DOI: https://doi.org/10.17816/DD625967

Оценка производительности программного обеспечения на основе технологии искусственного интеллекта при описании цифровых маммографических исследований

Ю.А. Васильев[1,2], А.В. Колсанов[3], К.М. Арзамасов[1], А.В. Владзимирский[1,4], О.В. Омелянская[1], С.С. Семёнов[1], Л.Е. Аксёнова[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий, Москва, Россия;
[2] Национальный медико-хирургический Центр имени Н.И. Пирогова, Москва, Россия;
[3] Самарский государственный медицинский университет, Самара, Россия;
[4] Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия

АННОТАЦИЯ

Обоснование. Цифровая скрининговая маммография — это основной инструмент для раннего выявления злокачественных новообразований молочной железы, позволяющий снизить смертность на 20–40%. На сегодняшний день разработано множество сервисов на основе искусственного интеллекта (ИИ), позволяющих автоматизировать анализ таких исследований.

Цель — сравнить результаты оценки цифровых маммографических исследований, выполненной тремя типами ИИ-сервисов в нескольких версиях, с заключениями врачей-рентгенологов.

Материалы и методы. Проведено сравнение бинарных шкал оценки маммографических исследований и нескольких типов и версий ИИ-сервисов по показателям диагностической точности, коэффициенту Мэтьюса и максимальному индексу Юдена.

Результаты. Сравнительный анализ показал, что выбор бинарной шкалы для оценки цифрового маммографического исследования влияет на количество выявляемых случаев патологии и точность результатов ИИ-сервисов. Кроме того, обнаружена зависимость показателей диагностической точности от порогового значения. Наилучшей производительностью обладает ИИ-сервис 1 в версии 3, что подтверждается большинством показателей диагностической точности.

Заключение. Полученные нами результаты могут быть полезны при выборе ИИ-сервисов для интерпретации данных скрининговой маммографии. Настройка ИИ-сервиса методом максимизации индекса Юдена позволяет получать сбалансированные значения чувствительности и специфичности, что не всегда целесообразно с клинической точки зрения.

Ключевые слова: злокачественные новообразования молочной железы; цифровая маммография; сервисы искусственного интеллекта; показатели диагностической точности; индекс Юдена.

Как цитировать:
Васильев Ю.А., Колсанов А.В., Арзамасов К.М., Владзимирский А.В., Омелянская О.В., Семёнов С.С., Аксёнова Л.Е. Оценка производительности программного обеспечения на основе технологии искусственного интеллекта при описании цифровых маммографических исследований // Digital Diagnostics. 2024. Т. 5, № 4. С. 695–711. DOI: https://doi.org/10.17816/DD625967

Рукопись получена: 24.01.2024 **Рукопись одобрена:** 10.10.2024 **Опубликована online:** 02.12.2024

DOI: https://doi.org/10.17816/DD625967

Evaluating the performance of artificial intelligence-based software for digital mammography characterization

Yuri A. Vasilev[1,2], Alexander V. Kolsanov[3], Kirill M. Arzamasov[1], Anton V. Vladzymyrskyy[1,4], Olga V. Omelyanskaya[1], Serafim S. Semenov[1], Lubov E. Axenova[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] National Medical and Surgical Center named after N.I. Pirogov, Moscow, Russia;
[3] Samara State Medical University, Samara, Russia;
[4] Sechenov First Moscow State Medical University, Moscow, Russia

ABSTRACT

BACKGROUND: Digital screening mammography is a key modality for early detection of breast cancer, reducing mortality by 20–40%. Many artificial intelligence (AI)-based services have been developed to automate the analysis of imaging data.

AIM: The aim of the study was to compare mammography assessments using three types of AI services in multiple versions with radiologists' conclusions.

MATERIALS AND METHODS: Binary mammography scoring scales were compared with several types and versions of AI services regarding diagnostic accuracy, Matthews correlation coefficient, and maximum Youden's index.

RESULTS: A comparative analysis showed that the use of a binary scale for evaluating digital mammography affects the number of detected abnormalities and accuracy of AI results. In addition, diagnostic accuracy was found to be threshold dependent. AI Service 1 in version 3 had the best performance, as confirmed by most diagnostic accuracy parameters.

CONCLUSION: Our results can be used to select AI services for interpreting mammography screening data. Using Youden's index maximization to set up an AI service provides a balance of sensitivity and specificity that is not always clinically relevant.

Keywords: malignant tumors of breast; digital mammography; artificial intelligence services; diagnostic accuracy; Youden's index.

To cite this article:
Vasilev YA, Kolsanov AV, Arzamasov KM, Vladzymyrskyy AV, Omelyanskaya OV, Semenov SS, Axenova LE. Evaluating the performance of artificial intelligence-based software for digital mammography characterization. *Digital Diagnostics.* 2024;5(4):695–711. DOI: https://doi.org/10.17816/DD625967

Received: 24.01.2024 **Accepted:** 10.10.2024 **Published online:** 02.12.2024

DOI: https://doi.org/10.17816/DD625967

基于人工智能技术的软件在描述数字乳房造影检查中的性能评估

Yuri A. Vasilev[1,2], Alexander V. Kolsanov[3], Kirill M. Arzamasov[1], Anton V. Vladzymyrskyy[1,4], Olga V. Omelyanskaya[1], Serafim S. Semenov[1], Lubov E. Axenova[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] National Medical and Surgical Center named after N.I. Pirogov, Moscow, Russia;
[3] Samara State Medical University, Samara, Russia;
[4] Sechenov First Moscow State Medical University, Moscow, Russia

摘要

论证。数字乳房造影筛查是早期发现乳腺恶性肿瘤的主要工具，可将死亡率降低20~40%。目前，已开发出许多基于人工智能（AI）的服务来自动分析此类检查。

目的 — 比较三种人工智能服务在不同版本中进行的乳房造影检查评估结果与放射科医生的意见。

材料和方法。比较了乳房造影检查二元评估量表与多种类型和版本的AI服务在诊断准确性指标、马修斯系数和最大尤登指数等方面的差异。

结果。比较分析表明，评估数字乳房造影检查的二元评估量表的选择会影响检测到的病理病例数量和AI服务结果的准确性。此外，还发现了诊断准确性指标对阈值的依赖性。版本3中的AI服务1实现了最佳性能，大多数诊断准确性指标都证实了这一点。

结论。我们的研究结果可能有助于选择AI服务来解读乳房造影筛查数据。通过最大化尤登指数来设置AI服务，可以获得灵敏度和特异性的平衡值，但从临床角度来说，并不总是合理的。

关键词：乳腺恶性肿瘤；数字乳房造影；人工智能服务；诊断准确性指标；尤登指数。

引用本文：
Vasilev YA, Kolsanov AV, Arzamasov KM, Vladzymyrskyy AV, Omelyanskaya OV, Semenov SS, Axenova LE. 基于人工智能技术的软件在描述数字乳房造影检查中的性能评估. *Digital Diagnostics.* 2024;5(4):695–711. DOI: https://doi.org/10.17816/DD625967

收到： 24.01.2024 **接受：** 10.10.2024 **发布日期：** 02.12.2024

ОБОСНОВАНИЕ

В рентгенологии цифровая маммография является основным инструментом для диагностики и единственным инструментом для скрининга злокачественных новообразований (ЗНО) молочной железы. Проведение скрининга позволяет значительно раньше обнаружить патологические изменения молочной железы, связанные со злокачественными новообразованиями, что снижает уровень смертности от ЗНО на 20–40% [1]. С развитием искусственного интеллекта (ИИ) появляется всё больше систем или сервисов на его основе, которые автоматизируют анализ изображений цифровой маммографии [2–4]. Часть исследований показывает, что сервисы искусственного интеллекта (ИИС) достигают высокой точности диагностики, которая в некоторых случаях превосходит результаты врачей-рентгенологов. Чаще всего это касается обнаружения признаков ЗНО на ранних стадиях развития опухоли и/или в случае преобладания фиброгландулярной ткани молочной железы в проекции патологических изменений. Однако есть исследования, указывающие на то, что при интерпретации маммографических изображений точность врачей-рентгенологов всё ещё превышает точность ИИС [5]. Модели машинного обучения — это основные функциональные компоненты ИИС, ответственные за детекцию и сегментацию областей интереса с патологическими изменениями, обработку и классификацию данных и вывод предсказаний или решений на основе этих данных. Сравнение моделей машинного обучения включает в себя расчёт показателей диагностической точности, таких как чувствительность (Sens — Sensetivity) и специфичность (Spec — Specificity), а также анализ площади под характеристической кривой (AUC — Area Under Curve) [6, 7].

Для оценки производительности ИИ необходимо выбрать истинное значение, с которым будут сравниваться результаты ИИС. В основном расчёты проводят относительно выходных данных модели и «золотого» стандарта, который формируется по результатам дополнительных исследований [8, 9]. Кроме того, возможна оценка результатов ИИ путём их сравнения с заключением врача [10, 11]. Основным преимуществом ИИ-систем является возможность их тонкой настройки, однако важным аспектом внедрения и использования ИИС в медицине является проверка точности программного обеспечения (ПО), которое выдаёт вероятность вместо традиционного бинарного результата.

Интерпретация вероятностных результатов требует определения порога отсечения, от которого зависит какая вероятность считается «патологией», а какая — «нормой». Определение оптимального порогового значения вероятности зависит от конкретных целей и контекста применения ИИС. Поскольку распределение вероятностей для несбалансированных данных имеет тенденцию к смещению в сторону класса «норма» [12], выбор значения 0,5 в качестве порога может оказаться неоптимальным. Для максимального выявления случаев ЗНО и сведения к минимуму количества ложно положительных результатов необходима балансировка между чувствительностью и специфичностью модели машинного обучения. Одним из подходов, которым часто пользуются для максимизации значений Sens и Spec, является максимизация их суммы с помощью индекса Юдена [7]. Кроме того, F. Chen и соавт. [13] предложили метод сравнения максимального значения индекса Юдена для нескольких диагностических тестов. Учитывая тот факт, что применение систем искусственного интеллекта в медицинской диагностике может иметь высокие риски при недостаточной их производительности, необходима методология полной оценки потенциала и ограничений в работе таких ИИ-систем.

ЦЕЛЬ

Сравнить результаты оценки цифровых маммографических исследований, выполненной ИИС в нескольких версиях, с заключениями врачей-рентгенологов.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования

Проведено обсервационное многоцентровое одномоментное выборочное исследование. Дизайн исследования, а также схема формирования наборов данных для проведения анализа представлены на рис. 1.

Критерии соответствия

Критерии включения. В выборку включали пациенток (без учёта их возраста или наличия сопутствующих заболеваний), проходивших цифровую маммографию в период с 22 июля 2020 г. по 29 декабря 2022 г., при наличии в составе медицинских данных изображения в формате DICOM и соответствующей информации для анализа ИИС.

Критерии невключения:

- отсутствие в составе медицинских данных результатов для обработки хотя бы одним из анализируемых ИИС;
- наличие технических дефектов изображений, мешающих корректной интерпретации (например, артефакты, частичное отсутствие данных);
- неполная информация о метаданных, необходимая для анализа.

Дополнительно. Исследования с участием пациенток с имплантами и пациенток после лучевой терапии не выделяли в отдельные подгруппы, и их количество в выборке не отслеживали.

Условия проведения

Данные в итоговой выборке включали в себя результаты обследований, проведённых в 123 амбулаторных медицинских организациях Департамента здравоохранения

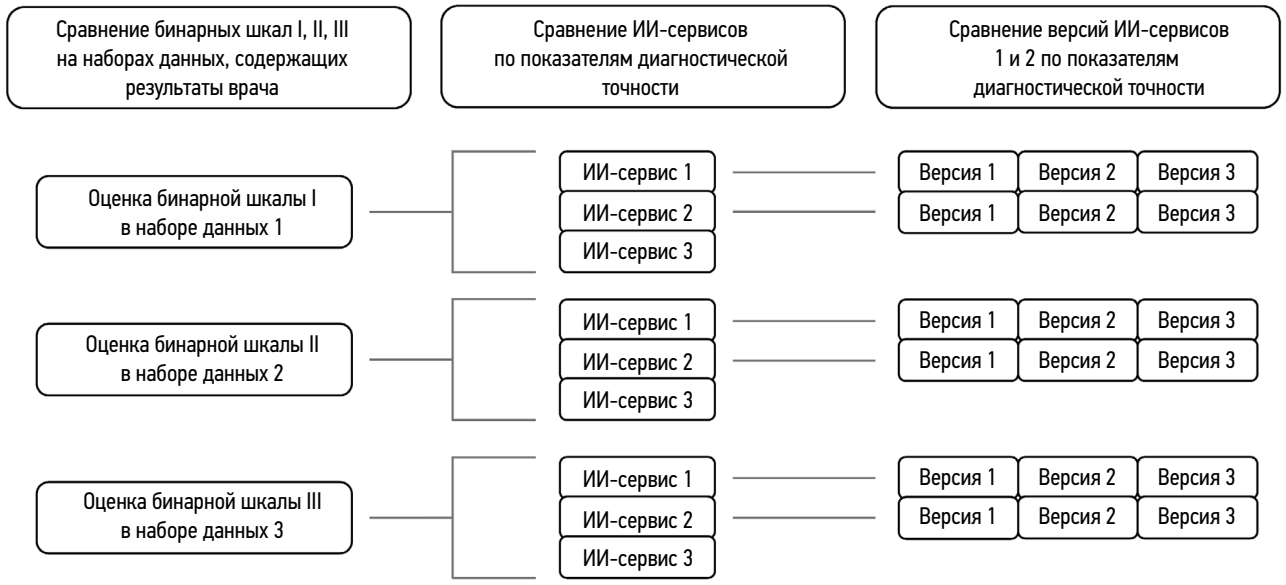

Рис. 1. Дизайн исследования и формирование наборов данных для анализа: ИИ-сервис — сервис искусственного интеллекта.

города Москвы. В исследовании приняли участие 531 врач-рентгенолог по субспециализации маммография, все врачи описывали исследования в медицинских организациях Департамента здравоохранения города Москвы. В качестве истинного значения для сравнения с результатами ИИС использовали заключение врача-рентгенолога по каждому исследованию. За указанный период времени каждый врач описал в среднем по 1250 исследований.

Формирование и анализ данных

В качестве эталонных значений для оценки точности результатов ИИС использовали заключения врачей-рентгенологов, взятые из медицинской документации. Заключения были представлены в виде категорий, соответствующих стандартной системе классификации и интерпретации результатов маммографических исследований BI-RADS 1–6 (Breast Imaging Reporting and Data System), отдельно для каждой молочной железы. Разделение на бинарные диагностические шкалы согласно вероятности наличия ЗНО по системе BI-RADS проводили тремя способами: шкала I — отнесение категории BI-RADS 1–2 к «норме», категории 3–6 — к «патологии»; шкала II — отнесение категории BI-RADS 1–3 к «норме», а категории 4–6 — к «патологии»; шкала III — отнесение категории BI-RADS 1–2 к «норме», а 4–6 — к «патологии» (категория BI-RADS 3 в данной шкале не учитывается).

В исследование было включено три ИИС: ТриоДМ-МТ® (AUC 0,90; специфичность 0,85; чувствительность 0,83; точность 0,84) (АО «Медицинские Технологии лтд», Россия), Цельс® (AUC 0,96) (ООО «Медицинские скрининг системы», Россия) и Lunit INSIGHT MMG® (AUC 0,96; чувствительность 0,89 при оценке исследований совместно с радиологом) (Lunit Inc., Южная Корея) [2–4]. Для каждого маммографического исследования результаты работы ИИС представлены в виде значений вероятности в диапазоне от 0% (низкое подозрение на злокачественность) до 100% (высокое подозрение на злокачественность). Далее по тексту торговые наименования ИИС представлены анонимно и рандомизированы.

Предварительная обработка данных включала удаление строк, в которых отсутствовал результат описания исследования врачом и/или отсутствовал результат работы ИИС. Кроме того, из набора данных исключили исследования, выполненные пациентам мужского пола; исследования, где возраст обследуемой составлял менее 40 или более 100 лет; исследования, где заключение врача не соответствовало системе BI-RADS 1–6 или ни одному из вышеперечисленных ИИС.

После предварительной обработки данных каждой маммографии рассчитывали показатели диагностической точности, включая AUC, чувствительность (Sens), специфичность (Spec), точность (Acc), положительную прогностическую ценность (PPV), коэффициент ложных отрицательных (FNR), коэффициент выявления случаев (CDR), долю ложноположительных заключений (AIR), коэффициент корреляции Мэттьюса (MCC) и индекс Юдена (J). В табл. 1 приведены описания каждой метрики и указаны диагностические шкалы, продемонстрировавшие максимальные значения этих метрик.

Во время исследования ИИС дорабатывали: осуществляли дообучение, тонкую настройку, вносили другие изменения. Каждое изменение версии ИИС соответствовало его доработке. В исследовании учитывали только затрагивающие ядро ИИС изменения, которые влияли на показатели диагностической точности. Таким образом, для ИИС-1 и ИИС-2 было выделено три версии, отражающие последовательные изменения лежащей в основе ПО модели и работающие в разное время. В ИИС-3 существенных изменений не вносили, поэтому отдельных версий не выделяли.

Таблица 1. Описание метрик диагностической точности и диагностические шкалы, продемонстрировавшие самые высокие значения этих метрик

Метрика	Описание метрики	Диагностическая шкала
AUC	Area Under the Curve — площадь под характеристической кривой; отражает способность различать класс, не чувствительна к дисбалансу классов	II и III
Sens	Sensitivity — чувствительность; отражает способность детектировать класс «патология»	III
Spec	Specificity — специфичность; отражает способность детектировать класс «норма»	II
Acc	Accuracy — точность; отражает долю правильно классифицированных объектов от общего числа объектов в выборке, чувствительна к дисбалансу классов	II
PPV	Positive Predictive Value — положительная прогностическая ценность; отражает соответствие детектированного класса «патология» действительно патологическому случаю	I
AIR	Abnormal Interpretation Rate — доля исследований, которые получили заключение «патология» и нуждаются в дополнительных диагностических процедурах; отражает наибольшее количество ложно положительных результатов	I
CDR	Case Detection Rate — коэффициент выявления случаев; отражает выявление случаев патологии независимо от общего числа ложно положительных результатов	I
FNR	False Negative Rate — коэффициент ложных отрицательных; оценивает количество случаев патологии, которые не были детектированы сервисом искусственного интеллекта	I
MCC	Matthews Correlation Coefficien — коэффициент корреляции Мэттьюса; оценивает качество классификации с учётом всех четырёх элементов матрицы ошибок, метрика не чувствительна к дисбалансу классов	I
J	Youden's Index — индекс Юдена	-

Для определения оптимального порога отсечения значения вероятности вычисляли AUC и максимальное значения индекса Юдена. Вычисления проводили с использованием WEB-инструмента, разработанного Московским центром диагностики и телемедицины[1]. Формула для расчёта значения индекса Юдена имеет вид:

$$J = Sens - Spec - 1 \quad (1)$$

где Sens — чувствительность; Spec — специфичность.

С использованием порога отсечения, были рассчитаны бинарные результаты для ИИС. Далее, для сравнения результатов ИИС с заключением врача вычисляли:

- TP — True Positive, количество истинно положительных случаев;
- TN — True Negative, количество истинно отрицательных случаев;
- FP — False Positive, количество ложно положительных случаев;
- FN — False Negative, количество ложно отрицательных случаев.

С использованием полученных значений TP, TN, FP и FN вычисляли следующие метрики точности ИИС (табл. 1) [14]:

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i) * (y_i + y_{i+1}) \quad (2)$$

где x — значения по оси X (например, ложные положительные), y — значения по оси Y (например, истинные положительные), n — общее количество точек на кривой, i — индекс текущей точки.

$$Sens = \frac{TP}{TP + FN} \quad (3)$$

$$Sens = \frac{TP}{TP + FN} \quad (4)$$

$$Acc = \frac{TP + TN}{FP + FN + TP + TN} \quad (5)$$

[1] С.П. Морозов, А.Е. Андрейченко, С.Ф. Четвериков, и др. Свидетельство о государственной регистрации программы для ЭВМ № 2022617324 Российская Федерация. Веб-инструмент для выполнения ROC анализа результатов диагностических тестов: № 2022616046: заявл. 05.04.2022: опубл. 19.04.2022. Режим доступа: https://roc-analysis.mosmed.ai/ Дата обращения: 20.08.2023 EDN: ECMPNH

$$PPV = \frac{TP}{TP + FP} \quad (6)$$

$$AIR = \frac{TP + FP}{TP + TN + FP + FN} \times 1000 \quad (7)$$

$$CDR = \frac{TP}{TP + TN + FP + FN} \times 1000 \quad (8)$$

$$FNR = \frac{FN}{FN + TP} \quad (9)$$

$$MCC = \frac{TP \times TN + FP + FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (10)$$

Для расчёта показателей точности, а также доверительных интервалов мы применяли метод бутстрэппинга (Bootstrapping), который заключается в формировании 100 выборок по 1000 образцов с соотношением категорий 0 («норма») и 1 («патология») равным 9:1 (для диагностической шкалы I), 33:1 (для шкалы II) и 31:1 (для шкалы III), что позволило сымитировать соотношение, рассчитанное в наборах данных 1–3.

Этическая экспертиза

Настоящая работа проведена в рамках ранее одобренного локальным этическим комитетом исследования «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» (Московский эксперимент); (протокол № NCT04489992 от 21 февраля 2020 года).

Статистический анализ

В настоящем исследовании сравнивали точность оценки наличия или отсутствия ЗНО молочной железы для трёх бинарных шкал, составленных на основе заключений врачей-рентгенологов и для трёх ИИС. Для проверки нормальности распределения сформированных наборов данных использовали непараметрический тест Колмогорова–Смирнова.

Чтобы оценить статистическую значимость различий между максимальными значениями индекса Юдена для различных типов и версий ИИС использовали метод, описанный F. Chen и соавт. [13]. Дисперсию (Variance — Var) разности двух независимых индексов Юдена измеряли по формуле:

$$Var(J_1 - J_2) = Var(J_1) + Var(J_2) \quad (11)$$

где J — значение индекса Юдена, а значение Var рассчитывается по формуле:

$$Var(J) = Spec^2 \times Var(Sens) + Sens^2 \times Var(Spec) \quad (12)$$

где Spec — специфичность; Senc — чувствительность.

Таким образом, уравнение имеет вид:

$$Var(J_1 - J_2) = Spec_1^2 \times Var(Sens_1) + Sens_1^2 \times \times Var(Spec_1) + Spec_2^2 \times Var(Sens_2) + Sens_2^2 \times \times Var(Spec_2) \quad (13)$$

Статистический тест и двусторонний доверительный интервал для оценки разности двух независимых индексов Юдена (*J*) рассчитывали на основе центральной предельной теоремы:

$$Z = \frac{J_1 - J_2}{\sigma_{D_J}} = \frac{J_1 - J_2}{\sqrt{Var(J_1 - J_2)}} \quad (14)$$

$$d \pm Z_{\alpha/2} \times \sigma_{D_J} = d \pm Z_{\alpha/2} \times \sqrt{Var(J_1 - J_2)} \quad (15)$$

где Z — стандартная нормальная случайная величина, характеризующая отклонение разности от нуля в стандартных отклонениях; Var — дисперсия; d — разность между двумя индексами Юдена; σ_{D_J} — стандартное отклонение разности индексов Юдена.

Статистически значимым выбрано значение *p* меньше 0,05. Доверительный интервал соответствует 95%. Для расчётов использовали библиотеки Pandas, Matplotlib и Seaborn, Scikit-learn, NumPy, Statistics (stats) языка программирования Python (Python Software Foundation, версия 3.11.0).

РЕЗУЛЬТАТЫ

Сравнение бинарных диагностических шкал, составленных на основе заключений врача

Оценка нормальности распределения категорий 1–6 по шкале BI-RADS, выставленных врачом, показала, что распределение данных показателей не соответствует нормальному. На рис. 2 представлены гистограммы распределения категорий BI-RADS. Пики на графике

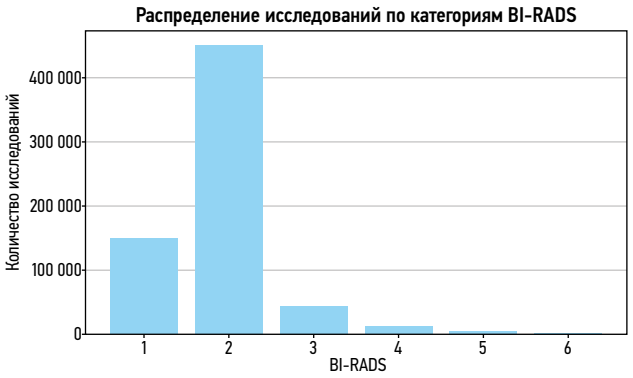

Рис. 2. Распределение категорий 1–6 по шкале BI-RADS, выставленных врачом в ходе описания цифровой маммографии для исследуемых наборов данных: по оси X — категория по шкале BI-RADS 1–6; по оси Y — количество исследований.

соответствуют наиболее вероятным категориям. В данном случае самый высокий пик соответствует категории BI-RADS 2 — «доброкачественные изменения молочной железы», что указывает на отсутствие признаков патологических изменений, ассоциированных с ЗНО, в большинстве исследований, входящих в выборку.

В набор данных 1 и 2 вошло 663 606 исследований, в набор данных 3 — 618 947 исследований, количество случаев патологии составило 64 100, 19 441 и 19 441, а количество случаев нормы —599 506, 644 165 и 599 506 соответственно. Таким образом, частота встречаемости ЗНО в исследуемой выборке данных составляет 9,66% для бинарной шкалы I и 2,9% для бинарной шкалы II и III (рис. 3). Подробная информация о наборах данных представлена в табл. 2 и 3.

Для оценки соответствия результатов врача и ИИС в трёх сформированных наборах данных были рассчитаны показатели диагностической точности (табл. 1 и 4). Можно отметить, что площадь под характеристической кривой AUC для шкалы I значительно отличалась от шкалы II и III, у которых AUC не различается между собой. Кроме того, чувствительность Sens была больше для шкалы III, в то время как специфичность Spec — для шкалы II. Наибольшее количество ложно положительных результатов AIR и самый высокий процент случаев заболевания, оставшихся не детектированными FNR, показала шкала I; самыми низкими перечисленные показатели были

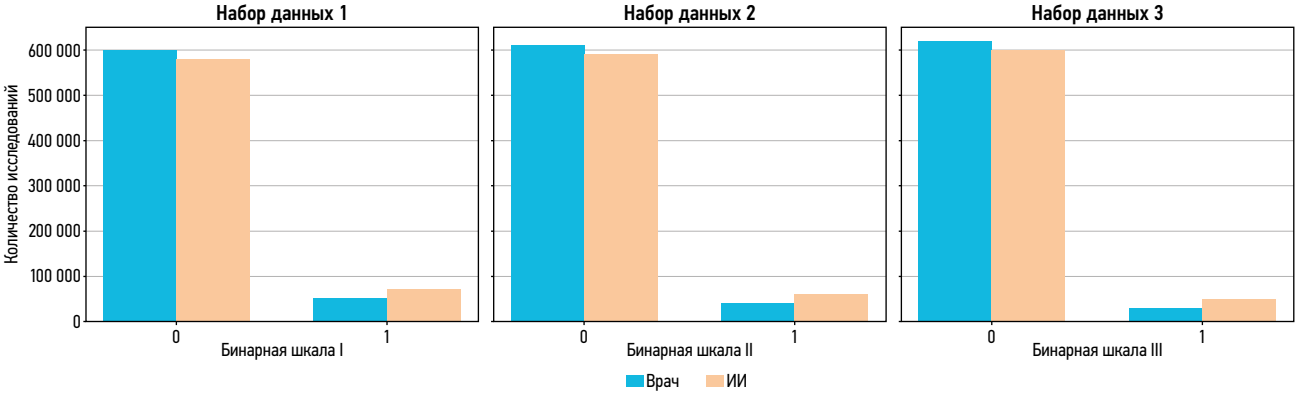

Рис. 3. Сравнение распределения категорий 0–1, выставленных врачами и сервисом искусственного интеллекта для трёх бинарных шкал: по оси X — бинарная шкала I–III; по оси Y — количество исследований; ИИ — искусственный интеллект.

Таблица 2. Количество случаев нормы и патологии в наборах данных 1-3

	Норма	Патология	Все исследования	Количество здоровых на 1 больного
Шкала I	599 506	64 100	663 606	9
Шкала II	644 165	19 441	663 606	33
Шкала III	599 506	19 441	618 947	31

Таблица 3. Количество исследований в наборах данных (период 2020–2022 гг.)

Шкалы	I–II							III					
Количество исследований	663 606							618 947					
Сервисы	1			2			3	1			2		
Количество исследований	545 362			108 763			9481	508 929			101 654		
Версии	1	2	3	1	2	3	-	1	2	3	1	2	3
Количество исследований	90 949	212 968	241 445	4922	46 851	56 990	-	83 828	198 231	226 870	4711	43 687	53 256

Таблица 4. Сервисы искусственного интеллекта и их версии, имевшие наибольшие значения метрик точности по сравнению с диагностическими шкалами

Метрика	Шкала сравнения	ИИС	Номер версии ИИС-1	Номер версии ИИС-2
AUC	I	1	3	1 и 3
	II	1	3	1 и 2
	III	1	3	1 и 2
Sens	I	1 и 2	2 и 3	2 и 3
	II	1 и 2	3	2
	III	1 и 2	3	2
Spec	I	1	3	1
	II	1	1 и 3	1
	III	1	1	1
Acc	I	1	3	1
	II	1	3	1
	III	1	1	1
PPV	I	1	3	1
	II	1	3	1
	III	1	2 и 3	1
AIR	I	3	1	2
	II	3	2	2
	III	3	3	2
CDR	I	1 и 2	2 и 3	2 и 3
	II	1 и 3	3	2
	III	1 и 3	3	2
FNR	I	3	1	1
	II	2	1	1 и 3
	III	2	1	1 и 3
MCC	I	1	3	1
	II	1	3	1
	III	1	3	1
Индекс Юдена	I	1	3	1
	II	1	3	2
	III	1	3	2

Примечание. ИИС — сервис искусственного интеллекта; AUC — площадь под характеристической кривой; Sens — чувствительность; Spec — специфичность; Acc — точность; PPV — положительная прогностическая ценность; AIR — доля исследований, получивших заключение «патология»; CDR — коэффициент выявления случаев; FNR — коэффициент ложных отрицательных; MCC — коэффициент корреляции Мэттьюса.

у шкалы II. Уровень согласованности, измеренный с помощью метрики MCC, также, как и метрики PPV и CDR, показали самые высокие значения в шкале I (табл. 5).

Сравнение ИИС между собой и со шкалами на основе заключений врача

Распределение значений вероятностей наличия патологии в исследовании для ИИС 1–3 представлены на рис. 4. Можно отметить, что распределение вероятностей ИИС наиболее схоже для шкалы II и III. При этом для категории «норма» распределение смещено вправо, особенно у ИИС-2 и -3, а для категории «патология» — для ИИС-1 распределение смещено влево, а для ИИС-2 и -3 — вправо.

Для оценки и сравнения между собой производительности ИИС-1, -2 и -3 использовали те же показатели диагностической точности. По показателям, которые указывают на соответствие результатов работы сервиса мнению врача в определении категорий «норма» и «патология», лидирует ИИС-1, в то время как количество ложноположительных результатов AIR и процент не детектированных реальных случаев заболевания FNR оказались самыми высокими для ИИС-2 и ИИС-3 (табл. 6).

Сравнение версий ИИС между собой

Показатели диагностической точности измерены и для оценки различных версий ИИС-1 и -2 (табл. 7 и 8). Большинство показателей диагностической точности отличались в зависимости от используемой шкалы, однако часть отличий оказалась статистически незначимыми. Таким образом, затруднительно сделать вывод о том какая версия ИИС является наиболее производительной.

Результаты оценки шкал, сервисов и версий, которые имели наибольшие показатели диагностической точности, приведены в табл. 4.

Для сравнения производительности различных типов и версий ИИС, кроме показателей диагностической точности, мы использовали максимальное значение индекса Юдена, которое позволяет оценить баланс между чувствительностью и специфичностью ИИС. Полученные результаты позволяют заключить, что по максимальному значению индекса Юдена лидирует ИИС-1 (табл. 4). При сравнении версий ИИС-1, наилучшие показатели диагностической точности получены для его 3-й версии. Однако, ИИС-2 по индексу Юдена показал наивысший результат для версии 1 по шкале I и для версии 2 — по шкалам II и III. Для всех результатов показаны статистически значимые различия.

ОБСУЖДЕНИЕ

В настоящем исследовании провели сравнение бинарных шкал оценки цифровой маммографии, трёх ИИС и трёх версий ИИС-1 и -2. Для этого рассчитывали классические показатели диагностической точности и индекс Юдена.

Отношение количества обнаруженных случаев ЗНО к общему числу выполненных исследований составляет

Таблица 5. Значения метрик диагностической точности, измеренные для результатов сервиса искусственного интеллекта (ИИС-1)

Метрика	Диагностическая бинарная шкала		
	I	II	III
Порог отсечения	62	74	68
AUC	0,659 [0,654; 0,663]	0,726 [0,717; 0,735] *	0,738 [0,730; 0,745] *
Sens	0,569 [0,560; 0,578]	0,626 [0,609; 0,644]	0,679 [0,664; 0,694] *
Spec	0,748 [0,746; 0,751]	0,826 [0,823; 0,828] *	0,796 [0,793; 0,798]
Acc	0,730 [0,728; 0,733]	0,820 [0,817; 0,822] *	0,792 [0,790; 0,795]
PPV	0,201 [0,198; 0,204] *	0,100 [0,098; 0,103]	0,099 [0,097; 0,101]
AIR	283,280 [280,570; 285,988] *	187,810 [185,199; 190,421]	219,240 [216,787; 221,693]
CDR	56,870 [55,960; 57,780] *	18,790 [18,273; 19,307]	21,730 [21,252; 22,208]
FNR	0,431 [0,422; 0,440] *	0,374 [0,356; 0,391]	0,321 [0,306; 0,336]
MCC	0,211 [0,205; 0,217] *	0,198 [0,190; 0,205]	0,202 [0,196; 0,209]

Примечание. Данные представлены в виде среднего значения [95%ДИ]; * — статистически значимые различия между шкалами I и II, I и III, II и III (ДИ не перекрываются); ДИ — доверительный интервал; AUC —площадь под характеристической кривой; Sens — чувствительность; Spec — специфичность; Acc — точность; PPV — положительная прогностическая ценность; AIR — доля исследований, получивших заключение «патология»; CDR — коэффициент выявления случаев; FNR — коэффициент ложных отрицательных; MCC — коэффициент корреляции Мэттьюса.

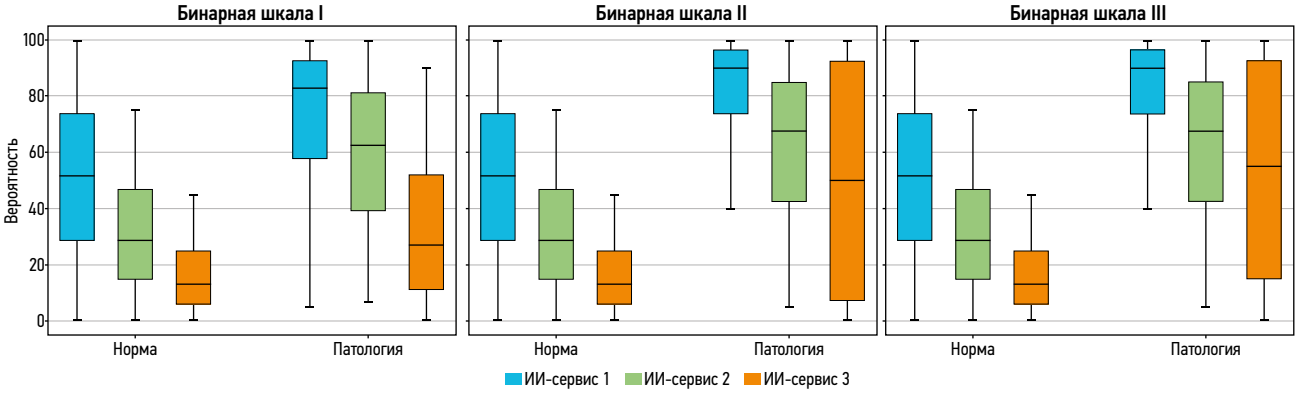

Рис. 4. Распределение результатов работы трёх сервисов искусственного интеллекта при проведении анализа трёх наборов данных: по оси X — варианты сервиса искусственного интеллекта; по оси Y —вероятность; данные представлены в виде: центральная линия — медиана; края «ящика» — первый (Q1) и третий (Q3) квартиль; «усы» — минимальное и максимальное значение данных; ИИ-сервис — сервис на основе искусственного интеллекта.

0,10, 0,03 и 0,03 для шкал I, II и III соответственно. Для скрининга важны различия в формировании диагностических шкал, поскольку необходимо оценивать риск пропуска патологии. Например, для категории BI-RADS 3 требуется дополнительное обследование, по результатам которого часть пациентов может быть отнесена к категориям с более высокой степенью злокачественности. Именно поэтому использование шкалы II позволит снизить вероятность пропуска патологии, так как в данной шкале категория BI-RADS 3 относится к группе «патология».

Примечательно, что в шкале I на частоту встречаемости патологии не влияет как присутствие категории BI-RADS 3 в группе «патология», так и полное отсутствие категории BI-RADS 3 в наборе данных. Однако, добавление этой категории в группу «норма» значительно увеличивает расчётную частоту встречаемости ЗНО в популяции скрининга. Такая же тенденция в категории BI-RADS 3 по шкале I наблюдается и для показателя площади под характеристической кривой AUC, который значимо не различается для шкал II и III.

Сравнение трёх выбранных для исследования ИИС по показателям диагностической точности выявило, что значения метрик AUC, специфичности, чувствительность, точности, PPV, CDR и MCC выше для ИИС-1 по сравнению со всеми шкалами, в то время как, метрика AIR имеет самые высокие значения для ИИС-3, а метрика FNR — для ИИС-2 и -3 в зависимости от бинарной шкалы. В целом эти результаты указывают на лучшую производительность ИИС-1. Сравнение максимальных значений индекса Юдена также показало наибольшую точность для ИИС-1 относительно всех шкал. Тем не менее оценка с помощью индекса Юдена показала статистически значимые различия между всеми шкалами и сервисами, тогда как при использовании доверительных интервалов,

Таблица 6. Значения метрик диагностической точности, измеренные для результатов трёх сервисов искусственного интеллекта относительно заключений врача

Метрика	Бинарная шкала	ИИС-1	ИИС-2	ИИС-3
Порог отсечения	I	64	32	10
	II	75	44	20
	III	74	44	20
AUC	I	0,671 [0,666; 0,676] *	0,647 [0,641; 0,652]	0,597 [0,592; 0,602]
	II	0,750 [0,740; 0,759] *	0,698 [0,689; 0,708]	0,713 [0,704; 0,722]
	III	0,755 [0,746; 0,763] *	0,708 [0,699; 0,717]	0,720 [0,713; 0,728]
Sens	I	0,610 [0,600; 0,620] *	0,602 [0,591; 0,613] *	0,578 [0,569; 0,587]
	II	0,687 [0,669; 0,706] *	0,609 [0,591; 0,627]	0,676 [0,658; 0,693] *
	III	0,689 [0,672; 0,705] *	0,616 [0,598; 0,633]	0,679 [0,664; 0,693] *
Spec	I	0,731 [0,728; 0,734] *	0,691 [0,688; 0,694]	0,616 [0,613; 0,619]
	II	0,812 [0,810; 0,815] *	0,788 [0,785; 0,790]	0,749 [0,746; 0,752]
	III	0,821 [0,819; 0,824] *	0,800 [0,798; 0,803]	0,762 [0,759; 0,765]
Acc	I	0,719 [0,717; 0,722] *	0,682 [0,679; 0,685]	0,612 [0,609; 0,615]
	II	0,809 [0,806; 0,811] *	0,782 [0,779; 0,785]	0,747 [0,744; 0,750]
	III	0,817 [0,815; 0,819] *	0,794 [0,792; 0,797]	0,760 [0,757; 0,762]
PPV	I	0,202 [0,199; 0,205] *	0,178 [0,175; 0,181]	0,143 [0,141; 0,145]
	II	0,102 [0,099; 0,105] *	0,082 [0,079; 0,084]	0,077 [0,075; 0,079]
	III	0,113 [0,110; 0,116] *	0,093 [0,090; 0,095]	0,086 [0,084; 0,088]
AIR	I	302,800 [299,851; 305,749]	338,140 [335,251; 341,029]	403,640 [400,720; 406,560] *
	II	202,600 [200,315; 204,885]	224,330 [221,624; 227,036]	263,320 [260,371; 266,269] *
	III	195,130 [192,823; 197,437]	212,940 [210,594; 215,286]	251,890 [249,124; 254,656] *
CDR	I	61,040 [60,041; 62,039] *	60,210 [59,113; 61,307] *	57,780 [56,889; 58,671]
	II	20,620 [20,062; 21,178] *	18,280 [17,735; 18,825]	20,270 [19,750; 20,790] *
	III	22,040 [21,507; 22,573] *	19,700 [19,131; 20,269]	21,720 [21,260; 22,180] *
FNR	I	0,390 [0,380; 0,400]	0,398 [0,387; 0,409]	0,422 [0,413; 0,431] *
	II	0,313 [0,294; 0,331]	0,391 [0,373; 0,409] *	0,324 [0,307; 0,342]
	III	0,311 [0,295; 0,328]	0,384 [0,367; 0,402] *	0,321 [0,307; 0,336]
MCC	I	0,223 [0,217; 0,230] *	0,186 [0,179; 0,193]	0,118 [0,113; 0,124]
	II	0,212 [0,204; 0,220] *	0,163 [0,155; 0,170]	0,165 [0,158; 0,172]
	III	0,227 [0,219; 0,234] *	0,179 [0,171; 0,187]	0,179 [0,173; 0,185]

Примечание. ИИС — сервис искусственного интеллекта; данные представлены в виде среднего [95% ДИ]; * — статистически значимые различия между сервисами 1 и 2, 1 и 3, 2 и 3 (ДИ не перекрываются); ДИ — доверительный интервал; AUC —площадь под характеристической кривой; Sens — чувствительность; Spec — специфичность; Acc — точность; PPV — положительная прогностическая ценность; AIR — доля исследований, получивших заключение «патология»; CDR — коэффициент выявления случаев; FNR — коэффициент ложных отрицательных; MCC — коэффициент корреляции Мэттьюса.

рассчитанных методом бутстрэппинга, метрика точности CDR и чувствительность значимо не различались между некоторыми сервисами.

При выборе ИИС и их версий так же, как и в случае бинарных шкал, важно учитывать контекст целей их использования. Например, при необходимости раннего обнаружения рентгенологических признаков ЗНО основным показателем диагностической точности будет чувствительность Sens, так как ИИС должен обнаруживать как можно больше действительно положительных, то есть патологических случаев. Второй важный показатель — это метрика FNR, минимизация которой позволит снизить количество пропущенных случаев патологии. Согласно результатам настоящего исследования, наибольшую чувствительность имеют ИИС-1 и ИИС-2. Однако, самое высокое значение показателя FNR было получено для ИИС-2 по отношению к шкале II и III. Что касается выбора версии, то в данном случае лучше всего соответствует цели применение ИИС-1 в версии 3, а ИИС-2 — в версии 2.

При необходимости получить максимальную точность

Таблица 7. Значения метрик диагностической точности, измеренные для результатов трёх версий сервиса искусственного интеллекта 1 относительно заключения врача

Шкала	ИИС	Версия ИИС	Порог отсечения	AUC	Sens	Spec	Acc
I	1	1	29	0,633 [0,627; 0,638]	0,617 [0,607; 0,627]	0,648 [0,645; 0,652]	0,645 [0,642; 0,648]
		2	68	0,674 [0,669; 0,679]	0,643 [0,634; 0,653] *	0,705 [0,702; 0,708]	0,699 [0,696; 0,702]
		3	66	0,702 [0,697; 0,706] *	0,647 [0,638; 0,656] *	0,757 [0,754; 0,760] *	0,746 [0,743; 0,748] *
II	1	1	57	0,684 [0,674; 0,693]	0,525 [0,506; 0,545]	0,842 [0,840; 0,844] *	0,833 [0,830; 0,835]
		2	78	0,750 [0,741; 0,760]	0,683 [0,664; 0,702]	0,817 [0,815; 0,819]	0,813 [0,811; 0,815]
		3	79	0,782 [0,774; 0,791] *	0,725 [0,708; 0,742] *	0,840 [0,837; 0,842] *	0,836 [0,834; 0,839] *
III	1	1	57	0,701 [0,692; 0,711]	0,548 [0,530; 0,566]	0,855 [0,853; 0,857] *	0,845 [0,843; 0,847] *
		2	78	0,756 [0,746; 0,766]	0,678 [0,659; 0,698]	0,833 [0,831; 0,836]	0,828 [0,826; 0,831]
		3	75	0,789 [0,783; 0,796] *	0,758 [0,745; 0,770] *	0,821 [0,819; 0,823]	0,819 [0,817; 0,821]

Шкала	ИИС	Версия ИИС	PPV	AIR	CDR	FNR	MCC
I	1	1	0,163 [0,161; 0,166]	378,070 [375,137; 381,003] *	61,690 [60,650; 62,730]	0,383 [0,373; 0,393] *	0,164 [0,157; 0,171]
		2	0,195 [0,193; 0,198]	329,660 [326,815; 332,505]	64,340 [63,384; 65,296] *	0,357 [0,347; 0,366]	0,223 [0,216; 0,229]
		3	0,228 [0,225; 0,231] *	283,710 [280,851; 286,569]	64,670 [63,768; 65,572] *	0,353 [0,344; 0,362]	0,269 [0,263; 0,275] *
II	1	1	0,093 [0,090; 0,097]	169,010 [166,841; 171,179]	15,760 [15,175; 16,345]	0,475 [0,455; 0,494] *	0,167 [0,158; 0,176]
		2	0,104 [0,101; 0,106]	198,140 [195,926; 200,354] *	20,500 [19,931; 21,069]	0,317 [0,298; 0,336]	0,214 [0,206; 0,222]
		3	0,123 [0,120; 0,126] *	177,240 [174,831; 179,649]	21,750 [21,234; 22,266] *	0,275 [0,258; 0,292]	0,253 [0,245; 0,260] *
III	1	1	0,111 [0,108; 0,115]	158,040 [155,998; 160,082]	17,540 [16,969; 18,111]	0,452 [0,434; 0,470] *	0,195 [0,186; 0,203]
		2	0,119 [0,115; 0,122] *	183,040 [180,595; 185,485]	21,710 [21,078; 22,342]	0,322 [0,302; 0,341]	0,233 [0,224; 0,242]
		3	0,123 [0,121; 0,125] *	197,480 [195,148; 199,812] *	24,240 [23.826; 24,654] *	0,242 [0,230; 0,255]	0,256 [0,250; 0,262] *

Примечание. ИИС — сервис искусственного интеллекта; данные представлены в виде среднего [95% ДИ]; * — ДИ не перекрываются, что указывает на статистическую значимость различий; ДИ — доверительный интервал; AUC —площадь под характеристической кривой; Sens — чувствительность; Spec — специфичность; Acc — точность; PPV — положительная прогностическая ценность; AIR — доля исследований, получивших заключение «патология»; CDR — коэффициент выявления случаев; FNR — коэффициент ложных отрицательных; MCC — коэффициент корреляции Мэттьюса.

DOI: https://doi.org/10.17816/DD625967

Таблица 8. Значение метрик диагностической точности, измеренные для результатов трёх версий сервиса искусственного интеллекта 2 относительно заключения врача

Шкала	ИИС	Версия ИИС	Порог отсечения	AUC	Sens	Spec	Acc
I	2	1	32	0,660 [0,655; 0,665] *	0,592 [0,581; 0,602]	0,728 [0,725; 0,731] *	0,714 [0,712; 0,71] *
		2	30	0,647 [0,642; 0,652]	0,621 [0.611; 0,631] *	0,673 [0.670; 0,676]	0,668 [0,665; 0,671]
		3	32	0,654 [0,649; 0,660] *	0,619 [0,609; 0,629] *	0,690 [0,687; 0,693]	0,683 [0,680; 0,685]
II	2	1	44	0,712 [0,702; 0,721] *	0,587 [0,568; 0,606]	0,837 [0,834; 0,839]*	0,829 [0,827; 0,832] *
		2	42	0,708 [0,700; 0,716] *	0,653 [0,637; 0,669] *	0,763 [0,760; 0,765]	0,759 [0,757; 0,762]
		3	44	0,688 [0,679; 0,696]	0,582 [0,566; 0,598]	0,793 [0,790; 0,795]	0,786 [0,784; 0,789]
III	2	1	44	0,709 [0,701; 0,718] *	0,576 [0,560; 0,593]	0,843 [0,841; 0,845] *	0,834 [0,832; 0,836] *
		2	39	0,722 [0,714; 0,731] *	0,688 [0,672; 0,705] *	0,756 [0,753; 0,759]	0,754 [0,751; 0,756]
		3	44	0,698 [0,689; 0,706]	0,587 [0,570; 0,604]	0,808 [0,806; 0,810]	0,801 [0,799; 0,803]

Шкала	ИИС	Версия ИИС	PPV	AIR	CDR	FNR	MCC
I	2	1	0,195 [0,192; 0,198] *	304,020 [301,225; 306,815]	59,190 [58,143; 60,237]	0,408 [0,398; 0,419] *	0,209 [0,202; 0,216] *
		2	0,174 [0,172; 0,177]	356,470 [353,548; 359,392] *	62,110 [61,135; 63,085] *	0,379 [0,369; 0,389]	0,184 [0,178; 0,191]
		3	0,182 [0,179; 0,184]	341,050 [338,332; 343,768]	61,880 [60,877; 62,883]*	0,381 [0,371; 0,391]	0,195 [0,189; 0,202]
II	2	1	0,100 [0,097; 0,104] *	175,920 [173,412; 178,428]	17,600 [17,029; 18,171]	0,413 [0,394; 0,432] *	0,190 [0,181; 0,199] *
		2	0,079 [0,077; 0,080]	249,890 [247,158; 252,622] *	19,590 [19,115; 20,065] *	0,347 [0,331; 0,363]	0,164 [0,158; 0,170]
		3	0,080 [0,078; 0,082]	218,500 [216,045; 220,955]	17,470 [16,991; 17,949]	0,418 [0,402; 0,434] *	0,155 [0,148; 0,162]
III	2	1	0,108 [0,105; 0,111] *	170,700 [168,638; 172,762]	18,440 [17,907; 18,973]	0,424 [0,407; 0,440] *	0,196 [0,188; 0,204] *
		2	0,085 [0,083; 0,088]	258,390 [255,703; 261,077] *	22,030 [21,511; 22,549] *	0,312 [0,295; 0,328]	0,179 [0,172; 0,186]
		3	0,092 [0,089; 0,095]	204,570 [202,244; 206,896]	18,790 [18,255; 19,325]	0,413 [0,396; 0,430] *	0,173 [0,165; 0,180]

Примечание. ИИС — сервис искусственного интеллекта; данные представлены в виде среднего [95% ДИ]; * — ДИ не перекрываются, что указывает на статистическую значимость различий; ДИ — доверительный интервал; AUC —площадь под характеристической кривой; Sens — чувствительность; Spec — специфичность; Acc — точность; PPV — положительная прогностическая ценность; AIR — доля исследований, получивших заключение «патология»; CDR — коэффициент выявления случаев; FNR — коэффициент ложных отрицательных; MCC — коэффициент корреляции Мэттьюса.

DOI: https://doi.org/10.17816/DD625967

интерпретации важна специфичность, для уменьшения числа ложно положительных результатов, и значение метрики PPV, которая позволяет убедиться, что большинство положительных результатов классификации действительно являются патологическими случаями. ИИС-1 имеет самую высокую эффективность в таких случаях. При этом для различных версий ИИС-1 значения метрик специфичности и PPV отличались для разных шкал.

Снижение метрики AIR позволяет сократить время, которое врачи-рентгенологи тратят на дополнительную интерпретацию исследования, если классификация ИИС ясна и надёжна. В нашем исследовании самое низкое значение AIR продемонстрировал ИИС-1 относительно всех шкал.

Чтобы получить общую оценку правильности классификации случаев «норма» и «патология» для ИИС и их версий необходимо обратить внимание на следующие метрики: точность Acc, высокое значение которой отражает в какой мере оба класса могут быть правильно классифицированы; коэффициент Мэтьюса, который оценивает общую производительность классификатора, учитывая все аспекты матрицы ошибок. В настоящем исследовании самую высокую общую оценку правильности классификации имеет ИИС-1 в версии 3. Важно отметить, что результаты работы ИИС сопоставляли с разметкой, выполненной врачом, и приведённой к соответствующему классу, что накладывает определенные ограничения, поскольку важно знать показатели диагностической точности врача-рентгенолога. Диагностическую точность врача можно оценить на эталонном наборе данных, в котором истинное значение определено по данным гистологии. Такое исследование уже было проведено и показало высокую диагностическую точность врачей-рентгенологов (AUC составляет 0,928) [15]. В настоящем исследовании мы получили более низкие значения AUC для выбранных ИИС, что свидетельствует о необходимости доработки решений, что и было выполнено в период с 2020 по 2022 год. С другой стороны, важно обращать внимание на значения показателей чувствительности и специфичности, а они уступают таковым для врача-рентгенолога [15]. В контексте исследования, описанного в настоящей статье, мы не ставили задачу оптимизировать настройку того или иного показателя. Важно заметить, что ИИС при одном и том же значении AUC может быть настроен на любое значение чувствительности. Например, настройка на чувствительность, близкую к 100%, позволит не пропускать патологию, но при этом даст большое количество ложно положительных результатов. В последующих работах мы планируем детально изучить возможности тонкой настройки ИИС с целью оптимизации показателей чувствительности и специфичности.

Применение технологий ИИ в маммографии в первую очередь видится в замене первого чтения, что будет способствовать повышению точности диагностики ЗНО молочных желёз [16], за счёт увеличения чувствительности. Возможен альтернативный способ применения ИИ — в качестве инструмента для сортировки исследований, когда настройки чувствительности близки к 100%. В этом случае врачам-рентгенологам можно не описывать исследования, которые ИИ классифицировал как «без патологии», а сразу передавать их в виде электронной медицинской записи. Такой способ показал свою перспективность для автономной сортировки результатов флюорографии в недавнем исследовании [17]. Что касается маммографии, такой сценарий может быть менее эффективным из-за наличия многочисленной группы доброкачественных изменений, которые также могут требовать внимания и дополнительного изучения.

Ограничения исследования

Данное исследование содержит результаты первых трёх лет масштабного исследования «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» [18] и не затрагивает вопросов оптимальной настройки ИИС. Ограничением исследования является то, что показатель AUC может быть недостаточно информативным для оценки производительности ИИС в клинической среде, поскольку выбор конкретного порога не всегда применим к реальным условиям использования сервиса. Кроме того, метрики чувствительности Sens и специфичности Spec не учитывают распространенность заболевания в популяции. Именно поэтому в дальнейших исследованиях мы планируем применять другие методы оценки эффективности использования ИИС в клинической практике, а также использовать результаты гистологической верификации в качестве истинных значений. Важно отметить, что в настоящее исследование включены только маммографические исследования, содержащие результаты обработки ИИС, при этом не анализировали маммограммы, по которым ИИС не вернул результата. Кроме того, в настоящем исследовании не уделяли внимания оценке качества работы ИИС с исследованиями молочных желёз при наличии инородных тел (имплантов) и с изменениями, вызванными лучевой терапией, что, несомненно, представляет большой практический интерес и будет являться целью одной из следующих работ.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании показано, что выбор способа формирования бинарной шкалы «норма/патология» влияет на результаты сравнительной оценки метрик диагностической точности различных типов и версий ИИС. В то же время индекс Юдена позволяет обнаружить статистически значимую разницу между значениями показателей точности ИИС и диагностических шкал, а выбор метрик для проведения сравнительной оценки ИИС зависит от клинической задачи. С другой

стороны, настройка ИИС методом максимизации индекса Юдена позволяет получать сбалансированные значения чувствительности и специфичности, что не всегда может быть целесообразно с клинической точки зрения.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ

Источник финансирования. Данная статья подготовлена авторским коллективом в рамках НИР/НИОКР «Научные методологии устойчивого развития технологий искусственного интеллекта в медицинской диагностике» (№ ЕГИСУ: 123031500004-5) в соответствии с Приказом от 21.12.2022 № 1196 «Об утверждении государственных заданий, финансовое обеспечение которых осуществляется за счёт средств бюджета города Москвы государственным бюджетным (автономным) учреждениям подведомственным Департаменту здравоохранения города Москвы, на 2023 год и плановый период 2024 и 2025 годов» Департамента здравоохранения города Москвы.

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

Вклад авторов. Авторы подтверждают соответствие своего авторства международным критериям ICMJE (все авторы внесли существенный вклад в разработку концепции, проведение исследования и подготовку статьи, прочли и одобрили финальную версию перед публикацией). Наибольший вклад распределён следующим образом: Ю.А. Васильев, А.В. Владзимирский, О.В. Омелянская, А.В. Колсанов — концепция исследования; К.М. Арзамасов — планирование и руководство исследованием; С.С. Семёнов — анализ данных; Л.Е. Аксёнова — анализ данных, написание текста публикации.

ADDITIONAL INFORMATION

Funding source. This article was prepared by a group of authors as a part of the research and development effort titled "Scientific methodologies for sustainable development of artificial intelligence technologies in medical diagnostics" (USIS No.: 123031500004-5) in accordance with the Order No. 1196 dated December 21, 2022 "On approval of state assignments funded by means of allocations from the budget of the city of Moscow to the state budgetary (autonomous) institutions subordinate to the Moscow Health Care Department, for 2023 and the planned period of 2024 and 2025" issued by the Moscow Health Care Department.

Competing interests. The authors declare that they have no competing interests.

Authors' contribution. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published and agree to be accountable for all aspects of the work. Yu.A. Vasiliev, A.V. Vladzimirskyy, O.V. Omelyanskaya, A.V. Kolsanov — research concept; K.M. Arzamasov — planning and directing the research; S.S. Semenov — data analysis; L.E. Axenova — data analysis, text writing.

СПИСОК ЛИТЕРАТУРЫ

1. Seely J.M., Alhassan T. Screening for breast cancer in 2018-what should we be doing today? // Curr Oncol. 2018. Vol. 25, Suppl. 1. P. S115–S124. doi: 10.3747/co.25.3770

2. Кандоба В.И. Искусственный интеллект в скрининговой маммографии. Клиническое использование, проблемы и направления развития [интернет]. Режим доступа: https://www.itmportal.ru/upload/iblock/69e/7q981uhfaxjhcntal0exngxtq43xeth2/2.2.3.-Kandoba-ITM_AI-2022.pdf Дата обращения: 20.08.2023

3. Цельс. Система поддержки принятия врачебных решений на базе технологий искусственного интеллекта для анализа цифровых медицинских изображений. Маммография [интернет]. Режим доступа: https://celsus.ai/products-mammography/ Дата обращения: 20.08.2023

4. Kim H.E., Kim H.H., Han B.K., et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study // Lancet Digit Health. 2020. Vol. 2, N 3. P. e138–e148. doi: 10.1016/S2589-7500(20)30003-0

5. Yoon J.H., Strand F., Baltzer P.A.T., et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis // Radiology. 2023. Vol. 307, N 5. ID: e222639. doi: 10.1148/radiol.222639

6. Zhou X.-H., Obuchowski N.A., McClish D.K. Statistical Methods in Diagnostic Medicine. NJ: John Wiley & Sons, Inc.; 2011. doi: 10.1002/9780470906514

7. Habibzadeh F., Habibzadeh P,, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results // Biochem Med (Zagreb). 2016. Vol. 26, N 3. P. 297–307. doi: 10.11613/BM.2016.034

8. Schaffter T., Buist D.S.M., Lee C.I., et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms // JAMA Netw Open. 2020. Vol. 3, N 3. ID: e200265. doi: 10.1001/jamanetworkopen.2020.0265

9. McKinney S.M., Sieniek M., Godbole V., et al. International evaluation of an AI system for breast cancer screening // Nature. 2020. Vol. 577, N 7788. P. 89–94. doi: 10.1038/s41586-019-1799-6

10. Nam J.G., Kim M., Park J., et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs // Eur Respir J. 2021. Vol. 57, N 5. ID: 2003061. doi: 10.1183/13993003.03061-2020

11. Сахнов С.Н., Аксенов К.Д., Аксенова Л.Е., и др. Разработка модели скрининга катаракты с использованием открытого набора данных и алгоритмов глубокого машинного обучения. Офтальмохирургия. 2022. № S4. С. 13–20. EDN: VEGPAW doi: 10.25276/0235-4160-2022-4S-13-20

12. King G., Zeng L. Logistic Regression in Rare Events Data // Political Analysis. 2001. Vol. 9, N 2. P. 137–163. doi: 10.1093/oxfordjournals.pan.a004868

13. Chen F., Xue Y., Tan M.T., Chen P. Efficient statistical tests to compare Youden index: accounting for contingency correlation // Stat Med. 2015. Vol. 34, N 9. P. 1560–1576. doi: 10.1002/sim.6432

14. Васильев Ю.А., Владзимирский А.В., Шарова Д.Е., и др. Клинические испытания систем искусственного интеллекта (лучевая

диагностика). Москва: Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», 2023. 40 с. EDN: PUIJLD
15. Арзамасов К.М., Васильев Ю.А., Владзимирский А.В., и др. Применение компьютерного зрения для профилактических исследований на примере маммографии // Профилактическая медицина. 2023. Т. 26, № 6. С. 117–123. EDN: YBKHPS doi: 10.17116/profmed202326061117
16. Васильев Ю.А., Тыров И.А., Владзимирский А.В., и др. Двойной просмотр результатов маммографии с применением технологий искусственного интеллекта: новая модель организации массовых профилактических исследований // Digital Diagnostics. 2023. Т. 4, № 2. С. 93–104. EDN: VRIEOH doi: 10.17816/DD321423
17. Васильев Ю.А., Тыров И.А., Владзимирский А.В., и др. Новая модель организации массовых профилактических исследований, основанная на автономном искусственном интеллекте для сортировки результатов флюорографии // Здоровье населения и среда обитания — ЗНиСО. 2023. Т. 31, № 11. С. 23–32. EDN: SYIQBX doi: 10.35627/2219-5238/2023-31-11-23-32
18. Владзимирский А.В., Васильев Ю.А., Арзамасов К.М., и др. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента. Москва: Издательские решения, 2022. EDN: FOYLXK

REFERENCES

1. Seely JM, Alhassan T. Screening for breast cancer in 2018-what should we be doing today? *Curr Oncol.* 2018;25(suppl 1):S115–S124. doi: 10.3747/co.25.3770
2. Artificial intelligence in mammography screening. *Clinical applications, issues and directions for development* [Internet; cited 20 August 2023]. Available from: https://www.itmportal.ru/upload/iblock/69e/7q981uhfaxjhcntal0exngxtq43xeth2/2.2.3.-Kandoba-ITM_AI-2022.pdf (in Russ.)
3. Celsus — *AI-software for analysis of X-ray and CT studies.* Mammography [Internet; cited 20 Aug 2023]. Available from: https://celsus.ai/products-mammography/
4. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health.* 2020;2(3):e138–e148. doi: 10.1016/S2589-7500(20)30003-0
5. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. *Radiology.* 2023;307(5):e222639. doi: 10.1148/radiol.222639
6. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine.* NJ: John Wiley & Sons, Inc.; 2011. doi: 10.1002/9780470906514
7. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb).* 2016;26(3):297–307. doi: 10.11613/BM.2016.034
8. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open.* 2020;3(3):e200265. doi: 10.1001/jamanetworkopen.2020.0265
9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature.* 2020;577(7788):89–94. doi: 10.1038/s41586-019-1799-6
10. Nam JG, Kim M, Park J, et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur Respir J.* 2021;57(5):2003061. doi: 10.1183/13993003.03061-2020
11. Sakhnov SN, Axenov KD, Axenova LE, et al. Development of a cataract screening model using an open dataset and deep machine learning algorithms. *Fyodorov Journal of Ophthalmic Surgery.* 2022;(S4):13–20. EDN: VEGPAW doi: 10.25276/0235-4160-2022-4S-13-20
12. King G, Zeng L. Logistic Regression in Rare Events Data. *Political Analysis.* 2001;9(2):137–163. doi: 10.1093/oxfordjournals.pan.a004868
13. Chen F, Xue Y, Tan MT, Chen P. Efficient statistical tests to compare Youden index: accounting for contingency correlation. *Stat Med.* 2015;34(9):1560–1576. doi: 10.1002/sim.6432
14. Vasiliev YuA, Vladzimirsky AV, Sharova DE, et al. *Clinical trials of artificial intelligence systems (radiation diagnostics).* Moscow: State budgetary healthcare institution of the city of Moscow «Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Department». 2023. 40 p. (In Russ.) EDN: PUIJLD
15. Arzamasov KM, Vasilev YuA, Vladzymyrskyy AV, et al. The use of computer vision for the mammography preventive research. *The Russian Journal of Preventive Medicine.* 2023;26(6):117–123. EDN: YBKHPS doi: 10.17116/profmed202326061117
16. Vasilev YuA, Tyrov IA, Vladzymyrskyy AV, et al. Double-reading mammograms using artificial intelligence technologies: A new model of mass preventive examination organization. *Digital Diagnostics.* 2023;4(2):93–104. EDN: VRIEOH doi: 10.17816/DD321423
17. Vasilev YuA, Tyrov IA, Vladzymyrskyy AV, et al. A New Model of Organizing Mass Screening Based on Stand-Alone Artificial Intelligence Used for Fluorography Image Triage. *Public Health and Life Environment — PH&LE.* 2023;31(11):23-32. EDN: SYIQBX doi: 10.35627/2219-5238/2023-31-11-23-32
18. Vladzimirskyy AV, Vasilev YuA, Arzamasov KM, et al. *Computer vision in radiology: the first stage of the Moscow experiment.* Moscow: Izdatel'skie resheniya; 2022. (In Russ.) EDN: FOYLXK

ОБ АВТОРАХ

*** Арзамасов Кирилл Михайлович,** канд. мед. наук;
адрес: Россия, 127051, Москва, ул. Петровка, д. 24, стр. 1;
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

AUTHORS' INFO

*** Kirill M. Arzamasov,** MD, Cand. Sci. (Medicine);
address: 24 bldg. 1 Petrovka str., 127051, Moscow, Russia;
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

Васильев Юрий Александрович, канд. мед. наук;
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: VasilevYA1@zdrav.mos.ru

Колсанов Александр Владимирович, д-р мед. наук, профессор;
ORCID: 0000-0002-4144-7090;
eLibrary SPIN: 2028-6609;
e-mail: a.v.kolsanov@samsmu.ru

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

Омелянская Ольга Васильевна;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: OmelyanskayaOV@zdrav.mos.ru

Семёнов Серафим Сергеевич;
ORCID: 0000-0003-2585-0864;
eLibrary SPIN: 4790-0416;
e-mail: SemenovSS3@zdrav.mos.ru

Аксёнова Любовь Евгеньевна;
ORCID: 0000-0003-0885-1355;
eLibrary SPIN: 7705-6293;
e-mail: AksenovaLE@zdrav.mos.ru

Yuriy A. Vasilev, MD, Cand. Sci. (Medicine);
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: VasilevYA1@zdrav.mos.ru

Alexander V. Kolsanov, MD, Dr. Sci. (Medicine), Professor;
ORCID: 0000-0002-4144-7090;
eLibrary SPIN: 2028-6609;
e-mail: a.v.kolsanov@samsmu.ru

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

Olga V. Omelyanskaya;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: OmelyanskayaOV@zdrav.mos.ru

Serafim S. Semenov;
ORCID: 0000-0003-2585-0864;
eLibrary SPIN: 4790-0416;
e-mail: SemenovSS3@zdrav.mos.ru

Lubov E. Axenova;
ORCID: 0000-0003-0885-1355;
eLibrary SPIN: 7705-6293;
e-mail: AksenovaLE@zdrav.mos.ru

* Автор, ответственный за переписку / Corresponding author