КРИТЕРИИ ПРИМЕНИМОСТИ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ ПРОФИЛАКТИЧЕСКИХ ИССЛЕДОВАНИЙ НА ПРИМЕРЕ РЕНТГЕНОГРАФИИ И ФЛЮОРОГРАФИИ ОРГАНОВ ГРУДНОЙ КЛЕТКИ

К.М. Арзамасов[1], С.С. Семенов[1], Д.Ю. Кокина[1], Т.М. Бобровская[1], Н.А. Павлов[1], Ю.С. Кирпичев[1,2], А.Е. Андрейченко[1,3,4], А.В. Владзимирский[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы, Москва

[2] ООО "Медскан", Москва

[3] Физико-технический факультет университета "ИТМО", Санкт-Петербург

[4] ООО "К-Скай", Петрозаводск

Цель: в условиях постоянного роста количества разработанных алгоритмов компьютерного зрения на основе искусственного интеллекта (ИИ) для медицинской диагностики возникает необходимость определения критериев для принятия решения о целесообразности их практического применения для массовых профилактических исследований населения.

Материал и методы: Исследование с участием нескольких врачей-рентгенологов проводилось на "Веб-платформе для оценки рентгенологических исследований" на размеченном наборе данных из набора рентгенограмм и флюорограмм в передней прямой проекции. На этом же наборе данных с помощью "Платформы тестирования при смене версионности" были получены ответы от двух коммерческих алгоритмов компьютерного зрения на основе ИИ, разработанных для анализа цифровых рентгенограмм. Оценка полученных от врачей и алгоритмов результатов (бинарных, в терминах "с патологией" и "без патологии") проводилась с помощью ROC-анализа. Для порогового значения, рассчитанного по методу Юдена, определялись метрики: чувствительность, специфичность и точность.

Результаты: Рассчитаны метрики диагностической точности для усредненной оценки врачей-рентгенологов и алгоритмов компьютерного зрения на основе ИИ при поиске патологических изменений на рентгенограммах органов грудной клетки в передней прямой проекции по данным ROC-анализа. Средние значения показателей диагностической точности врачей рентгенологов превзошли показатели ИИ алгоритмов.

Выводы: При принятии решения о внедрении в практику алгоритмов компьютерного зрения на основе ИИ для профилактических исследований следует руководствоваться метриками диагностической точности конкретного алгоритма, а в качестве целевых значений метрик использовать усредненный результат врачей при решении данной диагностической задачи.

Ключевые слова: *искусственный интеллект, лучевая диагностика, диагностическая точность, профилактические исследования*

DOI: 10.52775/1810-200X-2022-96-4-56-63

Введение

Повышение объемов лучевых диагностических исследований и количества диагностического оборудования значительно увеличивают нагрузку на врачей-рентгенологов [1]. Лавинообразное нарастание потока диагностических изображений, в первую очередь, рентгенографии и компьютерной томографии органов грудной клетки, отмечается в период пандемии COVID-19. Как правило, врачи-рентгенологи, работающие с рентгенографией и маммографией, имеют наибольшее количество исследований в смену [2].

Алгоритмы искусственного интеллекта (ИИ) активно интегрируются в рабочий процесс врачей-рентгенологов. Показатели точности алгоритмов, предоставляемые разработчиками, достаточно высоки [3–5], однако недостаточны для замены собой врача-рентгенолога [4]. Отдельные исследования указывают на возрастание точности интерпретации медицинских изображений при использовании сочетания методов ИИ и врача-рентгенолога [5].

Наибольшее количество скрининговых исследований в лучевой диагностике проводится по направлениям рентгенография (флюорография) органов грудной клетки и маммография. Особенностью этих исследований является наличие большой доли случаев без патологии. В этом ключе целесообразно рассматривать алгоритмы ИИ для классификации исследований в помощь врачу-рентгенологу.

Отдельные исследования [1, 6] посвящены оценке точности ИИ алгоритмов в анализе рентгенографии органов грудной клетки, тогда как подобных данных по работе с флюорографическими методиками практически нет.

Валидация описанных выше алгоритмов ИИ осуществляется на наборе тестовых данных, для которого рассчитываются метрики диагностической точности. При принятии решения о допуске конкретного ИИ алгоритма к практическому применению используют рекомендуемые усредненные пороговые значения, без привязки к специфике работы алгоритма ИИ или решаемой клинической задаче [7]. Однако для эффективного практического применения ИИ алгоритм должен отвечать определенным требованиям (согласно клинической задаче) и иметь диагностическую точность, сопоставимую с работой врача-рентгенолога.

Цель настоящего исследования: на основании диагностической точности экспертных оценок рентгенограмм врачами-рентгенологами разработать методологию оценки пороговых значений для алгоритмов компьютерного зрения на основе ИИ для анализа профилактической рентгенографии органов грудной клетки.

Материал и методы

Данное исследование основано на ранее зарегистрированном исследовании (https://clinicaltrials.gov/ct2/show/NCT04489992), одобренном локальным этическим комитетом.

Алгоритмы ИИ

В исследовании использовались лучшие по данным Лидерборда “Эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений” сервисы искусственного интеллекта (ИИ-алгоритмы) – по 2 коммерческих алгоритма от компаний-разработчиков для поиска и классификации патологических изменений на цифровой рентгенографии и флюорографии [8]. Ответы ИИ-алгоритмов получены на основании результатов тестирования при смене версионности на специально разработанной платформе тестирования. На рис. 1 приведены примеры обнаружения целевых патологий ИИ-алгоритмами. Ответы представлены в виде расчета вероятности наличия патологических изменений в исследовании вместе с пороговым значением, установленным для каждого алгоритма ИИ индивидуально в рамках настоящей работы. ИИ-алгоритмы при анализе изображений оценивают исследование по целому ряду патологических признаков, однако в качестве ответа получают одно вероятностное значение. Алгоритм, на основании которого принимается решение о выставлении единой вероятности патологии каждым ИИ-алгоритмом, предлагался компаниями самостоятельно и не являлся предметом изучения в настоящей работе. Бинарная классификация исследований по группам “без целевой патологии” и “с целевой патологией” осуществлялась по индивидуальному пороговому значению индекса Юдена для каждого алгоритма ИИ.

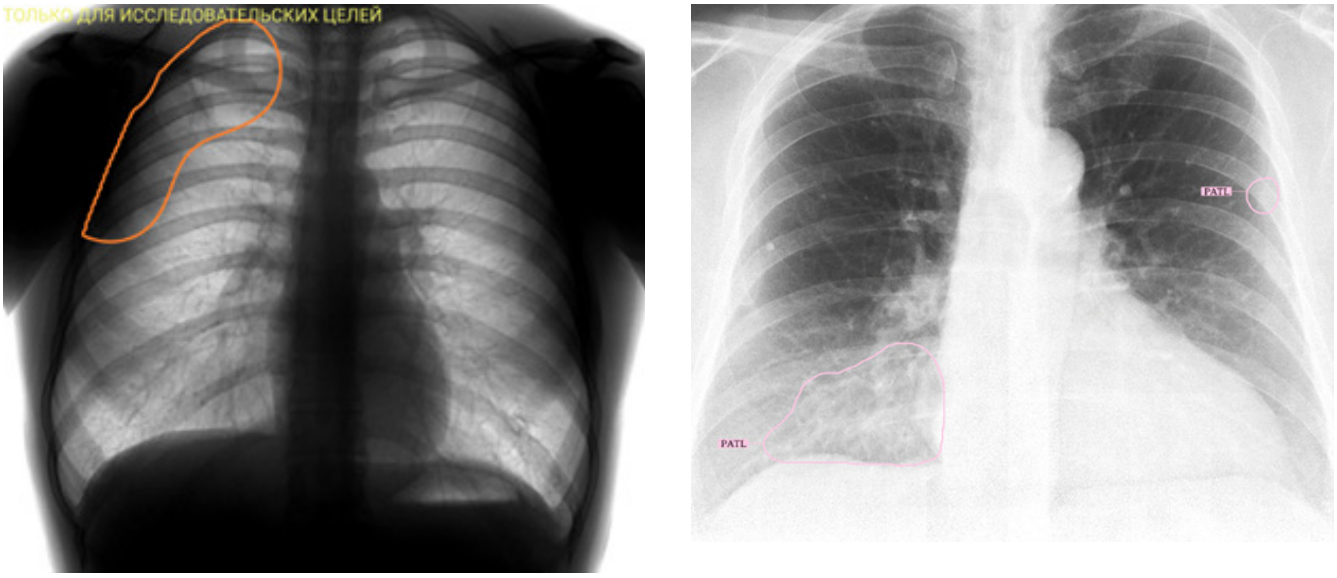

Рис. 1. *Примеры работы цифровой рентгенограммы органов грудной клетки в передней прямой проекции алгоритмом ИИ-1 (слева) и ИИ-2 (справа)*

Веб-платформа для сбора оценок врачей-рентгенологов

Для выполнения настоящего исследования была разработана “Веб-платформа для оценки рентгенологических исследований”, позволяющая участникам (врачам-рентгенологам) проводить оценку цифровых рентгенограмм и флюорограмм удалённо по сети Интернет. Серверная часть платформы представлена в виде программного интерфейса приложения (API) и выполнена на языке Python с использованием фреймворка Flask. Хранение результатов оценки и пользовательских данных осуществлялось в виде JSON-документов в NoSQL базе данных MongoDB. Браузерная часть сайта создана с помощью фреймворка Angular. Доступ на веб-платформу был организован с использованием имени пользователя и пароля (HTTP аутентификация), выданных каждому из участников. Доступ мог быть обеспечен с любого устройства, имеющего доступ к сети Интернет, при этом рекомендовалось использовать рабочее место врача-рентгенолога для визуализации исследования на качественном мониторе, используемом для просмотра медицинских исследований.

Для отображения исследований в формате DICOM был использован свободно распространяемый код системы передачи и архивации DICOM изображений Orthanc [(https://Www.Orthanc-Server.Com/Index.Php, n.d.)] и совместимый с ним веб-модуль просмотра изображений [(https://Www.Orthanc-Server.Com/Static.Php?Page=web-Viewer, n.d.)].

Для каждого исследования предоставлена информация о возрасте и поле обследуемого пациента. Дополнительная клиническая информация не предоставлена, что обеспечивает одинаковые условия для врачей и алгоритмов ИИ.

Для анализа рентгенограмм грудной клетки была использована только передняя прямая проекция.

У врачей-рентгенологов была возможность выбрать один из пяти вариантов ответов: 1) Определенно без патологии; 2) Возможно без патологии; 3) Затрудняюсь ответить; 4) Возможно с патологией; 5) Определенно с патологией. Каждый врач-рентгенолог мог проанализировать 20, 50 или 80 исследований.

Веб-платформа была доступна на протяжении всего времени исследования с 27 ноября по 13 декабря 2020 года.

Для решения задачи сравнения точности врачей и ИИ использованы оценки относительно наличия патологического процесса во всем исследовании в целом.

Врачи-рентгенологи

Разработанная веб-платформа была доступна врачам-рентгенологам из Российской Федерации и стран ближнего зарубежья. Врачи, не закончившие разметку выбранного количества исследований, исключались. В иссле-

дование были включены результаты 185 врачей-рентгенологов, проанализировавших рентгенограммы (РГ), и 69 врачей-рентгенологов, проанализировавших флюорограммы (ФЛГ). Среди врачей, оценивающих РГ преобладали врачи со стажем 1–5 лет – 60 врачей, по 36 врачей имели стаж 0–1 и 5–10 лет и 53 врача со стажем более 10 лет. Распределение по стажу среди врачей, описывающих ФЛГ, было следующим: 6 со стажем 0–1 год, 8 от 1 до 5 лет, 5 от 5 до 10 лет и 9 имели стаж более 10 лет.

Набор данных

В настоящее исследование было включено 140 цифровых рентгенограмм, из которых с патологическими изменениями было 47, а также 184 цифровых флюорограмм, с патологией – 84. Разметка исследований на "с целевой патологией" и "без целевой патологии" осуществлялась на основании консенсуса двух экспертов (врачей-рентгенологов со стажем более 5 лет). При этом в качестве целевой патологии на исследованиях рассматривались следующие патологические признаки: "Плевральный выпот", "Пневмоторакс", "Ателектаз", "Очаг затемнения", "Инфильтрация/консолидация", "Диссеминация", "Полость с распадом и уровнем жидкости", "Кальцинат", "Нарушение целостности кортикального слоя". Процентное содержание патологических находок и их классификация сопоставимы с показателями исследований, имеющихся в рутинной практике врачей-рентгенологов широкопрофильных медицинских организаций.

Коллективная оценка врачей

Коллективная оценка врачей выполнялась отдельно для рентгенографии, отдельно для флюорографии. Осуществлялась деперсонализация всех полученных ответов. Далее выполнялось разделение ответов по модальностям (рентгенография и флюорография). Осуществлялась балльная оценка и вычисление средней оценки среди ответов всех врачей по данному исследованию.

Статистическая обработка данных

Для уменьшения статистической значимости ответа одного врача в работу включались только те исследования, по которым была получена оценка от каждого алгоритма ИИ, а также содержащих не менее 5 оценок от врачей-рентгенологов. Оценка врача и алгоритмов осуществлялась на основании ROC-анализа. Это позволило минимизировать субъективность оценки исследования врачом-рентгенологом. Площадь под характеристической кривой рассчитывалась с 95 % доверительным интервалом методом Делонга (DeLong) [9]. Для определения оптимального порогового значения использовался максимум индекса Юдена (Youden) [10, 11], для данного порогового значения определялись метрики чувствительность, специфичность и точность.

Сравнительный анализ ROC AUC проводился с помощью перестановочного теста [12]. Проверялась нулевая гипотеза H_0 об отсутствии различий между ROC-кривыми против альтернативной H_1 о существовании различий. Уровень значимости p=0,05.

Результаты

Метрики диагностической точности на тестовом наборе данных приведены в табл. 1, характеристические кривые алгоритмов ИИ и "среднего" врача-рентгенолога представлены на рис. 2.

Как видно из приведённых данных, метрики диагностической точности "среднего" врача-рентгенолога для большинства показа-

Таблица 1

Метрики диагностической точности

	AUC (95 % ДИ)	Сравнение AUC ИИ и врачи, p	Чувствительность (95 % ДИ)	Специфичность (95 % ДИ)	Точность (95 % ДИ)
	Рентгенография				
Врачи	0,96 (0,93–1,0)		0,90 (0,82–0,99)	0,96 (0,9–1,0)	0,93 (0,88–0,98)
ИИ-1	0,91 (0,85–0,97)	0,019	0,86 (0,75–0,96)	0,89 (0,80–0,98)	0,88 (0,81–0,94)
ИИ-2	0,91 (0,84–0,98)	0,123	0,76 (0,63–0,89)	0,98 (0,94–1,0)	0,88 (0,81–0,94)
	Флюорография				
Врачи	0,97 (0,94–1,0)		1,0 (1,0–1,0)	0,91 (0,83–0,99)	0,96 (0,92–1,0)
ИИ-1	0,91 (0,86–0,97)	0,049	1,0 (1,0–1,0)	0,72 (0,60–0,85)	0,86 (0,79–0,93)
ИИ-2	0,95 (0,90–0,99)	0,459	0,89 (0,80–0,98)	0,96 (0,90–1,0)	0,92 (0,87–0,98)

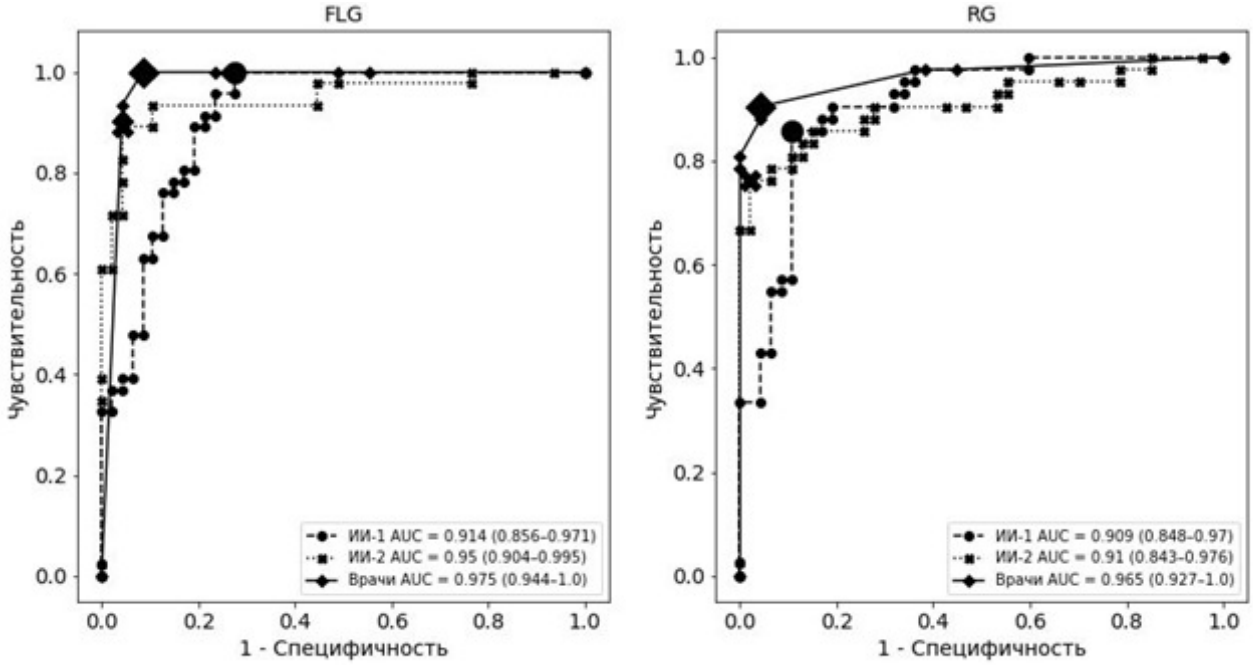

Рис. 2. *Результаты ROC-анализа оценок врачей и алгоритмов ИИ при анализе флюорограмм (FLG) и рентгенограмм (RG) органов грудной клетки для групп "с патологией" и "без патологии". Увеличенным маркером отмечена точка с оптимальным значением метрик (с максимизацией индекса Юдена)*

телей оказались выше, чем таковые для ИИ-1 и ИИ-2 алгоритмов, при этом в границах доверительных интервалов полученные величины имеют пересечения, что говорит об отсутствии статистически значимой разницы между этими показателями. Однако при сравнении ИИ-алгоритмов с помощью перестановочного теста мы наблюдаем статистически значимое (p<0,05) уменьшение ROC AUC ИИ-1 по сравнению с врачами как на рентгенографии, так и на флюорографии.

ИИ-1 и ИИ-2 для рентгенографии ОГК имели схожие средние значения AUC, а для флюорографии значения AUC у ИИ-2 оказались выше, однако и пересечение доверительных интервалов, и результаты перестановочного теста (при сравнении ИИ-алгоритмов между собой p=0,990 для рентгенографии и p=0,411 для флюорографии) указывают на отсутствие статистически значимой (p>0,05) разницы между алгоритмами.

При сопоставлении значений специфичности были получены значения, превышающие таковые для ИИ-1 и для врача-рентгенолога, но пересекающиеся в границах доверительных интервалов. Значения чувствительности для оптимального порога алгоритма ИИ-2 были существенно ниже таковых для ИИ-1 и врача-рентгенолога.

Показатели точности усредненной оценки врачей при поиске патологических изменений в органах грудной клетки превзошли показатели алгоритмов ИИ, при этом не было выявлено зависимости от модальности оцениваемого исследования – рентгенография или флюорография.

Обсуждение

В настоящее время в научной литературе встречается небольшое количество статей, целью которых является определение усредненных метрик диагностической точности для профилактических исследований в лучевой диагностике. Напротив, ежегодно публикуется большое количество работ, в том числе демонстрирующих высокие показатели диагностической точности и посвященных оценке метрик диагностической точности алгоритмов на основе ИИ для анализа диагностических исследований в лучевой диагностике [13–16],

Однако при принятии решения о масштабном применении ИИ в лучевой диагностике для профилактических исследований важно

понимать метрики диагностической точности ИИ- алгоритма и врача-рентгенолога. Наше исследование показало, что в границах доверительных интервалов AUC данные метрики для "среднего" врача-рентгенолога и ИИ-алгоритмов пересекаются. Реализованный формат многопользовательской оценки, доступной для широкой аудитории профессионального сообщества, позволил получить усредненную оценку диагностической точности врача-рентгенолога на основании независимых данных большого количества специалистов различной квалификации и различного опыта, полученные в сходных условиях, имитирующих реальный рабочий процесс.

Использование бинарных критериев "с патологией" и "без патологии" позволило унифицировать критерии оценки для врачей и ИИ-алгоритмов и обеспечить их корректное сравнение по решению задачи выявления патологических изменений в целом, что, однако, не позволяет сравнить точность в решении определения конкретной патологии.

При сравнении AUC ИИ-алгоритмов между собой также отсутствуют статистически значимые различия, при этом, если сопоставить отдельно метрики чувствительности и специфичности, то прослеживается тенденция на преобладание одной или другой метрики при одинаковой настройке по максимизации порога Юдена. Так, алгоритм ИИ-1 достигал чувствительности 100 %, но имел достоверно более низкую специфичность.

При принятии решения о выборе для практического применения алгоритма ИИ-1 или ИИ-2 нужно оценивать конкретную решаемую задачу. Для задач скрининга необходимо обеспечить максимально приближенную к чувствительности 100 %, т.к. это позволит минимизировать пропуски патологии [17]. В качестве консультанта (системы поддержки принятия врачебного решения) для рентгенолога рекомендуется минимизировать количество ложных срабатываний, т.к. врач-рентгенолог по результатам многих исследований имеет более высокую специфичность по сравнению с чувствительностью [18, 19], соответственно, необходим более специфичный алгоритм ИИ.

Таким образом, при сравнении алгоритмов ИИ по значению метрик диагностической точности недостаточно уделять внимание только AUC, необходимо оценивать оптимальную чувствительность или специфичность в зависимости от специфики клинической задачи. В связи с этим необходимо осуществлять тонкую настройку ИИ-алгоритма.

Следует обратить внимание и на то, что при более детальном сравнительном анализе мы выявили статистически значимые различия между врачами и одним из алгоритмов ИИ, при этом в сравнении между собой алгоритмы ИИ не показали статистически значимых различий. Это объясняется не только различиями в средних значениях AUC, но и разбросом. Данное наблюдение можно использовать, например, в контроле качества работы ИИ-алгоритмов и сравнивать их не только между собой, но и с врачами-рентгенологами.

В нашем исследовании мы не ставили задачу разделения врачей-рентгенологов по стажу, так как при масштабном внедрении технологий на основе ИИ не предполагается дифференцирование его применения по стажу пользователя. Также при проведении настоящего исследования не оценивалась визуализация результатов работы ИИ-алгоритма и корректность формирования текстового описания исследования, что важно при внедрении технологий ИИ в медицинскую практику. [20, 21]

Важным ограничением исследования является анализ врачами и алгоритмами только одной проекции рентгенологических исследований, это обусловлено функциональными возможностями ИИ алгоритмов. Также на момент завершения работы над настоящей рукописью стало известно о выходе обновленных версий рассмотренных алгоритмов ИИ.

Выводы

Сравнение алгоритмов ИИ в условиях многопользовательской оценки сопоставимо с процессом валидации алгоритмов при мультицентровых исследованиях. Подобный формат исследований является доступным и воспроизводимым, может быть использован для оценки различных алгоритмов на различных наборах данных с привлечением специалистов различных областей в равных условиях. Результаты исследования указали на недостаточность использования таких метрик как AUC и диагностическая точность, необходима обязательная оценка чувствительности и специфичности. Также для практического внедрения алгоритмов ИИ требуется настройка оптимального порога срабатывания таких алгоритмов.

Благодарности

Авторы выражают благодарность заведующей учебным центром НПКЦ ДиТ ДЗМ Трофименко Ирине Анатольевне за помощь в организации и проведении исследования по коллективной оценке врачей.

Список литературы

1. Yu K-H, Lee T-LM, Yen M-H, et al. Reproducible Machine Learning Methods for Lung Cancer Detection Using Computed Tomography Images: Algorithm Development and Validation. J Med Internet Res. 2020; 22(8): e16709. https://doi.org/10.2196/16709.
2. Herron J, Reynolds JH. Trends in the on-call workload of radiologists. Clinical Radiology. 2006; 61(1), 91-6. https://doi.org/10.1016/j.crad.2005.07.008.
3. Harris M, Qi A, Jeagal L, et al. A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis. PLOS ONE. 2019; 14(9): e0221339. https://doi.org/10.1371/journal.pone.0221339.
4. Rodriguez-Ruiz A, Leng K, Gubern-Merida A,et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. JNCI: Journal of the National Cancer Institute. 2019; 111(9): 916-22. https://doi.org/10.1093/jnci/djy222.
5. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. JAMA Network Open. 2020; 3(3): e200265. https://doi.org/10.1001/jamanetworkopen.2020.0265.
6. Wu JT, Wong KCL, Gur Y, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Network Open. 2020; 3(10): e2022779. https://doi.org/10.1001/jamanetworkopen.2020.22779.
7. Morozov SP, Vladzymyrskyy AV, Klyashtornyy VG, et al. Clinical acceptance of software based on artificial intelligence technologies (radiology).
8. Gusev A, Vladzymyrskyy A. и др. Развитие исследований и разработок в сфере технологий искусственного интеллекта для здравоохранения в Российской Федерации: итоги 2021 года. Digital Diagnostics, 2022; 3(2). https://doi.org/10.17816/DD107367.
9. Sun, X., & Xu, W. (2014). Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters, 21(11), 1389-93. https://doi.org/10.1109/LSP.2014.2337313.
10. Hu X, Li C, Chen J, Qin G. Confidence intervals for the Youden index and its optimal cut-off point in the presence of covariates. J Biopharm Statistics. 2021; 31(3). https://doi.org/10.1080/10543406. 2020.1832107.
11. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3(1). https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.
12. Pauly M, Asendorf T, Konietschke F. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. Biometrical J. 2016; 58(6): 1319-37. https://doi.org/10.1002/ bimj.201500105.
13. Adams SJ, Henderson RDE, Yi X, Babyn P. Artificial Intelligence Solutions for Analysis of X-ray Images. Canadian Association of Radiologists Journal. 2021; 72(1), 60-72. https://doi.org/10.1177/0846537120941671.
14. Alexander A, Jiang A, Ferreira C, Zurkiya D. An Intelligent Future for Medical Imaging: A Market Outlook on Artificial Intelligence for Medical Imaging. J Am Coll Radiol, 2020; 17(1), 165-70. https://doi.org/10.1016/j.jacr.2019.07.019.
15. Majkowska A, Mittal S, Steiner DF, et al. Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. Radiology, 2020; 294(2): 421-31. https://doi.org/10.1148/radiol.2019191293,
16. Wu JT, Wong KCL, Gur Y, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Network Open, 2020; 3(10): e2022779. https://doi.org/10.1001/jamanetworkopen.2020.22779.
17. Thrall JH, Li X, Li Q, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. J Am Coll Radiol. 2018; 15(3 Pt B), 504-8. https://doi.org/10.1016/J.JACR.2017.12.026.

18. Hwang EJ, Nam JG, Lim WH, et al. Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. Radiology. 2019; 293(3): 573-80. https://doi.org/10.1148/RADIOL. 2019191225.
19. Yoo H, Lee SH, Arru CD, et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol. 2021; 31(12): 9664-74. https:// doi.org/10.1007/S00330-021-08074-7.
20. Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы. Использование сервисов на основе технологии искусственного интеллекта при проведении описаний рентгенологических снимков. Лучшие практики лучевой и инструментальной диагностики. 2020.
21. Морозов СП, Владзимирский АВ, Шулькин ИМ, и др. Целесообразность применения технологий искусственного интеллекта в лучевой диагностике (результаты первого года Московского эксперимента по компьютерному зрению). Врач и Информационные Технологии. 2022; 12–29.

CRITERIA FOR THE APPLICABILITY OF COMPUTER VISION FOR PREVENTIVE STUDIES ON THE EXAMPLE OF CHEST X-RAY AND FLUOROGRAPHY

K.M. Arzamasov[1], S.S. Semenov[1], D.Yu. Kokina[1], T.M. Bobrovskaya[1], N.A. Pavlov[1], Y.S. Kirpichev[1,2], A.E. Andreychenko[1,3,4], A.V. Vladzymyrskyy[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia

[2] LLC Medscan, Moscow, Russia

[3] Department of Physics and Engineering, ITMO University, Saint Petersburg, Russia

[4] LLC K-SkAI, Petrozavodsk, Russia

Purpose: In the conditions of a constant increase in the number of computer vision algorithms developed based on artificial intelligence (AI) for medical diagnostics, it becomes necessary to determine criteria for deciding whether their practical application for mass preventive studies of the population is appropriate.

Materials and methods: The study with the participation of several radiologists was conducted on a "Web platform for evaluating radiological studies" on a marked data set containing digital radiographs and fluorograms in an anterior direct projection. On the same data set, using the "Versioning Testing Platform", responses were obtained from two commercial AI-based computer vision algorithms developed for the analysis of digital radiographs. Evaluation of the results obtained from doctors and algorithms (binary, in terms of "with pathology" and "without pathology") was carried out using ROC analysis. For the threshold value calculated by the Yuden method, the following metrics were determined: sensitivity, specificity and accuracy.

Results: diagnostic accuracy metrics were calculated for the average assessment of radiologists and AI-based computer vision algorithms when searching for pathological changes on chest X-rays in anterior direct projection according to ROC analysis. The average values of diagnostic accuracy indicators of radiologists exceeded the indicators of AI services.

Conclusions: when deciding on the implementation of AI-based computer vision algorithms for preventive research, One should be guided by the metrics of diagnostic accuracy of a particular algorithm and use the average result of doctors in solving this diagnostic problem as the target values of metrics.

Key words: *artificial Intelligence, diagnostic, radiology, diagnostic accuracy, screening*

E-mail: yukirpichev@gmail.com