изображения, исследована в одной работе [13]. При этом в анализ включали различные патологии, а их распределение не приведено в тексте публикации.

Диагностика патологий нервной системы на основе текста истории болезни и клинически значимых изменений по результатам рентгенологического исследования проведена в одной работе [14]. Классификация важности протоколов компьютерной томографии (КТ) головного мозга выполнена также в одном исследовании [15].

Оценка возможности принятия решения о выполнении инвазивного вмешательства на основе клинических данных и протокола КТ при остром коронарном синдроме проведена в одном исследовании [16].

В обзор включены публикации, описывающие пять многоцентровых исследований [9, 10, 12–14] и четыре одноцентровых [8, 11, 15, 16]. Исследования с проведением

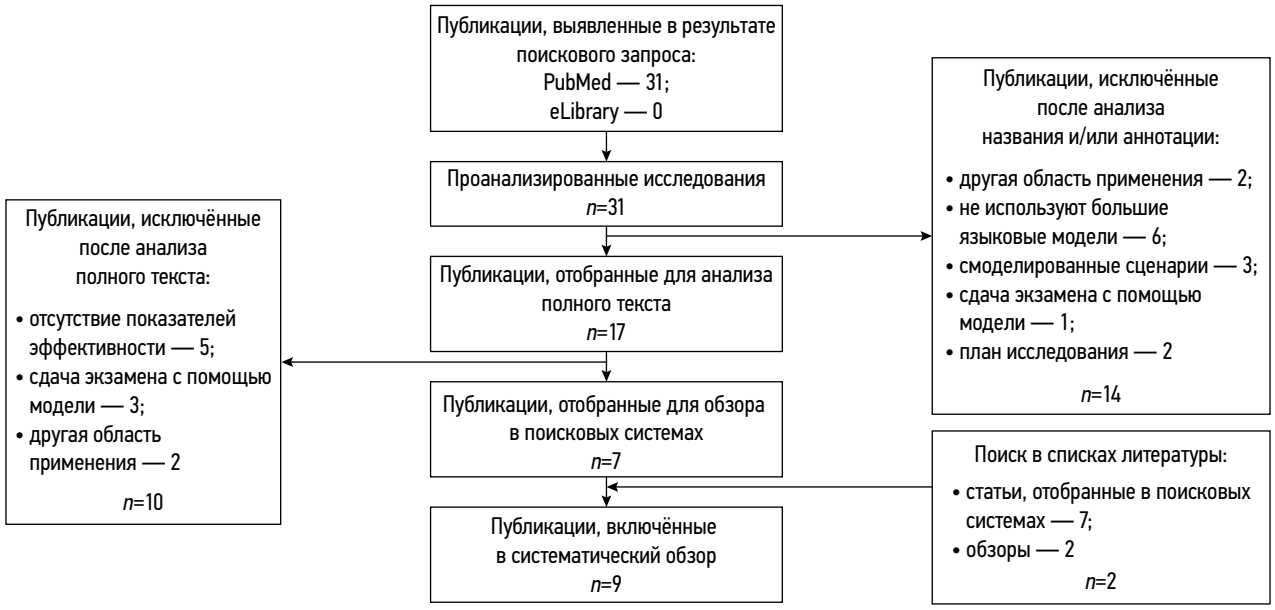

Рис. 1. Блок-схема первичного систематического поиска литературы.

Таблица 1. Список публикаций, включённых в обзор, и их базовые характеристики

Авторы	Год	Название	Журнал	ИФ	Задача исследования	Страна	Модальность	Исследуемая патология
Y. Mitsuyama и соавт. [9]	2024	Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors	European Radiology	4,7	Оценка диагностических возможностей GPT-4® (OpenAI, США) при анализе опухолей головного мозга, сравнение с заключениями нейрорентгенологов и рентгенологов общей практики	Япония	МРТ	Опухоли мозга
D. Horiuchi и соавт. [8]	2025	ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology	European Radiology	4,7	Оценка диагностической точность GPT-4® и GPT-4V® (OpenAI, США) и врачей-рентгенологов в области костно-мышечной рентгенологии	Япония	УЗИ, МРТ, КТ	Заболевания опорно-двигательной системы
E. Grolleau и соавт. [11]	2024	Incidental pulmonary nodules: Natural language processing analysis ofradiology reports	Respiratory Medicine and Research	2,2	Анализ выявления лёгочных узелков в течение года: частота, динамика и клинико-рентгенологические характеристики	Франция	КТ грудной клетки	Опухоли лёгкого
T. Han и соавт. [13]	2024	Comparative analysis of multimodal large language model performance on clinical vignette questions	JAMA	64	Оценка эффективности ответа больших языковых моделей на клинические вопросы	Германия	Неясно	Неясно
D. Horiuchi и соавт. [14]	2024	Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases	Neuroradiology	2,4	Оценка диагностической эффективности GPT-4® (OpenAI, США) в нейрорентгенологии	Япония	Неясно	Патологии нервной системы разного происхождения
T. Wataya и соавт. [15]	2024	Comparison of natural language processing algorithms in assessing the importance of head computed tomography reports written in Japanese	Japanese Journal of Radiology	2,9	Разработка 5-балльной шкалы типизации рентгенологических заключений и сравнение эффективности алгоритмов обработки естественного языка при их оценке на примере заключений КТ головы на японском языке	Япония	КТ	Патологии головы
K. Kaya и соавт. [10]	2024	Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study	Journal of Cardiovascular Magnetic Resonance	9,1	Оценка производительности GPT-4® (OpenAI, США) для принятия медицинских решений на основе заключений МРТ сердца при подозрении на миокардит	Германия	МРТ	Миокардиты
А.Н Хоружная и соавт. [12]	2024	Сравнение ансамбля алгоритмов машинного обучения и BERT для анализа текстовых описаний КТ головного мозга на предмет наличия внутричерепных кровоизлияний	Современные технологии в медицине	0,7 РИНЦ 1,243	Обучение и тестирование моделей машинного обучения, сравнение их производительности с моделью языка BERT® (Google, США), предварительно обученной на медицинских данных, для выполнения простой бинарной классификации	Россия	КТ	Внутричерепное кровоизлияние
A. Cagnina и соавт. [16]	2025	Assessing the need for coronary angiography in high-risk non-ST-elevation acute coronary syndrome patients using artificial intelligence and computed tomography	The International Journal of Cardiovascular Imaging	1,9	Оценка эффективности GPT-4® (OpenAI, США) в определении необходимости инвазивной коронарной ангиографии у пациентов с острым коронарным синдромом высокого риска без подъёма сегмента *ST* на основе как стандартных клинических данных, так и результатов коронарной КТ	Швейцария	КТ	Острый коронарный синдром

Примечание. ИФ — импакт-фактор журнала; МРТ — магнитно-резонансная томография; КТ — компьютерная томография; УЗИ — ультразвуковое исследование; РИНЦ — российский индекс научного цитирования; США — Соединённые Штаты Америки.

Таблица 2. Основные характеристики исследований, представленных в публикациях, включённых в систематический обзор

Авторы	Модели	Время исследования	Критерии включения	Критерии исключения	Дизайн исследования	Число врачей и их опыт	Референс-тест	Внешней валидации
Y. Mitsuyama и соавт. [9]	GPT-4® (OpenAI, США)	2017–2021 гг.	—	Случаи рецидива (96 рентгенологических заключений)	• многоцентровое (2 центра); • ретроспективное	• 3 нейрорентгенолога; • 4 врача-рентгенолога общей практики	Патологический диагноз опухоли, извлечённой нейрохирургическим путём	Нет
D. Horiuchi и соавт. [8]	GPT-4® (OpenAI, США); GPT-4V® (OpenAI, США)	С января 2014 по сентябрь 2023 г.	—	Отсутствие текстовых заключений визуализации (22 случая)	• одноцентровое; • ретроспективное	• 2 врача-рентгенолога с опытом работы 4 и 7 лет	Опубликованные диагнозы	Нет
E. Grolleau и соавт. [11]	BERT® (Google, США)	2020 год	—	—	• одноцентровое; • ретроспективное	—	Анализ заключений врачами	Нет
T. Han и соавт. [13]	GPT-4® (OpenAI, США); GPT-3.5® (OpenAI, США); Llama 2® (Meta, США); Med42® (Hippocratic AI, США); GPT-4V® (OpenAI, США); Gemini Pro® (Google DeepMind, Великобритания)	С января 2017 по август 2023 г.	Клинические задачи из журнала JAMA; клинические изображения из журнала New England Journal of Medicine	—	• многоцентровое; • ретроспективное	—	Неясно	Нет
D. Horiuchi и соавт. [14]	GPT-4® (OpenAI, США)	С октября 2021 по сентябрь 2023 г.	—	—	• многоцентровое; • ретроспективное	—	Опубликованные диагнозы	Нет
T. Wataya и соавт. [15]	BERT® (Google, США)	2020 год	Основной текст заключений	Отсутствие описаний мозга или черепа (10 заключений)	• одноцентровое; • ретроспективное	• 6 нейрорентгенологов с опытом 4,4, 4 и 6 лет (референс-тест)	Ручная аннотация заключений рентгенологами	Нет
K. Kaya и соавт. [10]	GPT-4® (OpenAI, США)	—	—	Существенные артефакты или плохое качество изображений, препятствующее установлению диагноза	• многоцентровое (8 центров); • ретроспективное	• 3 врач-рентгенолога с опытом работы 1, 2 и 4 года; • 2 врача-рентгенолога с опытом работы 8 и 10 лет (референс-тест)	Чтение заключений врачами-рентгенологами	Нет
А.Н Хоружая и соавт. [12]	• дерево принятия решений; • «случайного леса»; логистической регрессии; • ближайших соседей; • опорных векторов; • Catboost; • XGboost; MedRuBertTiny2	—	—	—	• многоцентровое (56 центров); • ретроспективное	—	Неясно	Нет
A. Cagnina и соавт. [16]	ChatGPT-4® (OpenAI, США)	2020 и 2022 гг.	Все пациенты с показанием к инвазивной коронарной ангиографии	—	• одноцентровое; • проспективное	—	Фактические результаты инвазивной коронарной ангиографии	Нет

DOI: https://doi.org/10.17816/DD678373

Таблица 3. Характеристики выборок и патология, представленные в исследованиях

Авторы	Пациенты/ заключения, *n*	Возрастные группы, лет	Соотношения полов	Раса	Представленность патологии	Конфликт интересов
Y. Mitsuyama и соавт. [9]	150 рентгенологических заключений	Институт A: 53,0±17,0; Институт B: 69,0±15,0	56 мужчин	Нет	**Институт A:** менингиома — 34; аденома гипофиза — 17; невринома — 12; ангиома — 5; краниофарингиома — 4; гемангиобластома — 4; глиома высокой степени злокачественности — 10; глиома низкой степени злокачественности — 3; эпидермальная киста — 2; саркома — 2; арахноидальная киста — 1; хордома — 1; лимфома — 1; метастатические опухоли — 1; киста расщелины Ратке — 1; центральная нейроцитома — 1; **Институт B:** менингиома — 16; аденома гипофиза — 6; невринома — 4; ангиома — 0; краниофарингиома — 0. гемангиобластома — 1; высокой степени злокачественности — 5; глиома низкой степени злокачественности — 1; эпидермальная киста — 0; саркома — 0; арахноидальная киста — 0; хордома — 0; лимфома — 10; метастатические опухоли — 6; киста расщелины Ратке — 2; центральная нейроцитома — 0	Нет
D. Horiuchi и соавт. [8]	106 случаев патологии опорно-двигательной системы из журнала Skeletal Radiology, раздел «Test Yourself»	Нет	Нет	Нет	Все случаи разделены на две группы: опухолевую (*n*=45) и неопухолевую (*n*=61) в соответствии с классификации опухолей мягких тканей и костей Всемирной организации здравоохранения 2020 года; Случаи в опухолевой группе разделены на: случаи опухолей костей (*n*=24) и мягких тканей (*n*=22); Случаи в неопухолевой группе разделены в зависимости от этиологии заболевания: мышечные/мягкие ткани/нервы (*n*=12), артрит/артропатия (*n*=10), инфекция (*n*=8), врождённая/развивающая аномалия и дисплазия (*n*=6), травма (*n*=6), метаболическое заболевание (*n*=5), анатомический вариант (*n*=4) и другие (*n*=10)	Нет
E. Grolleau и соавт. [11]	101 703 описаний снимков компьютерной томографии	64,7±19,6	55,2% мужчин	Нет	Упоминание узелков в заключении — 971 (48,8%); отсутствие упоминания — 1020 (51,2%)	Нет
T. Han и соавт. [13]	Неясно	Нет	Нет	Нет	—	Да
D. Horiuchi и соавт. [14]	100 описаний	Нет	Нет	Нет	Все случаи классифицированы по анатомическому местоположению: мозг (n=77), позвоночник (*n*=11), голова и шея (*n*=12); Случаи с поражением мозга разделены на группы: с опухолями центральной нервной системы (*n*=19) и другой локализации (*n*=58)	Нет
T. Wataya и соавт. [15]	3738 заключений компьютерной томографии головы	Нет	Нет	Нет	Доля каждого типа заключения во всём наборе данных составила: отсутствие патологических изменений — 15,0%; незначительные изменения — 26,7%; изменения, требующие планового наблюдения — 44,2%; изменения, требующие тщательного последующего наблюдения — 7,7%; изменения, требующие обследования и терапии пациента — 6,4%	Да
K. Kaya и соавт. [10]	396 пациентов	Миокардит: 38,6±17,7; его отсутствие: 44,4±17,6	Миокардит: 41 женщина и 130 мужчин; его отсутствие: 97 женщин и 128 мужчин	—	Неишемические кардиомиопатии — 60,1%; токсичность, вызванная химиотерапией — 0,3%; дилатационная кардиомиопатия — 6,1%; гипертрофическая кардиомиопатия — 3,0; некомпактная кардиомиопатия — 0,3%; миокардит — 41,2; перикардит — 5,8%; саркоидоз — 0,8%; кардиомиопатия Такоцубо — 1,8%; ишемическая кардиомиопатия — 5,8%; приобретённый порок сердца — 1,5%; неопределённые находки — 1,5%; нет находок — 36,1%	Да
А.Н Хоружая и соавт. [12]	34 188 заключений компьютерной томографии головного мозга	Нет	Нет	Нет	Из 1194 заключений тестовой выборки 927 без признаков внутричерепного кровоизлияния, 267 с признаками внутричерепного кровоизлияния	Нет
A. Cagnina и соавт. [16]	86 пациентов	62,0±13,0	27% женщин	Нет	—	Нет

внешней валидации не выявлены. Только одно исследование имело проспективный дизайн [16], остальные восемь были ретроспективные.

На больших по объёму выборках выполнено три исследования:

- 101 703 заключения КТ [11];
- 34 188 протоколов КТ головного мозга [12];
- 3738 протоколов КТ головы [15].

В пяти исследованиях выборки были относительно небольшими (86–396 случаев) [8–10, 14, 16]. В одном исследовании объём выборки не приведён [13]. Демографические характеристики описаны в четырёх публикациях [9–11, 16], при этом в большинстве случаев не сбалансированы по этим показателям, за исключением одного исследования [11]. Распределение патологий в выборке охарактеризовано в семи работах [8–12, 14, 15], однако во всех случаях наблюдали дисбаланс по представленности различных нозологических форм.

В двух исследованиях использовали протоколы, составленные по данным МРТ [9, 10]; четырёх — заключения КТ [11, 12, 15, 16]; в одном исследовании — любые доступные модальности [8]. В двух публикациях тип визуализации не уточнён [13, 14].

Самыми распространёнными архитектурами ИИ были модели на основе ChatGPT, в частности GPT-4® [OpenAI, Соединённые Штаты Америки (США)], использованная в пяти работах [8–10, 13, 14, 16]. Модель BERT® (Google, США) применяли в трёх исследованиях [11, 12, 15].

Сравнение разных БЯМ проводили в трёх работах [8, 12, 13]. Так, D. Horiuchi и соавт. [8] сравнивали эффективность модели GPT-4® (OpenAI, США), функционирующей на основе текстового ввода, и её мультимодального аналога GPT-4V, способного анализировать как текстовую, так и визуальную информацию. T. Han и соавт. [13] провели сравнение эффективности четырёх БЯМ — GPT-4® и GPT-3.5® (OpenAI, США), Llama 2® (Meta, США), Med42® (Hippocratic AI, США) (две последние в открытом доступе), а также двух мультимодальных моделей, обрабатывающих как текст, так и изображения — GPT-4V® (OpenAI, США), Gemini Pro (Google DeepMind, Великобритания). А.Н. Хоружая и соавт. [12] сравнивали диагностические показатели эффективности семи моделей на основе методов машинного обучения, их ансамблевого объединения и модели BERT® (Google, США).

Сравнительная оценка диагностических показателей БЯМ и врачей проведена в четырёх исследованиях [8–10, 13].

О конфликте интересов заявлено в трёх работах [10, 13, 15], в шести — об его отсутствии. Включённые исследования не анализировали временные затраты и потенциальную экономическую эффективность внедрения БЯМ в клиническую практику.

Результаты дополнительного поиска

Результаты дополнительного систематического поиска литературы продемонстрированы на рис. 2. Основные характеристики отобранных публикаций и соответствующих исследований представлены в Приложении 1. Большая часть этих исследований проведена в США (78 публикаций, 36,1%). Второе место по количеству публикаций занимают Германия и Япония (по 23 работы, или 10,6%), за которыми следует Китайская Народная Республика (20 работ, 9,3%). Всего в анализ включены исследования, проведённые в 25 странах, представленные в 216 публикациях.

Типы данных, которые анализировали во включённых исследованиях, представлены преимущественно текстовой информацией (185 работ, 85,6%). Медицинские изображения использовали в 16 исследованиях (7,4%), сочетание текстовых и визуальных данных — 14 (6,5%), в одном случае анализировали аудиозаписи (0,5%).

Среди модальностей лучевой диагностики наибольшее число публикаций связано с использованием КТ (80 работ, 37,0%). Второе место заняла магнитно-резонансная томография (МРТ) (55 работ, 25,5%), третье — рентгенография (39 работ, 36,1%). В отдельных исследованиях также применяли данные ультразвукового исследования (УЗИ), маммографии, позитронно-эмиссионной томографии, совмещённой с компьютерной томографией (ПЭТ/КТ), и сцинтиграфии.

Рис. 2. Блок-схема дополнительного систематического поиска литературы.

Таблица 4. Области применения больших языковых моделей в лучевой диагностике

Сценарий	Число работ, *n* (%)	Положительный эффект	Ограничения
Упрощение текстов рентгенологических протоколов	76 (35,2)	Повышение доступности информации для пациентов, снижение барьеров в коммуникации	• потеря точности и детализации при упрощении сложной медицинской информации
Выявление патологических признаков	31 (14,3)	Раннее выявление заболеваний, увеличение диагностической точности	• ложноположительные и ложноотрицательные результаты; • необходимость клинической верификации
Проведение дифференциальной диагностики	27 (12,5)	Поддержка в сложных диагностических случаях, уменьшение времени до постановки диагноза	• риск предложений неверных диагнозов; • зависимость от качества обучающих данных
Извлечение структурированных данных	22 (10,2)	Ускорение анализа больших объёмов данных, улучшение аналитики	• ошибки при автоматической разметке; • необходимость валидации данных экспертами
Генерация рентгенологических протоколов	15 (6,9)	Снижение времени подготовки заключений, стандартизация и повышение производительности труда	• ошибки в сложных случаях; возможная потеря индивидуального подхода
Классификация исследований по системам оценки данных (например, PI-RADS, BI-RADS)	11 (5,1)	Повышение точности диагностической категоризации, стандартизация клинических решений	• риск ошибок в нестандартных случаях; • зависимость от качества исходных данных
Автоматизация обработки и интерпретации данных в рентгенологии	6 (2,8)	Автоматизация сложных процессов обработки медицинских текстов и изображений, повышение эффективности работы	• необходимость строгого мониторинга корректности работы моделей, риск неправильной интерпретации из-за ошибок в клинических данных
Прогноз исходов заболеваний	4 (1,8)	Улучшение планирования лечения	• проблемы с обобщаемостью моделей на разные популяции; необходимость больших данных
Выявление ошибок в рентгенологических протоколах	3 (1,4)	Повышение качества медицинской документации, снижение вероятности ошибок	• не все ошибки возможно автоматически обнаружить; • риск пропуска редких ошибок

Примечание. PI-RADS (Prostate Imaging Reporting and Data System) — система оценки и представления данных магнитно-резонансной томографии предстательной железы; BI-RADS (Breast Imaging Reporting and Data System) — система оценки и представления данных визуализации молочных желёз.

Области практического применения БЯМ в лучевой диагностике классифицированы на 9 основных сценариев. Наиболее распространённым среди них оказалось переформулирование текстов рентгенологических заключений с целью повышения их доступности и понятности для пациентов (табл. 4). Кроме того, 21 исследование (9,7%) посвящено сравнительному анализу различных БЯМ в контексте выполнения диагностических задач. Включение этих работ в обзор обусловлено их вкладом в определение оптимальных моделей для задач лучевой диагностики. При этом важно учитывать, что их эффективность может зависеть от специфики поставленной задачи, характеристик исходных данных и степени адаптации модели под клинический контекст. Результаты таких исследований свидетельствуют о наличии закономерности: более крупные и современные модели, как правило, демонстрируют более высокую точность.

Следует отметить, что значительная часть исследований, посвящённых применению БЯМ в медицине, характеризуется низким качеством как методологического исполнения, так и представления результатов. Несмотря на существование нескольких групп рекомендаций по стандартизации проведения и отчётности исследований в области ИИ в здравоохранении [17, 18], их соблюдение зафиксировано только в 2,3% работ, включённых в обзор на основании дополнительного поиска.

Диагностическая точность моделей и врачей

В проанализированных работах наблюдают большое разнообразие как в формулировке задач, решаемых с помощью БЯМ в лучевой диагностике, так и в спектре исследуемых патологий. Рассмотрим подробнее диагностические показатели, полученные для каждой из задач.

Диагностика на основе рентгенологических протоколов проведена в трёх исследованиях [8–10]. При этом полученные значения диагностических показателей варьируют от низких до высоких. Во всех трёх работах анализировали диагностические возможности модели GPT-4® (OpenAI, США). В исследовании Y. Mitsuyama и соавт. [9], выполненном в области нейрорентгенологии (диагностика опухолей головного мозга), данная модель продемонстрировала достаточно высокие показатели: точность составила 73,0% при установлении заключительного диагноза и 94,0% — при проведении дифференциальной диагностики (среди трёх возможных вариантов). При этом показатели точности врачей-рентгенологов были сопоставимыми и составили

65,0–79,0 и 73,0–89,0% соответственно. В исследовании D. Horiuchi и соавт [8], посвящённом патологиям опорно-двигательной системы, модель GPT-4® (OpenAI, США) показала более низкие результаты: точность составила 43,0 и 58,0% при установлении заключительного диагноза и проведении дифференциальной диагностики (среди трёх вариантов) соответственно. Аналогично, врачи-рентгенологи продемонстрировали низкие показатели: 41,0–53,0 и 58,0–67,0% соответственно. Различия точности между моделью и врачами не достигали статистической значимости (p=0,78 и p=0,99 при сравнении со стажёром-рентгенологом; p=0,22 и p=0,26 при сравнении с сертифицированным врачом-рентгенологом, χ^2-тест). Использование мультимодальной модели GPT-4V® (OpenAI, США) в данном исследовании оказалась неэффективным: её точность составила 8,0 и 14,0% соответственно. В то же время применение GPT-4® (OpenAI, США) в качестве вспомогательного инструмента привело к незначительному улучшению точности врачей (на 3,0–6,0%), однако статистическая значимость прироста не указана. Использование GPT-4V® (OpenAI, США) в аналогичном качестве не дало прироста точности ни в одном из случаев.

В исследовании K. Kaya и соавт. [10], посвящённом диагностике миокардита, модель GPT-4® (OpenAI, США) продемонстрировала высокие показатели диагностической эффективности: чувствительность, специфичность и точность составили 90,0, 78,0 и 83,0% соответственно. Эти показатели сопоставимы с результатами врача с опытом работы один год (90,0, 84,0 и 86,0% соответственно; p=0,14, χ^2-тест), однако статистически значимо уступали метрикам более опытных специалистов: 86,0–85,0, 91,0–96,0 и 89,0–91,0% соответственно (p <0,01).

В двух исследованиях, посвящённых автоматическому выявлению патологических находок в протоколах, использовали модель BERT® (Google, США) [11, 12]. В обоих случаях БЯМ продемонстрировала высокие диагностические показатели. Так, в работе E. Grolleau и соавт. [11], направленной на детекцию лёгочных узелков, модель достигла чувствительности, специфичности и точности 98,0, 99,0 и 99,0% соответственно. Аналогично, в исследовании А.Н. Хоружая и соавт. [12], посвящённом выявлению признаков внутричерепного кровоизлияния, чувствительность и специфичность модели BERT® (Google, США) составили 97,0 и 90,0% соответственно. Эти значения статистически значимо превосходили аналогичные показатели, полученные при применении других методов машинного обучения, обученных на тех же данных. В частности, медианные значения чувствительности и специфичности для всех использованных моделей машинного обучения с разными способами векторизации текста составили 94,0 и 78,0% соответственно, а для ансамблевых моделей машинного обучения — 94,0 и 84,0%. Различия были статистически значимыми (p <0,05, тест МакНемара).

T. Han и соавт. [13] оценивали эффективность ответов БЯМ на клинические вопросы, включающие как текстовые данные, так и изображения. Однако более детальная информация о содержании вопросов, используемых для тестирования моделей, в статье не приводится. В сравнительном анализе участвовали четыре БЯМ: GPT-4® и GPT-3.5® (OpenAI, США), Llama 2® (Meta, США), Med42® (Hippocratic AI, США), а также две мультимодальные модели, способные анализировать как текст, так и изображения — GPT-4V® (OpenAI, США) и Gemini Pro® (Google DeepMind, Великобритания). Согласно результатам, точность GPT-4V® (OpenAI, США) при ответе на 140 вопросов журнала *JAMA* составила 73,3%, а при ответе на 348 вопросов журнала *NEJM (The New England Journal of Medicine)* — 88,7%, что статистически значимо превосходит как другие модели (медианные значения точности составили 53,6 и 51,4% соответственно; p <0,05, t-тест), так и показатели врачей (51,4% при ответе на вопросы журнала *NEJM*). Высокие показатели точности GPT-4V® (OpenAI, США), продемонстрированные в данном исследовании, резко контрастируют с результатами работы D. Horiuchi и соавт. [8].

D. Horiuchi и соавт. [14] отметили невысокие значения точности модели GPT-4® (OpenAI, США) в отношении диагностики патологий нервной системы различной этиологии на основе анализа текста истории болезни и описаний клинически значимых находок по данным рентгенологических исследований: 50,0 и 63,0% для установления заключительного диагноза и проведения дифференциальной диагностики соответственно. Эти результаты заметно уступают показателям, продемонстрированным в работе Y. Mitsuyama и соавт. [9]. D. Horiuchi и соавт. [14] не выявили статистически значимых различий точности модели в зависимости от анатомической локализации патологического процесса (головной мозг, спиной мозг, голова и шея, p=0,89, точный тест Фишера). С другой стороны, авторы зафиксировали статистически значимые различия в точности в зависимости от типа новообразования: при опухолях центральной нервной системы точность составила 16,0 и 26,0% при установлении заключительного диагноза и проведении дифференциальной диагностики соответственно, тогда как при опухолях другой локализации — 62,0 и 74,0% соответственно (p <0,01).

T. Wataya и соавт. [15] сравнивали возможности нескольких моделей ИИ классифицировать рентгенологические протоколы КТ головного мозга по степени их клинической значимости, определяемой по специально разработанной 5-уровневой шкале:

- категория 0 — отсутствие патологических изменений;
- категория 1 — незначительные изменения;
- категория 2 — изменения, требующие планового наблюдения;
- категория 3 — изменения, требующие тщательного последующего наблюдения;
- категория 4 — изменения, требующие обследования и терапии пациента.

Таблица 5. Оценка качества методологии исследований с использованием модифицированного опросника QUADAS-CAD

Вопросы	Y. Mitsuyama и соавт. [9]	D. Horiuchi и соавт. [8]	E. Grolleau и соавт. [11]	T. Han и соавт. [13]	D. Horiuchi и соавт. [14]	T. Wataya и соавт. [15]	K. Kaya и соавт. [10]	А.Н. Хоружая и соавт. [12]	A. Cagnina и соавт. [16]
Отбор пациентов (D1)									
Были ли данные (обучающие и тестовые наборы) сбалансированы по тяжести (включая отсутствие) целевой патологии?	Нет	Нет	Да	Неясно	Нет	Нет	Нет	Да	Нет
Были ли данные (обучающие и тестовые наборы) сбалансированы с точки зрения демографических факторов?	Нет	Нет	Да	Неясно	Нет	Нет	Нет	Нет	Нет
Избежало ли исследование неуместных исключений?	Неясно	Неясно	Неясно	Неясно	Да	Неясно	Да	Да	Да
Индексный тест (D2)									
Если использовали нейронную сеть, были ли наборы данных обучения и тестирования не пересекающимися?	Да	Неясно	Да	Неясно	Да	Неясно	Да	Да	Да
Если использовали нейронную сеть, был ли рационализирован размер каждого набора данных?	Нет	Нет	Нет	Нет	Нет	Нет	Нет	Да	Нет
Если использовали порог патологии, был ли он установлен заранее?	Неясно	Неясно	Неясно	Неясно	Неясно	Неясно	Неясно	Да	Да
Если использовали порог принятия решения (для ИИ), был ли он задан заранее?	Неясно	Неясно	Неясно	Неясно	Неясно	Да	Да	Да	Да
Референсный стандарт (D3)									
Может ли референсный стандарт правильно классифицировать целевое состояние?	Да	Неясно	Да	Неясно	Да	Да	Да	Да	Да
Были ли результаты референсных стандартов подготовлены или проверены с необходимым уровнем экспертизы?	Да	Неясно	Да	Неясно	Да	Да	Да	Неясно	Да
Получение результатов									
Была ли прозрачность в том, как были получены результаты?	Да	Да	Да	Неясно	Да	Да	Да	Да	Да
Все ли данные пациентов имели один и тот же референсный стандарт?	Да	Нет	Да	Неясно	Неясно	Да	Да	Неясно	Да

Примечание. Жирным шрифтом выделены сигнальные вопросы; ИИ — искусственный интеллект; QUADAS-CAD (Quality Assessment of Diagnostic Accuracy Studies Computer-Aided Detection) — специализированный модифицированный опросник для оценки риска систематических ошибок и применимости исследований в области технологий искусственного интеллекта.

Сравнивали четыре модели:

- модель логистической регрессии;
- модель рекуррентной нейронной сети BiLSTM;
- общая модель BERT® (Google, США), обученная с использованием немедицинских японских текстов из раздела Википедии;
- домен-специфичная модель BERT® (Google, США), обученная с использованием медицинских записей на японском языке из базы данных больницы Токийского университета.

В работе проводили парные множественные сравнения без корректировки уровня значимости на множественные тесты. Оба варианта языковой модели BERT® (Google, США) показали статистически значимое превосходство по сравнению с логистической регрессией (78,7%) и моделью BiLSTM (76,5%): точность составила 81,6% для общей модели и 84,3% — для домен-специфичной модели BERT® (Google, США) (p=0,001–0,02 для разных попарных сравнений, U-тест Манна–Уитни). Различие в точности между двумя моделями BERT® (Google, США) не достигло статистической значимости (p=0,06), однако авторы отмечают, что домен-специфичная модель BERT® (Google, США) продемонстрировала наилучший результат.

В исследовании A. Cagnina и соавт. [16] оценивали диагностические возможности GPT-4® (OpenAI, США) в контексте принятия решения о необходимости проведения инвазивной процедуры на основе клинических данных и протокола КТ при остром коронарном синдроме. Модель показала высокие значения чувствительности (95,0%) и точности (86,0%) при умеренном уровне специфичности (64,0%).

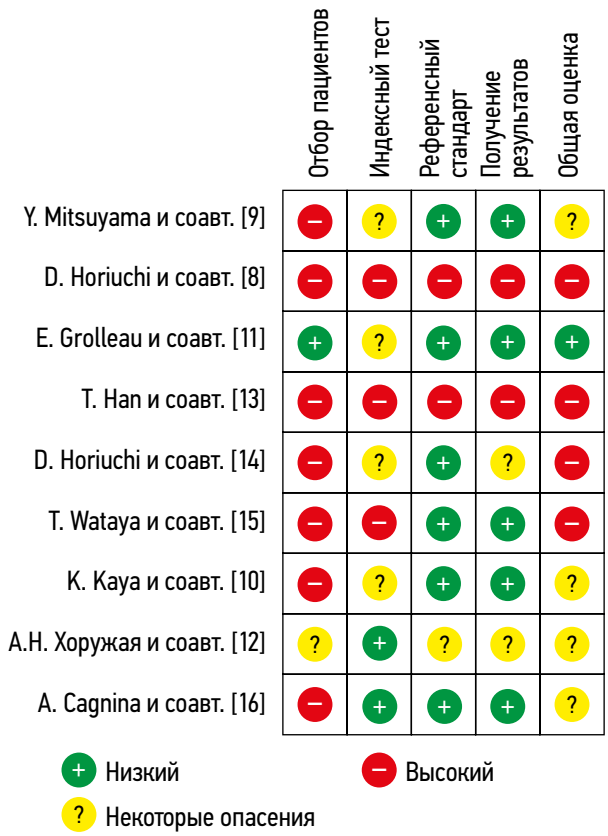

Рис. 3. Оценки риска систематической ошибки с помощью модифицированного опросника QUADAS-CAD: QUADAS-CAD (Quality Assessment of Diagnostic Accuracy Studies Computer-Aided Detection) — специализированный модифицированный опросник для оценки риска систематических ошибок и применимости исследований в области технологий искусственного интеллекта.

Оценка качества методологии исследований

Оценка качества методологии проанализированных исследований с использованием модифицированного опросника QUADAS-CAD представлена в табл. 5.

В большинстве (88,9%) проанализированных работ есть вероятность систематической ошибки и завышения диагностических показателей БЯМ (рис. 3). Рассмотрим подробнее причины:

- во-первых, это объём и состав выборок. Значительная часть исследований (66,7%) выполнена на выборках с малым числом наблюдений. Кроме того, во многих случаях выборки не сбалансированы как по демографическим характеристикам (88,9%), так и по представленности патологий и их тяжести (77,8%). В одной работе вовсе не указано, какие именно патологии легли в основу клинических вопросов [13], что является сигнальным признаком риска систематической ошибки в домене D1 (Patient Selection). В 55,6% исследований отсутствует ясность в отношении критериев отбора данных: либо они описаны недостаточно подробно, либо отсутствуют;
- во-вторых, в некоторых работах (33,3%), где использовали данные из открытых источников, обучающие и тестовые выборки могут пересекаться, о чём сообщают авторы этих исследований [8, 13, 15], что, очевидно, завышает диагностические показатели эффективности БЯМ. Этот вопрос является сигнальным в домене D2 (Index Test);
- в-третьих, в большинстве работ объём включённых наборов данных не обосновали или не рационализировали. Формирование выборок, как правило, происходило случайным образом или исходя из реалий доступности данных;
- в-четвёртых, в некоторых исследованиях не до конца ясна процедура формирования референсного стандарта: в 22,2% случаев не уточняли, каким образом он получен [8, 13]; в 44,4% — не указан факт применения единого референсного стандарта [8, 12–14]; в 33,3% — нет информации об уровне квалификации специалистов, участвовавших в подготовке и верификации референсных стандартов [8, 12, 13].

В большинстве проанализированных работ (в 6 из 9, 66,7%) присутствует высокая вероятность ошибки первого рода — нахождения статистически значимых различий между группами там, где их нет, — что обусловлено отсутствием коррекции уровня значимости при множественных сравнениях [19, 20]. Ни в одной из работ, где проводили попарные множественные сравнения [8, 10–15], такая коррекция не выполнена.

ОБСУЖДЕНИЕ

Используемые в нашей работе два варианта поиска и анализа литературы — первичный и дополнительный — позволили, с одной стороны, детально рассмотреть исследования с наиболее хорошо проработанной методологией и прозрачными результатами, с другой стороны — максимально охватить разнообразие сценариев применения БЯМ в лучевой диагностике.

Область применения больших языковых моделей в лучевой диагностике

Несмотря на небольшое число публикаций, отобранных для оценки предметного поля на этапе первичного поиска, задачи лучевой диагностики, решаемые с помощью БЯМ, отличаются разнообразием. Наиболее распространённой является диагностика на основе анализа текста рентгенологических протоколов, истории болезни пациента и клинически значимых патологических изменений. Кроме того, БЯМ использовали для детекции клинически значимых изменений в рентгенологических протоколах, классификации протоколов, принятия решений о необходимости инвазивных процедур на основе клинических данных и протоколов, а также для оценки точности ответов на клинические вопросы. Разнообразие диагностируемых патологий также достаточно велико: в проанализированных исследованиях рассматривали диагностику патологий нервной системы, включая опухоли и внутричерепные кровоизлияния, заболеваний опорно-двигательной системы, злокачественных новообразований лёгких и миокардитов.

Наиболее часто используемые в работах модели — GPT-4® (OpenAI, США) и BERT® (Google, США). Кроме того, применяли GPT-3.5® (OpenAI, США), Llama 2® (Meta, США), Med42® (Hippocratic AI, США), а также мультимодальные модели, способные работать не только с текстом, но и с изображениями, такие как GPT-4V® (OpenAI, США) и Gemini Pro® (Google DeepMind, Великобритания).

Дополнительный поиск позволил расширить спектр задач, решаемых с помощью БЯМ, и оценить частоту их постановки на большом объёме данных. Так, самой распространённой задачей в этом случае является упрощение текстов рентгенологических протоколов, что способствует снижению коммуникационных барьеров между медицинскими специалистами и пациентами, например путём создания соответствующих чат-ботов. Задачи по выявлению патологических признаков и проведению дифференциальной диагностики встречали более чем в два раза реже и занимали второе место по частоте. Наименее распространёнными сценариями применения БЯМ были прогнозирование исходов заболеваний и выявление ошибок в рентгенологических протоколах. Каждому из рассмотренных сценариев присущи определённые ограничения, включая риски ложноположительных и ложноотрицательных результатов, снижение точности, а также утрату персонализированного подхода. Для преодоления приведённых ограничений необходима детальная разработка методологии применения БЯМ в каждом конкретном сценарии и повышение качества соответствующих исследований.

В большинстве случаев при работе с БЯМ используют текстовый формат данных, реже — изображения либо их комбинацию. Аудио формат данных встречается крайне редко. По распределению используемых модальностей лидирует КТ, за ней следуют МРТ и рентгенография. Реже используют УЗИ, маммографию, ПЭТ/КТ и сцинтиграфию. Подобное распределение характерно для исследований в области лучевой диагностики в целом.

Диагностическая точность больших языковых моделей в лучевой диагностике

Диагностическая точность БЯМ сильно варьирует между работами, при этом даже эффективность одинаковой модели сильно различается при решении разных задач. Так, модель GPT-4® (OpenAI, США) продемонстрировала высокую диагностическую точность при диагностике опухолей мозга и миокардитов на основе рентгенологических протоколов. При принятии решений о проведении инвазивной процедуры на основе клинических данных и рентгенологических протоколов при остром коронарном синдроме модель GPT-4® (OpenAI, США) показала высокую чувствительности и точности при сравнительно низкой специфичности. В то же время её диагностическая эффективность была невысокой при распознавании патологий нервной системы различной этиологии (на основе текста истории болезни и рентгенологических данных), а также заболеваний опорно-двигательной системы (по протоколам визуализации). При диагностике последней группы заболеваний GPT-4V® (OpenAI, США) продемонстрировала особенно низкую точность и значительно уступала GPT-4® (OpenAI, США). Однако при тестировании качества ответов на клинические вопросы GPT-4V® (OpenAI, США), напротив, показала наилучшие результаты среди шести протестированных моделей, включая GPT-4® (OpenAI, США).

Во всех исследованиях с использованием модели BERT® (Google, США) продемонстрирована высокая диагностическая точность: при детекции лёгочных узелков, признаков внутричерепного кровоизлияния, а также при классификации протоколов КТ головного мозга по степени их клинической значимости.

В двух исследованиях, где сравнивали диагностическую точность BERT® (Google, США) с другими моделями машинного обучения (в задачах выявления признаков внутричерепного кровоизлияния и классификации протоколов КТ головного мозга), БЯМ показала превосходство по точности.

Сравнение диагностической точности БЯМ и врачей-рентгенологов проводили в четырёх исследованиях. В двух из них сравнение проводили с врачами разного уровня квалификации. При диагностике миокардита модель

показала сопоставимую точность с врачами начального уровня квалификации, уступая специалистам высокой квалификации. В случае опорно-двигательной патологии диагностическая точность GPT-4® (OpenAI, США) не имела отличий от результатов как стажёров-рентгенологов, так и сертифицированных врачей-рентгенологов. В других двух работах, где опыт врачей неизвестен, GPT-4® (OpenAI, США) показала сопоставимую точность при диагностике опухолей головного мозга, а GPT-4V® (OpenAI, США) при ответе на клинические вопросы превзошла врачей-рентгенологов.

В одной работе показано, что диагностическая точность GPT-4® (OpenAI, США) в контексте диагностики патологий нервной системы может сильно варьировать в зависимости от её типа. Так, точность диагностики опухолевых заболеваний центральной нервной системы по рентгенологическим протоколам была существенно ниже, чем другой локализации.

Оценка качества методов исследований, посвящённых применению больших языковых моделей в лучевой диагностике

Оценка рисков систематической ошибки в работах, отобранных при первичном поиске, показала, что в большинстве из них (88,89%) присутствует вероятность систематической ошибки и завышения диагностических показателей, что обусловлено определёнными причинами.

Большинство исследований выполнено на небольших выборках. Кроме того, использованные выборки не сбалансированы по демографическим показателям и по представленности патологий в них. В некоторых случаях распределение патологий в выборках вообще неясно.

Пересечение обучающей и тестовой выборок способно существенно завысить диагностические показатели тестируемых моделей. При использовании данных из открытых источников необходимо формировать выборки таким образом, чтобы исключить их перекрёстность и избежать включения случаев, потенциально присутствующих как в обучающей, так и тестовой выборке.

Во многих проанализированных исследованиях представлены исключительно значения точности — наименее информативного из диагностических показателей, — при этом не приведены другие ключевые метрики, такие как чувствительность и специфичность. Отсутствие этих показателей затрудняет оценку доли правильно классифицированных истинно положительных и истинно отрицательных результатов, что принципиально важно для медицинской практики.

В большинстве работ проводили попарные сравнения диагностических показателей между разными моделями, между моделями и врачами, а также разными патологиями. Однако ни в одном из этих исследований не применяли коррекцию уровня значимости на множественные сравнения. Это может существенно завышать уровень значимости обнаруженных различий. Необходимо как минимум указывать и интерпретировать результаты с учётом статистических поправок на множественные сравнения. Кроме того, число проведённых сравнений варьировало в разных работах, что затрудняет обобщение и сопоставление их результатов.

Соблюдение рекомендаций и чек-листов [21–25], разработанных для оценки диагностической точности инструментов на основе ИИ, может способствовать устранению указанных проблем и значительно повысить качество публикаций.

Обзор 216 исследований, отобранных при дополнительном поиске, показал, что большинство из них характеризуется низким качеством с точки зрения методологического соответствия современным стандартам в исследовании БЯМ и представлении результатов.

Ограничения систематического обзора

Поиск исследований осуществляли с помощью двух поисково-аналитических систем на двух языках (английском и русском), а также в списках литературы отобранных статей. Предпринятый подход не позволяет охватить все существующие публикации в данной области, а лишь формирует репрезентативную выборку, отражающую общую тенденцию. Исследования отличаются большим разнообразием решаемых задач при одновременно невысоком их качестве, что на данном этапе затрудняет формулирование обобщённых выводов о диагностической точности БЯМ. Необходимо дальнейшее накопление данных [26, 27] и их систематический анализ по каждому из рассмотренных направлений.

ЗАКЛЮЧЕНИЕ

БЯМ демонстрируют перспективные результаты в лучевой диагностике, успешно решая широкий спектр задач — от упрощения текстов протоколов до поддержки клинических решений. Однако текущие данные об их диагностической точности варьируют в зависимости от области применения и часто получены в условиях высокого риска систематической ошибки. В связи с этим на данном этапе преждевременно говорить о полноценном внедрении БЯМ в клиническую практику. Необходимы дальнейшее улучшение качества исследований и стандартизация методов оценки диагностической эффективности для формирования надёжной доказательной базы.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Приложение 1. Список включённых в обзор исследований, отобранных при дополнительном поиске, и их основные характеристики.
doi: 10.17816/DD678373-4340320

Вклад авторов. Ю.А. Васильев, А.В. Владзимирский, О.В. Омелянская — разработка концепции исследования; Р.В. Решетников, О.Г. Нанова, К.М. Арзамасов, М.Р. Коденко, Р.А. Ерижоков, А.П. Памова, С.Р. Сераджи, И.А. Блохин, А.П. Гончар, П.Б. Гележе, Д.А. Ахмедзянова, Ю.Ф. Шумская — сбор и анализ литературных данных, написание текста рукописи;

Р.В. Решетников, О.Г. Нанова — редактирование текста рукописи. Все авторы одобрили рукопись (версию для публикации), а также согласились нести ответственность за все аспекты работы, гарантируя надлежащее рассмотрение и решение вопросов, связанных с точностью и добросовестностью любой её части.

Этическая экспертиза. Неприменимо.

Источник финансирования. Данная статья подготовлена авторским коллективом в рамках научно-практического проекта в сфере медицины (№ ЕГИСУ: 125051305989-8) «Перспективный АРМ врача-рентгенолога на основе генеративного искусственного интеллекта».

Раскрытие интересов. Авторы заявляют об отсутствии отношений, деятельности и интересов за последние три года, связанных с третьими лицами (коммерческими и некоммерческими), интересы которых могут быть затронуты содержанием статьи.

Оригинальность. При создании настоящей работы авторы не использовали ранее опубликованные сведения (текст, иллюстрации, данные).

Доступ к данным. Все данные, полученные в настоящем исследовании, доступны в статье и в приложениях к ней.

Генеративный искусственный интеллект. При создании настоящей статьи технологии генеративного искусственного интеллекта не использовали.

Рассмотрение и рецензирование. Настоящая работа подана в журнал в инициативном порядке и рассмотрена по обычной процедуре. В рецензировании участвовали два члена редакционной коллегии и научный редактор издания.

ADDITIONAL INFORMATION

Supplement 1: List of the included studies from the additional search and their basic characteristics.
doi: 10.17816/DD678373-4340320

Author contributions: Yu.A. Vasilev, A.V. Vladzymyrskyy, O.V. Omelyanskaya: development of the research concept; R.V. Reshetnikov; O.G. Nanova, K.M. Arzamasov, M.R. Kodenko, R.A. Erizhokov, A.P. Pamova, S.R. Seradzhi, I.A. Blokhin, A.P. Gonchar, P.B. Gelezhe, D.A. Akhmedzyanova, Yu.F. Shumskaya: literature review, data analysis, writing the text of the manuscript; R.V. Reshetnikov; O.G. Nanova: text editing. All the authors approved the version of the manuscript to be published and agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics approval: Not applicable.

Funding sources: This article was prepared by a team of authors within the framework of a scientific and practical project in the field of medicine (No. EGISU: 125051305989-8) "A promising automated workplace of a radiologist based on generative artificial intelligence".

Disclosure of interests: The authors have no relationships, activities, or interests for the last three years related to for-profit or not-for-profit third parties whose interests may be affected by the content of the article.

Statement of originality: No previously published material (text, images, or data) was used in this work.

Data availability statement: All data generated during this study are available in the article and its supplementary material.

Generative AI: No generative artificial intelligence technologies were used to prepare this article.

Provenance and peer-review: This paper was submitted unsolicited and reviewed following the standard procedure. The peer review process involved two members of the editorial board and the in-house science editor.

СПИСОК ЛИТЕРАТУРЫ | REFERENCES

1. Cherif H, Moussa C, Missaoui AM, et al. Appraisal of ChatGPT's aptitude for medical education: comparative analysis with third-year medical students in a pulmonology examination. *JMIR Medical Education.* 2024;10:e52818. doi: 10.2196/52818 EDN: OFMTDE

2. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean Dental licensing examination: a comparative study. *International Dental Journal.* 2025;75(1):176–184. doi: 10.1016/j.identj.2024.09.002 EDN: JDFMDL

3. Busch F, Hoffmann L, dos Santos DP, et al. Large language models for structured reporting in radiology: past, present, and future. *European Radiology.* 2024;35(5):2589–2602. doi: 10.1007/s00330-024-11107-6 EDN: PNFKNR

4. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. *Diagnostic and Interventional Imaging.* 2023;104(6):269–274. doi: 10.1016/j.diii.2023.02.003EDN: FGMMTY

5. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine.* 2018;169(7):467–473. doi: 10.7326/M18-0850

6. Vasilev YuA, Vladzymyrskyy AV, Omelyanskaya OV, et al. *Methodological recommendations for preparing a systematic review.* Moscow: Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies; 2023. (In Russ.) EDN: XKXHDA

7. Kodenko MR, Vasilev YA, Vladzymyrskyy AV, et al. Diagnostic accuracy of ai for opportunistic screening of abdominal aortic aneurysm in CT: a systematic review and narrative synthesis. *Diagnostics.* 2022;12(12):3197. doi: 10.3390/diagnostics12123197 EDN: ERWYPX

8. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *European Radiology.* 2024;35(1):506–516. doi: 10.1007/s00330-024-10902-5 EDN: JAHWFM

9. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *European Radiology.* 2024;35(4):1938–1947. doi: 10.1007/s00330-024-11032-8 EDN: UHMLBQ

10. Kaya K, Gietzen C, Hahnfeldt R, et al. Generative Pre-trained Transformer 4 analysis of cardiovascular magnetic resonance reports in suspected myocarditis: A multicenter study. *Journal of Cardiovascular Magnetic Resonance.* 2024;26(2):101068. doi: 10.1016/j.jocmr.2024.101068 EDN: TSRLJX

11. Grolleau E, Couraud S, Jupin Delevaux E, et al. Incidental pulmonary nodules: Natural language processing analysis of radiology reports. *Respiratory Medicine and Research.* 2024;86:101136. doi: 10.1016/j.resmer.2024.101136 EDN: DHDPIX

12. Khoruzhaya AN, Kozlov DV, Arzamasov KM, Kremneva EI. Comparison of an ensemble of machine learning models and the BERT language model for analysis of text descriptions of brain CT reports to determine the presence of intracranial hemorrhage. *Sovremennye tehnologii v medicine.* 2024;16(1):27–36. doi: 10.17691/stm2024.16.1.03 EDN: AXXVVD

13. Han T, Adams LC, Bressem KK, et al. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA.* 2024;331(15):1320–1321. doi: 10.1001/jama.2023.27861 EDN: KPFLZG

14. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology.* 2023;66(1):73–79. doi: 10.1007/s00234-023-03252-4 EDN: SRFGAA

15. Wataya T, Miura A, Sakisuka T, et al. Comparison of natural language processing algorithms in assessing the importance of head computed tomography reports written in Japanese. *Japanese Journal of Radiology.* 2024;42(7):697–708. doi: 10.1007/s11604-024-01549-9 EDN: VAKPBV

16. Cagnina A, Salihu A, Meier D, et al. Assessing the need for coronary angiography in high-risk non-ST-elevation acute coronary syndrome

patients using artificial intelligence and computed tomography. *The International Journal of Cardiovascular Imaging.* 2024;41(1):55–61. doi: 10.1007/s10554-024-03283-9 EDN: JMBFSX

17. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine.* 2025;31(1):60–69. doi: 10.1038/s41591-024-03425-5 EDN: KAPIXF

18. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. *NEJM AI.* 2025;2(6). doi: 10.1056/AIp2401106

19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 1995;57(1):289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

20. Hollestein LM, Lo SN, Leonardi-Bee J, et al. MULTIPLE ways to correct for MULTIPLE comparisons in MULTIPLE types of studies. *British Journal of Dermatology.* 2021;185(6):1081–1083. doi: 10.1111/bjd.20600 EDN: QQWVVP

21. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi: 10.1136/bmj-2023-078378 EDN: WSTQKK

22. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799. doi: 10.1136/bmjopen-2016-012799

23. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ.* 2015;351:h5527. doi: 10.1136/bmj.h5527

24. Vasiliev YuA, Vlazimirsky AV, Omelyanskaya OV, et al. Methodology for testing and monitoring artificial intelligence-based software for medical diagnostics. *Digital Diagnostics.* 2023;4(3):252–267. doi: 10.17816/DD321971 EDN: UEDORU

25. Vasilev YuA, Bobrovskaya TM, Arzamasov KM, et al. Medical datasets for machine learning: fundamental principles of standartization and systematization. *Manager Zdravookhranenia.* 2023; (4):28–41. doi: 10.21045/1811-0185-2023-4-28-41 EDN: EPGAMD

26. Vinogradova IA, Nizovtsova LA, Omelyanskaya OV. Innovative strategic session in the scientific activity of the Center for Diagnostics and Telemedicine. *Digital Diagnostics.* 2022;3(4):414–420. doi: 10.17816/DD111833 EDN: DLRLVI

27. Kalinina ML, Svitachev AP, Biswas D, Vishnu P. Comparison of awareness and attitudes toward artificial intelligence among Russian- and English-speaking students at Orenburg State Medical University. *Digital Diagnostics.* 2023;4(1S):62–65. doi: 10.17816/DD430346 EDN: DIKOYA

ОБ АВТОРАХ

*** Нанова Ольга Геннадьевна,** канд. биол. наук;
адрес: Россия, 127051, г. Москва, ул. Петровка, д. 24, стр. 1;
ORCID: 0000-0001-8886-3684;
eLibrary SPIN: 6135-4872;
e-mail: nanova@mail.ru

Васильев Юрий Александрович, канд. мед. наук;
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: npcmr@zdrav.mos.ru

Решетников Роман Владимирович, канд. физ.-мат. наук;
ORCID: 0000-0002-9661-0254;
eLibrary SPIN: 8592-0558;
e-mail: ReshetnikovRV1@zdrav.mos.ru

Владзимирский Антон Вячеславович, д-р мед. наук;
ORCID: 0000-0002-2990-7736;
eLibrary SPIN: 3602-7120;
e-mail: VladzimirskijAV@zdrav.mos.ru

Арзамасов Кирилл Михайлович, д-р мед. наук;
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

Омелянская Ольга Васильевна;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: o.omelyanskaya@npcmr.ru

Коденко Мария Романовна, канд. техн. наук;
ORCID: 0000-0002-0166-3768;
eLibrary SPIN: 5789-0319;
e-mail: KodenkoMR@zdrav.mos.ru

Ерижоков Рустам Арсеньевич;
ORCID: 0009-0007-3636-2889;
eLibrary SPIN: 2274-6428;
e-mail: ErizhokovRA@zdrav.mos.ru

AUTHORS' INFO

*** Olga G. Nanova,** Cand. Sci. (Biology);
address: 24 Petrovka st, bldg 1, Moscow, Russia, 127051;
ORCID: 0000-0001-8886-3684;
eLibrary SPIN: 6135-4872;
e-mail: nanova@mail.ru

Yuriy A. Vasilev, MD, Cand. Sci. (Medicine);
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: npcmr@zdrav.mos.ru

Roman V. Reshetnikov, Cand. Sci. (Physics and Mathematics);
ORCID: 0000-0002-9661-0254;
eLibrary SPIN: 8592-0558;
e-mail: ReshetnikovRV1@zdrav.mos.ru

Anton V. Vladzymyrskyy, MD, Dr. Sci. (Medicine);
ORCID: 0000-0002-2990-7736;
eLibrary SPIN: 3602-7120;
e-mail: VladzimirskijAV@zdrav.mos.ru

Kirill M. Arzamasov, MD, Dr. Sci. (Medicine);
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

Olga V. Omelyanskaya;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: o.omelyanskaya@npcmr.ru

Maria R. Kodenko, Cand. Sci. (Engineering);
ORCID: 0000-0002-0166-3768;
eLibrary SPIN: 5789-0319;
e-mail: KodenkoMR@zdrav.mos.ru

Rustam A. Erizhokov, MD;
ORCID: 0009-0007-3636-2889;
eLibrary SPIN: 2274-6428;
e-mail: ErizhokovRA@zdrav.mos.ru

Памова Анастасия Петровна, канд. мед. наук;
ORCID: 0000-0002-0041-3281;
eLibrary SPIN: 5146-4355;
e-mail: PamovaAP@zdrav.mos.ru

Сераджи Сеал Рахмануддин;
ORCID: 0009-0000-3990-6668;
e-mail: SeradzhiSR@zdrav.mos.ru

Блохин Иван Андреевич, канд. мед. наук;
ORCID: 0000-0002-2681-9378;
eLibrary SPIN: 3306-1387;
e-mail: BlokhinIA@zdrav.mos.ru

Гончар Анна Павловна, канд. мед. наук;
ORCID: 0000-0001-5161-6540;
eLibrary SPIN: 3513-9531;
e-mail: GoncharAP@zdrav.mos.ru

Гележе Павел Борисович, канд. мед. наук;
ORCID: 0000-0003-1072-2202;
eLibrary SPIN: 4841-3234;
e-mail: GelezhePB@zdrav.mos.ru

Ахмедзянова Дина Альфредовна;
ORCID: 0000-0001-7705-9754;
eLibrary SPIN: 6983-5991;
e-mail: AkhmedzyanovaDA@zdrav.mos.ru

Шумская Юлия Федоровна;
ORCID: 0000-0002-8521-4045;
eLibrary SPIN: 3164-5518;
e-mail: shumskayayf@zdrav.mos.ru

Anastasia P. Pamova, MD, Cand. Sci. (Medicine);
ORCID: 0000-0002-0041-3281;
eLibrary SPIN: 5146-4355;
e-mail: PamovaAP@zdrav.mos.ru

Seal R. Seradzhi;
ORCID: 0009-0000-3990-6668;
e-mail: SeradzhiSR@zdrav.mos.ru

Ivan A. Blokhin, MD, Cand. Sci. (Medicine);
ORCID: 0000-0002-2681-9378;
eLibrary SPIN: 3306-1387;
e-mail: BlokhinIA@zdrav.mos.ru

Anna P. Gonchar, MD, Cand. Sci. (Medicine);
ORCID: 0000-0001-5161-6540;
eLibrary SPIN: 3513-9531;
e-mail: GoncharAP@zdrav.mos.ru

Pavel B. Gelezhe, MD, Cand. Sci. (Medicine);
ORCID: 0000-0003-1072-2202;
eLibrary SPIN: 4841-3234;
e-mail: GelezhePB@zdrav.mos.ru

Dina A. Akhmedzyanova, MD;
ORCID: 0000-0001-7705-9754;
eLibrary SPIN: 6983-5991;
e-mail: AkhmedzyanovaDA@zdrav.mos.ru

Yuliya F. Shumskaya, MD;
ORCID: 0000-0002-8521-4045;
eLibrary SPIN: 3164-5518;
e-mail: shumskayayf@zdrav.mos.ru

* Автор, ответственный за переписку / Corresponding author