



Подходы к оценке точности измерений, выполненных ИИ-сервисами, в лучевой диагностике

Евгения Александровна Крылова^{1✉}, Юрий Александрович Васильев²,
Вера Владимировна Соболева³, Ирина Андреевна Разницына⁴,
Татьяна Михайловна Бобровская⁵, Кирилл Михайлович Арзамасов⁶

1, 2, 4–6 Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы, Москва, Россия

³ Российский национальный исследовательский медицинский университет им. Н. И. Пирогова, Москва, Россия

⁶ МИРЭА – Российский технологический университет, Москва, Россия

✉ KrylovaEA13@zdrav.mos.ru

Аннотация. Роль количественных измерений в лучевой диагностике возрастает, как и число внедряемых в область сервисов искусственного интеллекта (ИИ). Однако оценка точности их измерений остается недостаточно стандартизированной и опирается на статистические показатели, не всегда применимые для задач области. Целью исследования стали разработка и первичная апробация автоматизированного метода оценки качества количественных измерений, выполняемых ИИ-сервисами, на примере линейных измерений образований надпочечников на данных компьютерной томографии органов грудной клетки (КТ ОГК).

Набор данных содержал 90 КТ ОГК с разметкой: бинарный признак наличия образования и его измерение. Апробация была выполнена на основе результатов работы двух ИИ-сервисов. Предложенный подход включал оценку каждого измерения в зависимости от близости к эталону (0, 0,5 или 1 балл).

Общий балл рассчитывали как среднее; порогом достаточной точности было принято значение 0,81. Допустимые отклонения определялись эмпирически: 2 ($\pm 0,2$) и 4 ($\pm 0,4$) мм для границ корректного и частично корректного измерения. Общий балл ИИ-1 составил 0,914 (0,853–0,966), ИИ-2 – 0,913 (0,857–0,968) при ICC 0,94 и 0,93, коэффициенте Гвета AC2 0,97 и 0,96 соответственно; систематического смещения обнаружено не было. Разработанный подход обеспечил оценку отдельных измерений и общего качества работы ИИ-сервиса, и показал себя применимым для оценки ИИ-сервисов в задаче измерения образований надпочечников на КТ ОГК.

Ключевые слова и фразы: Искусственный интеллект, контроль качества, лучевая диагностика, статистический анализ

Благодарности: Данная статья подготовлена авторским коллективом в рамках НИР «Искусственный интеллект для одновременного анализа изображений и ЭМК (мультимодальный ИИ)», в соответствии с приказом от 27.11.2025 № 1213 «Об утверждении государственных заданий, финансовое обеспечение которых осуществляется за счет средств бюджета города Москвы, государственным бюджетным (автономным) учреждениям, подведомственным Департаменту здравоохранения города Москвы, на 2026 год и плановый период 2027 и 2028 годов» Департамента здравоохранения города Москвы

Для цитирования: Крылова Е. А., Васильев Ю. А., Соболева В. В., Разницына И. А., Бобровская Т. М., Арзамасов К. М. Подходы к оценке точности измерений, выполненных ИИ-сервисами, в лучевой диагностике // Программные системы: теория и приложения. 2026. Т. 17. № 3(72). С. 105–126. https://psta.psiras.ru/read/psta2026_3_105-126.pdf

Введение

Количественные измерения в лучевой диагностике все чаще встречаются в рутинной практике врача-рентгенолога [1, 2]. Фиксируемые на изображении, они отражают морфологическое и функциональное состояние тканей и анатомических структур. Рост встречаемости конкретных числовых значений в медицинской документации рентгенолога обусловлен совершенствованием лучевых методов, позволяющих фиксировать больше точных характеристик, увеличением частоты выполнения исследований сложных модальностей, таких как компьютерная (КТ) и магнитно-резонансная томография (МРТ) [1, 2, 4]. Это отражается в рекомендациях [5], международных профессиональных стандартах [6] и конкретных порогах, определяющих ведение пациента [7].

Однако измерения остаются одной из наиболее трудозатратных частей интерпретации лучевого исследования, увеличивая время описания и нагрузку на врачей-рентгенологов в условиях кадрового дефицита [8]. Кроме того, ручные измерения подвержены вариабельности, как между разными специалистами, так и между измерениями одного и того же врача в разные моменты времени [9, 10]. Это снижает воспроизводимость количественных признаков и их клиническую интерпретируемость.

Искусственный интеллект (ИИ) позволяет уменьшить влияние человеческого фактора и автоматизировать выполнение измерений [4]. Число решений на основе технологий искусственного интеллекта (ТИИ) с данным функционалом растет. Тем не менее, оценка качества числовых показателей, возвращаемых ИИ-сервисами, остается нестандартизированной областью.

Задача оценки точности измерений в медицине является широко распространенной, однако исторически чаще встречается в контексте лабораторных исследований, для которых количественные показатели сменили качественные еще до появления лучевых методов [11, 12]. Более того, для лабораторной диагностики свойственен прямой анализ объекта исследования, в то время как лучевые методы основываются на косвенной оценке посредством визуализации. Эти факторы ограничивают перенос существующих стандартов оценки точности измерений на область лучевой диагностики и рентгенологии в частности.

В настоящее время методы оценки того, насколько корректные значения возвращает ИИ-сервис, работающий с лучевыми изображениями, значительно отличаются от исследования к исследованию. Авторы, основываясь на различных гипотезах и источниках, реализуют собственные

методики контроля качества измерений, используя различные сочетания классических статистических показателей и метрик. Помимо этого, крайне разнообразны и критерии качества, основанных на интерпретации полученных значений этих метрик — одно и то же значение в разных работах может быть названо отличным и неудовлетворительным [13–15]. Отдельную сложность представляет оценка воспроизводимости таких методик — обоснованию применимости метода на конкретных данных часто не уделяется достаточное внимание, что может ставить под сомнение достоверность получаемого результата [16].

Более того, в условиях практического внедрения ИИ-сервисов в клиническую практику, необходимо не только характеризовать работу программного обеспечения (ПО) в общем, но и понимать качество каждого отдельного измерения [17]. Статистические характеристики дают лишь совокупную оценку, не позволяя выделить конкретные исследования, для которых сервис совершил ошибку, что, в свою очередь, дает разработчику возможность эффективно дорабатывать сервис, анализируя источники ошибок [18].

Таким образом, представляется актуальным создание подхода к оценке точности измерений в лучевой диагностике, который позволял бы не только получать общую оценку работы ИИ-сервиса, но и анализировать качество каждого отдельного измерения, что необходимо для корректного сравнения результатов работы различных сервисов и будущей автоматизации этапа тестирования и мониторинга ПО.

Целью настоящей работы стала разработка и первичная апробация автоматизированного метода оценки качества количественных измерений, выполняемых ИИ-сервисами, на примере линейных измерений образований надпочечников на КТ ОГК.

1. Материалы и методы

1.1. Дизайн исследования

Был разработан подход к оценке точности измерений, выполняемых ИИ-сервисами в лучевой диагностике, позволяющий оценивать каждое конкретное измерение. В ходе первичной апробации разработанного подхода, балльная оценка была вычислена для результатов работы двух ИИ-сервисов, выполняющих задачу поиска и измерения образований надпочечников на КТ ОГК. Полученные баллы были проанализированы относительно результатов классических тестов согласованности. Первичная апробация предлагаемого подхода проводилась с целью оценки его применимости и считалась успешной, если выполнялись следующие условия:

- (1) Результаты интерпретации баллов каждого ИИ-сервиса не противоречат статистическим метрикам.
- (2) Баллы и статистические метрики для ИИ-сервисов относительно друг друга показывают схожие тенденции.

1.2. Набор данных

На основе материалов Московского эксперимента был сформирован набор данных (НД), содержащий 90 КТ ОГК, выполненных с октября 2022 года по январь 2023 в медицинских организациях Москвы. Критериями включения были: возраст пациента старше 18 лет, наличие двух надпочечников в зоне сканирования. Критериями исключения стало наличие артефактов и дефектов укладки на исследовании, отсутствие результатов интерпретации от ИИ-сервисов. НД включал 31 исследование мужчин (34,4%), и 59 — женщин (65,6%); медиана возраста составила 69 лет (минимум 28 лет, максимум — 96 лет).

Для каждого исследования была получена экспертная разметка, выполненная 3 врачами-рентгенологами с опытом работы не менее 5 лет и включавшая в себя бинарную классификацию по признаку «наличие/отсутствие образования» для каждого надпочечника независимо, а также размер образования по короткой оси на аксиальном срезе. Аналогичные данные были получены от исследуемых ИИ-сервисов.

Настоящая работа является пилотным исследованием, и объем общей выборки был ограничен наибольшим доступным числом размеченных исследований.

1.3. Эталонные значения

В качестве эталонного значения для наличия образования выступило мнение двух экспертов с привлечением третьего в роли арбитра.

Каждое образование оценивалось независимо. Эталонным измерением считалось среднее от измерения двух (если один из экспертов не нашел образование) или трех врачей. Полученные значения, несмотря на использование в роли эталонных, в силу субъективности интерпретации изображений лучевой диагностики рассматривались не как истинная величина, а исключительно как наиболее возможное приближение к согласованной оценке, отражающей консенсусное мнение экспертов.

1.4. ИИ-сервисы

В качестве оцениваемых ИИ-сервисов выступили два ПО на основе ТИИ, участвующих в Московском эксперименте: АИ Диагностик КТ ОГК комплекс (ООО «Интел диагностик») и Chest-IRA (ООО «АЙРА Лабс»). Для обеспечения независимости анализа сервисы были анонимизированы и рандомизированы как ИИ-1 и ИИ-2.

1.5. Статистические методы

Для оценки согласованности разметки и качества выполняемых ИИ-сервисами измерений с точки зрения статистики использовались следующие методы:

- (1) MD (mean difference; средняя разность) — показатель, отражающий направленное среднее смещение одного ряда измерений относительно другого.
- (2) ICC (intraclass correlation coefficient; коэффициент внутриклассовой корреляции) модели ICC (2, 1) для абсолютного согласия.
- (3) Показатели регрессии Пассинга-Баблока, отражающие пропорциональный и систематический компоненты отличия рядов измерений (наклон и смещение соответственно).
- (4) Коэффициент Гвета AC2, позволяющий оценить согласованность двух рядов измерений, непрерывных по смыслу, но высокодискретных на практике, что приближает данные к категориальным.

Помимо этого, в качестве вспомогательных методов применялись вычисление среднего, доверительных интервалов (ДИ), медианы, стандартного отклонения, квартилей, поправка Бонферони на множественные гипотезы, тест Шапиро-Уилка и Брауна-Форсайта для проверки гипотез о характеристиках выборок, а также метод бутстрепа с возвращением (1000 повторений) для вычисления границ ДИ, значения которых указывались в скобках.

Для оценки достаточности объема выборки было рассмотрено значение минимально фиксируемой средней разности между эталоном и результатами ИИ-сервисов при $\alpha = 0,05$, $\beta = 0,20$ для двустороннего теста [19]. Минимальным необходимым фиксируемым значением MD было принято 0,5 мм как половина шага измерения. Минимальная выборка для оценки ИИ-сервисов рассчитывалась по данным стандартного отклонения разностей попарных измерений экспертов.

1.6. Программные методы

Для вычислений использовались библиотеки `pandas`, `numpy`, `scipy`, `scikit-bio`, `scikit-learn`, `statsmodels` языка Python 3.12, а также программный продукт MedCalc 23.4.4.^{[URL](#)}

2. Результаты

2.1. Описание разработанного подхода

Был разработан подход на основе балльной оценки отдельных измерений. Предлагаемый подход основывался на требовании оценки каждого конкретного измерения с возможностью получения совокупной метрики для результатов работы ИИ-сервиса.

Каждое измерение получало балл в зависимости от близости к эталонному. Было определено два значения отклонения от эталонного значения *REF* — малое *r*₁ и большое *r*₂. Соответствие баллов и интервалов попадания представлено в Таблице 1.

Таблица 1. Предлагаемые интервалы для оценки значений и баллы, соответствующие измерению в случае попадания в конкретный интервал

Значения	Измерение	Балл
$< (REF - r_2)$	Некорректное	0
$[(REF - r_2); (REF - r_1))$	Частично корректное	0,5
$[(REF - r_1); (REF + r_1)]$	Корректное	1
$((REF + r_1); (REF + r_2)]$	Частично корректное	0,5
$> (REF + r_2)$	Некорректное	0

Общий балл по данным определенного набора измеряемых объектов вычисляется как среднее арифметическое от оценок всех его измерений:

$$\text{Общий балл} = \frac{\text{Корректные} + 0,5 \times \text{Частично корректные}}{\text{Все}}$$

Порог достаточной точности метода измерений по данной балльной метрике был предварительно определен как 0,81 в соответствии с минимальной долей корректных измерений по нормативно-правовой документации Московского эксперимента [20]. Для вычисления значений *r*₁ и *r*₂ была использована полная совокупность абсолютных расхождений для всех возможных попарных экспертных измерений. *r*₁ определялось как 68-й перцентиль этой совокупности, *r*₂ — как 95-й. Для *r*₁ и *r*₂ при помощи бутстрапа рассчитывалось стандартное отклонение, что позволило оценить устойчивость выбранных порогов в рамках конкретной клинической задачи.

2.2. Оценка применимости предлагаемого подхода

По результатам экспертной разметки наличие образования было определено для 75 из 180 надпочечников (41,7%). Для измерений, сделанных экспертами, не было обнаружено статистически значимых отклонений от нормального, различий дисперсий также выявлено не было. Описательная статистика представлена в Таблице 2.

ТАБЛИЦА 2. Описательная статистика для измерений образований врачами

Разметчик	N	Me (Q_1 ; Q_3)
Врач 1	72	15 (12; 19)
Врач 2	65	17 (13; 21)
Врач 3	75	16 (12; 19)

Применимость разметки для оценки ИИ-сервисов была проанализирована посредством анализа согласованности экспертов-разметчиков друг с другом попарно и в совокупности (Таблица 3). Стандартное отклонение попарных разностей составило 1,3 мм, минимальный размер выборки для оценки ИИ-сервисов составил 53 измерения.

ТАБЛИЦА 3. Результаты оценки согласованности врачей друг с другом при измерении образований надпочечников

Показатель	Врач 1 & 2	Врач 1 & 3	Врач 2 & 3
Общее число измерений	60	71	64
MD [мм]	-1,1 (-1,7; -0,3)	-0,3 (-0,5; 0)	0,5 (0,2; 0,9)
ICC (попарное)	0,88 (0,75; 0,94)	0,98 (0,98; 0,99)	0,95 (0,91; 0,97)
ICC (совместное)	0,94 (0,89; 0,96)		
Коэффициент Гвета AC2	0,93 (0,87; 0,95)	0,99 (0,98; 0,99)	0,97 (0,95; 0,98)
Регрессия Пассинга-Баблока:			
смещение	-3,9 (-5,8; -1,0)	0,0 (-1,1; 0,0)	1,0 (0,0; 2,7)
наклон	1,2 (1,0; 1,3)	1,0 (1,0; 1,5)	0,9 (0,9; 1,0)

По совокупности полученных данных разметка была признана корректной, несмотря на ненулевое смещение для отдельных пар. Соответственно, все сделанные измерения были использованы при расчете ряда эталонных измерений, а также r_1 и r_2 . Значение последних составили соответственно $2 (\pm 0,2)$ и $4 (\pm 0,4)$ мм (Рисунок 1).

Число измерений для обоих ИИ-сервисов превышало минимальное необходимое. Баллы для ИИ-сервисов и результаты вычисления статистических коэффициентов представлены в Таблице 4.

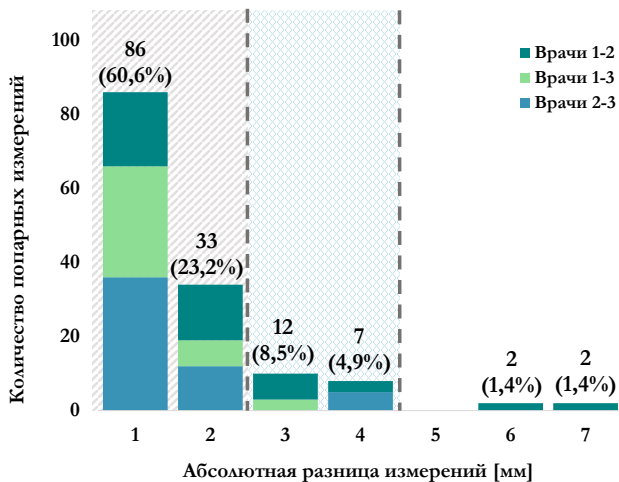


РИСУНОК 1. Гистограмма, отображающая количество парных измерений и абсолютные разницы. Серая область — значения, входящие в 68%, голубая — в 95%.

ТАБЛИЦА 4. Результаты оценки согласованности ИИ-сервисов с эталоном при измерении образований надпочечников и балльная оценка

Показатель	ИИ-1	ИИ-2
Общее число измерений	58	63
Корректных	50 (86,2%)	54 (85,7%)
Частично корректных	6 (10,3%)	7 (11,1%)
Некорректных	2 (3,5%)	2 (3,2%)
Общий балл	0,914 (0,853; 0,966)	0,913 (0,857; 0,968)
MD [мм]	0,2 (−0,3; 0,7)	−0,6 (−1,0; 0)
ICC	0,94 (0,90; 0,96)	0,93 (0,88; 0,96)
Коэффициент Гвета AC2	0,97 (0,95; 0,98)	0,96 (0,92; 0,97)
Регрессия Пассинга-Баблока:		
смещение	1,9 (0,0; 3,6)	0,0 (−0,1; 1,1)
наклон	0,9 (0,8; 1,0)	1,0 (0,9; 1,0)

По данным MD для обоих ИИ-сервисов не было обнаружено значимого смещения. Аналогичный результат был получен при построении регрессии

Пассинга-Баблока, не показавшей также и пропорциональной ошибки. Коэффициенты ICC и Гвета AC2 показали значения, соответствующие высокой согласованности результатов ИИ-сервисов и эталонных значений [16, 21].

Полученные результаты свидетельствовали о высокой точности измерений ИИ-сервисов. Баллы обоих ИИ-сервисов превышали предложенный порог 0,81. Значения рассчитанных коэффициентов для двух ИИ-сервисов были достаточно близки, при этом их доверительные интервалы перекрывались, что было верно и для общего балла предлагаемого подхода.

Таким образом, предложенный подход показал себя применимым в рамках задачи оценки точности измерения образований надпочечников и показывающим результат, согласующийся с классическими статистическими показателями.

3. Обсуждение

Подходы к оценке ИИ-сервисов, выполняющих количественные измерения, в лучевой диагностике и смежных биомедицинских областях основываются на ряде метрик, включающем в себя коэффициенты внутри-классовой корреляции (ICC) [9, 22] и конкордации Лина (CCC) [14, 15], график Блэнда–Альмана [13, 23], линейную регрессию [24], исследования корреляции [24], а также оценку абсолютных расхождений: среднюю разницу, среднее абсолютное расхождение MAE, среднее квадратичное расхождение MSE, корень из него RMSE [1, 23, 25]. Однако применение статистических коэффициентов для оценки качества измерений в лучевой диагностике связано с рядом сложностей.

Медицинские данные характеризуются распределениями, не всегда удовлетворяющими условиям нормальности и гомоскедастичности, что значительно влияет на ряд показателей, используемых для вывода о качестве инструмента измерения; в частности, это верно для ICC, CCC, показателей порогов согласия [26, 27]. Подобные методы зависят от того, для какой выборки рассчитывались, и исследования показывают, что сравнение значений, например, ICC, полученных на разных наборах данных, некорректно [28]. Это накладывает ограничения на возможность сравнения результатов ИИ-сервисов друг с другом и оценку динамики качества их работы во времени — что является одной из важных целей оценки.

Переход от числовых значений метрик к выводу о качестве ИИ-сервиса может быть затруднен — любые точные числовые значения, позволяющие отделить «высокую» согласованность от «низкой», являются эмпирическими и определяются в соответствии с тем, какой именно метод оценивается и какие распределения данных характерны для области [29]. При этом в области лучевой диагностики такие пороги до сих пор не общеприняты. Более того, разные клинические задачи, вероятно, требуют разной точности выполнения измерений, что также препятствует прямой интерпретации статистических коэффициентов [30].

Балльная оценка каждого отдельного измерения в данных условиях остается легко интерпретируемой — как разработчиками, так и врачами.

Использование среднего в качестве эталонного значения является распространенным методом объединения нескольких повторных измерений одного и того же объекта [31, 32]. Именно данный метод соответствует ГОСТ, описывающему порядок многократных измерений [33]. Аналогичный подход широко используется в области рентгенологии, в том числе для последующей оценки автоматизированных методов измерения и моделей ИИ [13, 34, 35]. При этом среднее объединяет всю имеющуюся информацию разметки, трактуя измерения как равновесные, не меняет единицы измерения, уменьшает влияние случайной ошибки, а сам подход остается легко интерпретируемым и понятным [30].

Значения баллов для измерений разной степени корректности основываются на методических рекомендациях по оценке сервисов искусственного интеллекта в задачах рентгенологии [17]. Необходимо подчеркнуть, что настоящие величины баллов могут быть изменены в зависимости от клинической задачи выполняющихся измерений, однако в рамках общего описания подхода и применительно к измерениям образований надпочечников завышение и занижение размеров имеют одинаковый штраф в баллах. Аналогично может быть откорректировано и пороговое значение в зависимости от предполагаемой области применения, целевой аудиторией и пациентской когорты.

Для задачи оценки качества измерения размеров анатомических структур в рентгенологии вывод о неправильности значения при наличии любого по величине расхождения с эталоном был бы некорректным, учитывая, что формирование самого эталона опирается на ручные измерения и является условным. Следовательно, предполагается необходимым определить интервал, попадание в который является допустимым для каждого конкретного значения. Свойственное области небольшое

количество врачей-разметчиков [36] и вероятные ошибки экспертов не позволяют считать ряд измерений каждого объекта исчерпывающим. Таким образом, допустимый интервал основывался не на разбросе значений для конкретного объекта, а на общей совокупности абсолютных попарных расхождений. Это позволяет выявить, насколько в конкретной клинической задаче врачи отклоняются друг от друга при выполнении измерения в условиях целенаправленной разметки, и перенести аналогичные тенденции на оценку ИИ-сервисов.

Определение конкретных значений допустимых отклонений от эталона сильно зависит от многих факторов, в частности от фактического размаха размера, точности измерения, критичности выявляемого числового значения для маршрутизации пациента и т. д. [28, 30]. Образования надпочечников измерялись с точностью до целых и, учитывая, что все экспертные измерения принимались как корректные и не включали значимых выбросов, предлагаются эмпирические значения 68-го и 95-го перцентилей. При этом стандартное отклонение получившихся значений не превышает 10%, что позволило подтвердить применимость таких порогов в текущей задаче.

Рассчитанные таким образом баллы для ИИ-сервисов показали близкие и высокие абсолютные значения, однако при этом балл ИИ-1 был выше, что соответствовало результатам рассчитанных статистических критериев, для которых ИИ-1 показывал незначимо более высокие результаты. При этом интерпретация сервисов как достаточно или недостаточно хорошо справляющихся с задачей измерения совпадала для всех значений — оба сервиса показали высокие абсолютные результаты и, более того, согласовывались с эталоном не хуже, чем врачи друг с другом.

Таким образом, настоящая работа предлагает легко реализуемый и интерпретируемый подход, который позволяет как оценивать каждое измерение, выполненное ИИ-сервисом, так и получать общую единую метрику, отражающую близость автоматизированных значений к эталонным. Подход показал себя применимым в задаче определения размеров образований надпочечников на КТ ОГК, что позволило сделать вывод о практической применимости, адекватности возвращаемых результатов и предложенных порогов.

4. Ограничения

Результаты и выводы настоящей работы ограничены объемом набора данных, определявшимся доступными исследованиями, и числом врачей, выполнявших разметку — при больших данных значения допустимых отклонений могут быть уточнены. Использование подхода для любой иной задачи требует дополнительной направленной апробации, так как задачи клинической медицины могут включать в себя требования и условия, которые необходимо учитывать как при подготовке эталонного набора данных, так и при определении пороговых значений. Кроме того, предлагаемый подход не учитывает бинарную точность выявления патологии, выполняемого ИИ-сервисами и включает рассмотрение только истинно-положительных находок.

Заключение

В настоящем исследовании был представлен подход к оценке измерений ИИ-сервисов в лучевой диагностике, который позволяет оценивать измерения на основании их близости к эталонным значениям с помощью балльной оценки. Описанный в общем виде, подход был реализован на данных КТ ОГК с задачей определения размеров образования надпочечников, что позволило сделать вывод о его применимости в условиях реальных данных.

Согласие этического комитета

Исследование основывается на данных, полученных в ходе Московского эксперимента (выписка из протокола №2 НЭК МРО РОПР от 20 февраля 2020 года).

Список использованных источников

- [1] Barat M., Jannot A. S., Dohan A., Soyer P. *How to report and compare quantitative variables in a radiology article* // Diagnostic and Interventional Imaging.— 2022.— Vol. **103**.— No. 12.— Pp. 571–573.  [↑106, 113](#)
- [2] Marek J., Bachurska D., Wolak T., Borowiec A., Sajdek M., Maj E. *Quantitative brain volumetry in neurological disorders: from disease mechanisms to software solutions* // Polish Journal of Radiology.— 2025.— Vol. **90**.— Pp. e299–e306.  [↑106](#)

- [3] Weingärtner S., Desmond K. L., Obuchowski N. A., Baessler B., Zhang Y., Biondetti E., Ma D., Golay X., Boss M. A., Gunter J. L., Keenan K. E., Hernando D., ISMRM Quantitative MR Study Group *Development, validation, qualification, and dissemination of quantitative MR methods: Overview and recommendations by the ISMRM quantitative MR study group* // Magnetic Resonance in Medicine.– 2022.– Vol. **87**.– No. 3.– Pp. 1184–1206.  ↑
- [4] Шихмурадов Д. У., Васильев Ю. А., Владзимирский А. В., Омелянская О. В., Арзамасов К. М., Доможирова А. С., Пестренин Л. Д., Бобровская Т. М. *Сравнительная оценка диагностической точности врача-рентгенолога и программного обеспечения на основе технологий искусственного интеллекта в выявлении инциденталом надпочечников при компьютерной томографии* // Вестник Российского научного центра рентгенорадиологии.– 2026.– Т. **26**.– № 1.– С. 25–40.   ↑¹⁰⁶
- [5] Гомболевский В. А., Харламов К. А., Масри А. Г., Ким С. Ю., Морозов С. П. *Общие рекомендации по описанию первичных и повторных КТ, МРТ, рентгенологических исследований, Лучшие практики лучевой и инструментальной диагностики*.– Т. **2**.– М.: ГБУЗ «НПКЦ ДиТ ДЗМ».– 2017.– 19 с.  ↑¹⁰⁶
- [6] Fassnacht M., Tsagarakis S., Terzolo M., Tabarin A., Sahdev A., Newell-Price J., Pelsma I., Marina L., Lorenz K., Bancos I., Arlt W., Dekkers O. *European Society of Endocrinology clinical practice guidelines on the management of adrenal incidentalomas, in collaboration with the European Network for the Study of Adrenal Tumors* // European Journal of Endocrinology.– 2023.– Vol. **189**.– No. 1.– Pp. G1–G42.  ↑¹⁰⁶
- [7] Eisenhauer E. A., Therasse P., Bogaerts J., Schwartz L. H., Sargent D., Ford R., Dancey J., Arbuck S., Gwyther S., Mooney M., Rubinstein L., Shankar L., Dodd L., Kaplan R., Lacombe D., Verweij J. *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)* // European Journal of Cancer.– 2009.– Vol. **45**.– No. 2.– Pp. 228–247.  ↑¹⁰⁶
- [8] *Искусственный интеллект в лучевой диагностике: Per Aspera Ad Astra*/ ред. Васильев Ю. А., Владзимирский А. В.– М.: Издательские решения.– 2025.– ISBN 978-5-0067-5622-9.– 491 с.  ↑¹⁰⁶
- [9] Han K., Ryu L. *Statistical methods for the analysis of inter-reader agreement among three or more readers* // Korean Journal of Radiology.– 2024.– Vol. **25**.– No. 4.– Pp. 325–327.  ↑^{106, 113}
- [10] Quinn L., Tryposkiadis K., Deeks J., De Vet H. C. W., Mallett S., Mokkink L. B., Takwoingi Y., Taylor-Phillips S., Sitch A. *Interobserver variability studies in diagnostic imaging: a methodological systematic review* // The British journal of radiology.– 2023.– Vol. **96**.– No. 1148.– id. 20220972.  ↑¹⁰⁶
- [11] McBride G. B. *A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficients*, NIWA Client Rep.: HAM2005–062.– 2005.– 18 pp. ↑¹⁰⁶
- [12] Lin L., Hedayat A. S., Wu W. *Statistical Tools for Measuring Agreement*.– New York: Springer.– 2012.– ISBN 978-1-4614-0561-0.– 161 pp.  ↑¹⁰⁶

- [13] Iuga A. I., Carolus H., Caldeira L. L., Kottlors J., Rinneburger M., Weisthoff M., Fervers P., Rauen P., Ficht F., Goertz L., Niederau P., Siedek F., Heneweier C., Gietzen C., Pennig L., Dobrostal A., Laqua F., Woznicki P., Maintz D., Baessler B., Persigehl T. *Fully automated AI-based lymph node measurements in chest CT: Accuracy and reproducibility compared with multi-reader assessment* // *Diagnostics* (Basel).– 2026.– Vol. **16**.– No. 7.– id. 967.– 14 pp.  ↑_{107, 113, 114}
- [14] Watson T., Calder A., Barber J. L. *Quantitative bowel apparent diffusion coefficient measurements in children with inflammatory bowel disease are not reproducible* // *Clinical Radiology*.– 2018.– Vol. **73**.– No. 6.– Pp. 574–579.  ↑_{107, 113}
- [15] Taye M. R., Abdel-Aziz M. A., Hameed D. A., Behnsawy H. M., Moeen A. M., Samy M. *External validation of automated CT volumetry in renal tumor volume measurement and its surgical implications* // *Kidney Cancer*.– 2025.– Vol. **9**.– id. 9.  ↑_{107, 113}
- [16] Koo T. K., Li M. Y. *A guideline of selecting and reporting intraclass correlation coefficients for reliability research* // *Journal of Chiropractic Medicine*.– 2016.– Vol. **15**.– No. 2.– Pp. 155–163.  ↑_{107, 113}
- [17] Васильев Ю. А., Владимирский А. В., Омелянская О. В., Шулькин И. М., Арзамасов К. М., Никитин Н. Ю., Пестренин Л. Д., Шарова Д. Е. *Оценка зрелости технологий искусственного интеллекта для здравоохранения: методические рекомендации, Лучшие практики лучевой и инструментальной диагностики*.– Т. **123**.– М.: ГБУЗ «НПКЦ ДиТ ДЗМ».– 2023.– 28 с.  ↑_{107, 114}
- [18] Васильев Ю. А., Владимирский А. В., Шарова Д. Е., Зинченко В. В., Заюнчковский С. Ю., Ахмад Е. С., Гусев А. В., Арзамасов К. М. *Клинические испытания систем искусственного интеллекта (мучевая диагностика), Лучшие практики лучевой и инструментальной диагностики*.– Т. **113**.– М.: ГБУЗ «НПКЦ ДиТ ДЗМ».– 2023.– 40 с.  ↑₁₀₇
- [19] Ryan T. P. *Sample Size Determination and Power*, Wiley series in probability and statistics.– Hoboken, N.J.: John Wiley & Sons, Inc.– 2013.– ISBN 978-1118437605.– 378 pp.  ↑₁₀₉
- [20] Приказ Департамента здравоохранения города Москвы от 09.02.2024 № 89 «Об утверждении Порядка и условий дальнейшего проведения эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения этих технологий в системе здравоохранения».– 2024 (24.05.2026).  ↑₁₁₀
- [21] Landis J. R., Koch G. G. *The measurement of observer agreement for categorical data* // *Biometrics*.– 1977.– Vol. **33**.– No. 1.– Pp. 159–174.  ↑₁₁₃
- [22] Kim T. M., Lee Y., Seo J. Y., Cho J. Y., Kim S. Y., Kim Y. G. *Automated deep learning for detection and measurement of adrenal masses in contrast-enhanced abdominal CT* // *European Radiology*.– 2026.– Vol. **36**.– No. 6.– Pp. 4527–4538. 

- [23] Васильев Ю. А., Блохин И. А., Коденко М. Р., Сучилова М. М., Владимирский А. В., Омелянская О. В., Решетников Р. В. *Оценка надежности измерений легочных узлов при компьютерной томографии: ретроспективное исследование* // Медицинский совет.– 2026.– Т. 19.– № 21.– С. 123–129.   ^{↑113}
- [24] O'Shaughnessy E., Detrinidad E., Soyer P., Lecler A. *An introductory guide to statistics for the radiologist* // Diagnostic and interventional imaging.– 2025.– Vol. 106.– No. 2.– Pp. 49–52.  ^{↑113}
- [25] ГОСТ Р 59898-2021. *Оценка качества систем искусственного интеллекта. Общие положения.*– М.: Российский институт стандартизации.– 2022.– 26 с. ^{↑113}
- [26] Costa-Santos C., Bernardes J., Ayres-de-Campos D., Costa A., Amorim-Costa C. *The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement* // Journal of clinical epidemiology.– 2011.– Vol. 64.– No. 3.– Pp. 264–269.  ^{↑113}
- [27] Taffé P., Zuppingen C., Burger G. M., Nusslé S. G. *The Bland-Altman method should not be used when one of the two measurement methods has negligible measurement errors* // PloS One.– 2022.– Vol. 17.– No. 12.– id. e0278915.  ^{↑113}
- [28] Mehta S., Bastero-Caballero R. F., Sun Y., Zhu R., Murphy D. K., Hardas B., Koch G. *Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies* // Statistics in Medicine.– 2018.– Vol. 37.– No. 18.– Pp. 2734–2752.  ^{↑113, 115}
- [29] Wadoux A. M. J.-C., Minasny B. *Some limitations of the concordance correlation coefficient to characterise model accuracy* // Ecological Informatics.– 2024.– Vol. 83.– id. 102820.– 4 pp.  ^{↑114}
- [30] Anvari A., Halpern E. F., Samir A. E. *Essentials of statistical methods for assessing reliability and agreement in quantitative imaging* // Academic Radiology.– 2018.– Vol. 25.– No. 3.– Pp. 391–396.  ^{↑114, 115}
- [31] Park J. E., Han K., Sung Y. S., Chung M. S., Koo H. J., Yoon H. M., Choi Y. J., Lee S. S., Kim K. W., Shin Y., An S., Cho H. M., Park S. H. *Selection and reporting of statistical methods to assess reliability of a diagnostic test: conformity to recommended methods in a peer-reviewed journal* // Korean Journal of Radiology.– 2017.– Vol. 18.– No. 6.– Pp. 888–897.  ^{↑114}
- [32] Iannessi A., Beaumont H. *Breaking down the RECIST 1.1 double read variability in lung trials: What do baseline assessments tell us?*Frontiers in Oncology.– 2023.– Vol. 13.– id. 988784.– 14 pp.  ^{↑114}
- [33] ГОСТ Р 8.736-2011. *Измерения прямые многократные. Методы обработки результатов измерений. Основные положения.*– М.: Российский институт стандартизации.– 2013.– 23 с. ^{↑114}
- [34] Yao H., Zhang Z., Cheng G., Chen X., He L., Wang W., Zhou S., Wang P. *Automatic measurement of anatomical parameters of the lumbar vertebral body and the intervertebral disc on radiographs by deep learning* // Quantitative Imaging in Medicine and Surgery.– 2024.– Vol. 14.– No. 8.– Pp. 5877–5890.  ^{↑114}

- [35] Carabetta N., Panuccio G., Giordano S., Sorrentino S., Mazza G. A., Sabatino J., Canino G., Leo I., Salerno N., Strangio A., Petullà M., Torella D., De Rosa S. *Reliability of immersive virtual reality for pre-procedural planning for TAVI: A CT-based validation* // Journal of Cardiovascular Development and Disease.– 2025.– Vol. **12**.– No. 12.– id. 481.– 14 pp. doi ↑114
- [36] Nagendran M., Chen Y., Lovejoy C. A., Gordon A. C., Komorowski M., Harvey H., Topol E. J., Ioannidis J. P. A., Collins G. S., Maruthappu M. *Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies* // BMJ.– 2020.– Vol. **368**.– id. m689.– 12 pp. doi ↑115

Поступила в редакцию 16.07.2026;
одобрена после рецензирования 16.08.2026;
принята к публикации 09.09.2026;
опубликована онлайн 12.09.2026.

Рекомендовал к публикации

к.т.н. Гулиев Ядулла Иман оглы

Информация об авторах:



Евгения Александровна Крылова

Техник отдела медицинской информатики, радиомики и радиогенетики ГБУЗ «НПКЦ ДиТ ДЗМ».

Научные интересы: системы поддержки принятия врачебных решений, информатизация здравоохранения, применение статистических методов в биомедицинских задачах.



0009-0008-0421-3448

e-mail: KrylovaEA13@zdrav.mos.ru



Юрий Александрович Васильев

д.м.н, доцент, директор ГБУЗ «НПКЦ ДиТ ДЗМ». Главный внештатный специалист по лучевой и инструментальной диагностике Департамента здравоохранения Москвы, член правления Московского регионального отделения Российского общества рентгенологов и радиологов.

Научные интересы: внедрение технологий искусственного интеллекта в здравоохранение, цифровизация инструментальной диагностики, биомедицинская техника и физика.



0000-0002-5283-5961

e-mail: VasilevYA1@zdrav.mos.ru

**Вера Владимировна Соболева**

к.б.н., доцент, кафедра медицинской кибернетики и информатики им. С.А. Гаспаряна МБФ РНИМУ им. Н.И. Пирогова.

Научные интересы: математическое моделирование, системный анализ, методы оптимизации, методы поддержки принятия решений.



0009-0006-5089-7065

e-mail: sobolevav2010@yandex.ru

**Ирина Андреевна Разницына**

к.ф.-м.н, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ».

Научные интересы: технологии искусственного интеллекта в здравоохранении, медицинская физика, системы оценки качества.



0000-0003-4145-6947

e-mail: RaznitsynaIA@zdrav.mos.ru

**Татьяна Михайловна Бобровская**

к.м.н, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ».

Научные интересы: методологическое обоснование применения инновационных технологий в области здравоохранения, порядок формирования медицинских наборов данных.



0000-0002-2746-7554

e-mail: BobrovskayaTM@zdrav.mos.ru

**Кирилл Михайлович Арзамасов**

д.м.н., руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ».

Научные интересы: системы поддержки принятия врачебных решений, анализ медицинских изображений, телемедицинские диагностические системы.



0000-0001-7786-0349

e-mail: ArzamasovKM@zdrav.mos.ru

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: благополучие авторов не зависит от результатов исследования.



Approaches to Assessing the Accuracy of Measurements Performed by AI-Services in Radiology

Evgeniya Alexandrovna **Krylova**¹, Yuriy Alexandrovich **Vasilev**²,
Vera Vladimirovna **Soboleva**³, Irina Andreevna **Raznitsyna**⁴,
Tatiana Mikhailovna **Bobrovskaya**⁵, Kirill Mikhailovich **Arzamasov**⁶

^{1,2,4-6} Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Department of Health, Moscow, Russia

³ Pirogov Russian National Research Medical University, Moscow, Russia

⁶ MIREA - Russian Technological University, Moscow, Russia

[✉] KrylovaEA13@zdrav.mos.ru

Abstract. The importance of quantitative measurements in radiology is increasing, as is the number of artificial intelligence (AI) services being introduced into the field. However, assessment of their measurement accuracy remains insufficiently standardized and relies on statistical metrics that are not always applicable to domain-specific tasks. This study aimed to develop and conduct a preliminary evaluation of an automated method for assessing the quality of quantitative measurements performed by AI-services in radiology, using linear measurements of adrenal lesions on chest CT as an example.

The dataset comprised 90 chest CT examinations with annotations including a binary indicator of lesion presence and its measurement. The approach was evaluated using the outputs of two AI-services. Each measurement was scored according to its proximity to the reference standard (0, 0.5, or 1 point). The overall score was calculated as the mean score, with 0.81 adopted as the threshold for sufficient accuracy. Acceptable deviations were determined empirically: 2 (± 0.2) and 4 (± 0.4) mm defined the boundaries for correct and partially correct measurements, respectively.

The overall score was 0.914 (0.853–0.966) for AI-1 and 0.913 (0.857–0.968) for AI-2, while the intraclass correlation coefficient was 0.94 and 0.93, Gwet's AC2 was 0.97 and 0.96, respectively. No systematic bias was identified. The proposed approach enabled assessment of individual measurements as well as the overall performance quality of an AI-service and proved applicable for the task of measuring adrenal lesions on chest CT. (*In Russian*).

Key words and phrases: Artificial intelligence, quality control, radiology, statistical analysis

2020 Mathematics Subject Classification: 68U10; 62P10, 92C50

Acknowledgments: This paper was prepared by a group of authors as a part of the research and development effort titled "Artificial intelligence for synchronous evaluation of images and EHR (multimodal AI)", in accordance with the Order No. 1213 dated November 27, 2025 "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 2026 and the planned period of 2027 and 2028" issued by the Moscow Health Care Department.

For citation: Evgeniya A. Krylova, Yuriy A. Vasilev, Vera V. Soboleva, Irina A. Raznitsyna, Tatiana M. Bobrovskaya, Kirill M. Arzamasov. *Approaches to Assessing the Accuracy of Measurements Performed by AI-Services in Radiology*. Program Systems: Theory and Applications, 2026, **17**:3(72), pp. 105–126. (*In Russ.*).
https://psta.psiras.ru/read/psta2026_3_105-126.pdf

References

- [1] M. Barat, A. S. Jannot, A. Dohan, P. Soyer. “How to report and compare quantitative variables in a radiology article”, *Diagnostic and Interventional Imaging*, **103**:12 (2022), pp. 571–573.
- [2] J. Marek, D. Bachurska, T. Wolak, A. Borowiec, M. Sajdek, E. Maj. “Quantitative brain volumetry in neurological disorders: from disease mechanisms to software solutions”, *Polish Journal of Radiology*, **90** (2025), pp. e299–e306.
- [3] S. Weingärtner, K. L. Desmond, N. A. Obuchowski, B. Baessler, Y. Zhang, E. Biondetti, D. Ma, X. Golay, M. A. Boss, J. L. Gunter, K. E. Keenan, D. Hernando, Quantitative MR Study Group ISMRM. “Development, validation, qualification, and dissemination of quantitative MR methods: Overview and recommendations by the ISMRM quantitative MR study group”, *Magnetic Resonance in Medicine*, **87**:3 (2022), pp. 1184–1206.
- [4] D. U. Shixmuradov, Yu. A. Vasil’ev, A. V. Vladzimirskij, O. V. Omelyanskaya, K. M. Arzamasov, A. S. Domozhirova, L. D. Pestrenin, T. M. Bobrovskaya. “Comparative evaluation of the diagnostic accuracy of a radiologist and artificial intelligence-based software in the detection of adrenal incidentalomas on computed tomography”, *Vestnik Rossijskogo nauchnogo centra rentgenoradiologii*, **26**:1 (2026), pp. 25–40.
- [5] V. A. Gombolevskij, K. A. Xarlamov, A. G. Masri, S. Yu. Kim, S. P. Morozov. *General recommendations for reporting initial and follow-up CT, MRI, and X-ray examinations*, Luchshie praktiki luchevoj i instrumental’noj diagnostiki, vol. **2**, GBUZ «NPKC DiT DZM», M., 2017, 19 pp. (in Russian).
- [6] M. Fassnacht, S. Tsagarakis, M. Terzolo, A. Tabarin, A. Sahdev, J. Newell-Price, I. Pelsma, L. Marina, K. Lorenz, I. Bancos W. Arlt, O. Dekkers. “European Society of Endocrinology clinical practice guidelines on the management of adrenal incidentalomas, in collaboration with the European Network for the Study of Adrenal Tumors”, *European Journal of Endocrinology*, **189**:1 (2023), pp. G1–G42.
- [7] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij. “New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)”, *European Journal of Cancer*, **45**:2 (2009), pp. 228–247.
- [8] *Artificial Intelligence in Diagnostic Imaging: Per Aspera Ad Astra*, ed. Vasil’ev Yu. A., Vladzimirskij A. V., Izdatel’skie resheniya, M., 2025, ISBN 978-5-0067-5622-9, 491 pp. (in Russian).
- [9] K. Han, L. Ryu. “Statistical methods for the analysis of inter-reader agreement among three or more readers”, *Korean Journal of Radiology*, **25**:4 (2024), pp. 325–327.

- [10] L. Quinn, K. Tryposkiadis, J. Deeks, Vet H. C. W. De, S. Mallett, L. B. Mookink, Y. Takwoingi, S. Taylor-Phillips, A. Sitch. “Interobserver variability studies in diagnostic imaging: a methodological systematic review”, *The British journal of radiology*, **96**:1148 (2023), id. 20220972. 
- [11] G. B. McBride. *A proposal for strength-of-agreement criteria for Lin’s concordance correlation coefficients*, NIWA Client Rep.: HAM2005–062, 2005, 18 pp.
- [12] L. Lin, A. S. Hedayat, W. Wu. *Statistical Tools for Measuring Agreement*, Springer, New York, 2012, ISBN 978-1-4614-0561-0, 161 pp. 
- [13] A. I. Iuga, H. Carolus, L. L. Caldeira, J. Kottlors, M. Rinneburger, M. Weisthoff, P. Fervers, P. Rauen, F. Fichter, L. Goertz, P. Niederau, F. Siedek, C. Heneweier, C. Gietzen, L. Pennig, A. Dobrostal, F. Laqua, P. Woznicki, D. Maintz, B. Baessler, T. Persigehl. “Fully automated AI-based lymph node measurements in chest CT: Accuracy and reproducibility compared with multi-reader assessment”, *Diagnostics (Basel)*, **16**:7 (2026), id. 967, 14 pp. 
- [14] T. Watson, A. Calder, J. L. Barber. “Quantitative bowel apparent diffusion coefficient measurements in children with inflammatory bowel disease are not reproducible”, *Clinical Radiology*, **73**:6 (2018), pp. 574–579. 
- [15] M. R. Taye, M. A. Abdel-Aziz, D. A. Hameed, H. M. Behnsawy, A. M. Moeen, M. Samy. “External validation of automated CT volumetry in renal tumor volume measurement and its surgical implications”, *Kidney Cancer*, **9** (2025), id. 9. 
- [16] T. K. Koo, M. Y. Li. “A guideline of selecting and reporting intraclass correlation coefficients for reliability research”, *Journal of Chiropractic Medicine*, **15**:2 (2016), pp. 155–163. 
- [17] Yu. A. Vasil’ev, A. V. Vladzimirskij, O. V. Omelyanskaya, I. M. Shul’kin, K. M. Arzamasov, N. Yu. Nikitin, L. D. Pestrenin, D. E. Sharova. *Assessment of Artificial Intelligence Technology Maturity for Healthcare: Methodological Recommendations*, Luchshie praktiki luchevoj i instrumental’noj diagnostiki, vol. **123**, GBUZ “NPKC DiT DZM”, M., 2023, 28 pp. (in Russian).
- [18] Yu. A. Vasil’ev, A. V. Vladzimirskij, D. E. Sharova, V. V. Zinchenko, S. Yu. Zayunchkovskij, E. S. Axmad, A. V. Gusev, K. M. Arzamasov. *Clinical Trials of Artificial Intelligence Systems (Diagnostic Imaging)*, Luchshie praktiki luchevoj i instrumental’noj diagnostiki, vol. **113**, GBUZ “NPKC DiT DZM”, M., 2023, 40 pp. (in Russian).
- [19] T. P. Ryan. *Sample Size Determination and Power*, Wiley series in probability and statistics, John Wiley & Sons, Inc., Hoboken, N.J., 2013, ISBN 978-1118437605, 378 pp. 
- [20] *Prikaz Departamenta zdravooxraneniya goroda Moskvy ot 09.02.2024 No 89 “Ob utverzhdenii Poryadka i uslovij dal’nejshego provedeniya eksperimenta po ispol’zovaniyu innovacionnyx tehnologij v oblasti komp’yuternogo zreniya dlya analiza medicinskih izobrazhenij i dal’nejshego primeneniya etix tehnologij v sisteme zdravooxraneniya”*, 2024 (24.05.2026). 

- [21] J. R. Landis, G. G. Koch. “The measurement of observer agreement for categorical data”, *Biometrics*, **33**:1 (1977), pp. 159–174. [doi](#)
- [22] T. M. Kim, Y. Lee, J. Y. Seo, J. Y. Cho, S. Y. Kim, Y. G. Kim. “Automated deep learning for detection and measurement of adrenal masses in contrast-enhanced abdominal CT”, *European Radiology*, **36**:6 (2026), pp. 4527–4538. [doi](#)
- [23] Yu. A. Vasil’ev, I. A. Bloxin, M. R. Kodenko, M. M. Suchilova, A. V. Vladzimirskij, O. V. Omelyanskaya, R. V. Reshetnikov. “Lung nodule measurement reliability in computed tomography: A retrospective study”, *Medicinskij sovet*, **19**:21 (2026), pp. 123–129 (in Russian). [doi](#)
- [24] E. O’Shaughnessy, E. Detrinidad, P. Soyer, A. Lecler. “An introductory guide to statistics for the radiologist”, *Diagnostic and interventional imaging*, **106**:2 (2025), pp. 49–52. [doi](#)
- [25] *GOST R 59898-2021. Ocenka kachestva sistem iskusstvennogo intellekta. Obshhie polozheniya*, Rossijskij institut standartizacii, M., 2022, 26 pp. (in Russian).
- [26] C. Costa-Santos, J. Bernardes, D. Ayres-de-Campos, A. Costa, C. Amorim-Costa. “The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement”, *Journal of clinical epidemiology*, **64**:3 (2011), pp. 264–269. [doi](#)
- [27] P. Taffé, C. Zuppinger, G. M. Burger, S. G. Nusslé. “The Bland-Altman method should not be used when one of the two measurement methods has negligible measurement errors”, *PloS One*, **17**:12 (2022), id. e0278915. [doi](#)
- [28] S. Mehta, R. F. Bastero-Caballero, Y. Sun, R. Zhu, D. K. Murphy, B. Hardas, G. Koch. “Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies”, *Statistics in Medicine*, **37**:18 (2018), pp. 2734–2752. [doi](#)
- [29] A. M. J.-C. Wadoux, B. Minasny. “Some limitations of the concordance correlation coefficient to characterise model accuracy”, *Ecological Informatics*, **83** (2024), id. 102820, 4 pp. [doi](#)
- [30] A. Anvari, E. F. Halpern, A. E. Samir. “Essentials of statistical methods for assessing reliability and agreement in quantitative imaging”, *Academic Radiology*, **25**:3 (2018), pp. 391–396. [doi](#)
- [31] J. E. Park, K. Han, Y. S. Sung, M. S. Chung, H. J. Koo, H. M. Yoon, Y. J. Choi, S. S. Lee, K. W. Kim, Y. Shin, S. An, H. M. Cho, S. H. Park. “Selection and reporting of statistical methods to assess reliability of a diagnostic test: conformity to recommended methods in a peer-reviewed journal”, *Korean Journal of Radiology*, **18**:6 (2017), pp. 888–897. [doi](#)
- [32] A. Iannessi, H. Beaumont. Breaking down the RECIST 1.1 double read variability in lung trials: What do baseline assessments tell us? *Frontiers in Oncology*, **13** (2023), id. 988784, 14 pp. [doi](#)

- [33] *GOST R 8.736-2011. Izmereniya pryamyie mnogokratnye. Metody obrabotki rezul'tatov izmerenij. Osnovnye polozeniya*, Rossijskij institut standartizacii, M., 2013, 23 pp. (in Russian).
- [34] H. Yao, Z. Zhang, G. Cheng, X. Chen, L. He, W. Wang, S. Zhou, P. Wang. “Automatic measurement of anatomical parameters of the lumbar vertebral body and the intervertebral disc on radiographs by deep learning”, *Quantitative Imaging in Medicine and Surgery*, **14**:8 (2024), pp. 5877–5890. 
- [35] N. Carabetta, G. Panuccio, S. Giordano, S. Sorrentino, G. A. Mazza, J. Sabatino, G. Canino, I. Leo, N. Salerno, A. Strangio, M. Petullà, D. Torella, Rosa S. De. “Reliability of immersive virtual reality for pre-procedural planning for TAVI: A CT-based validation”, *Journal of Cardiovascular Development and Disease*, **12**:12 (2025), id. 481, 14 pp. 
- [36] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. A. Ioannidis, G. S. Collins, M. Maruthappu. “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies”, *BMJ*, **368** (2020), id. m689, 12 pp. 