DOI: https://doi.org/10.17816/DD60622

Вариабельность заключений при интерпретации КТ-снимков: один за всех и все за одного

Н.С. Кульберг[1, 2], Р.В. Решетников[1, 3], В.П. Новик[1], А.Б. Елизаров[1], М.А. Гусев[1, 4], В.А. Гомболевский[1], А.В. Владзимирский[1], С.П. Морозов[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы, Москва, Российская Федерация
[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Российская Федерация
[3] Первый Московский государственный медицинский университет имени И.М. Сеченова (Сеченовский Университет), Москва, Российская Федерация
[4] Московский политехнический университет, Москва, Российская Федерация

АННОТАЦИЯ

Обоснование. Разметка наборов медицинских изображений во многом полагается на субъективную интерпретацию наблюдаемых подозрительных структур. На настоящий момент не существует рекомендованного протокола по определению эталонных данных (ground truth), основанных на врачебных описаниях.

Цель — анализ правильности и согласованности оценок рентгенологов, принимавших участие в подготовке общедоступного набора данных CTLungCa-500; определение взаимосвязи этих показателей с количеством специалистов, проводящих независимую интерпретацию изображений, полученных при компьютерно-томографическом (КТ) исследовании.

Материал и методы. Набор данных, в разметке которого принимали участие 34 рентгенолога, включает 536 КТ-исследований пациентов из группы риска развития рака лёгкого. Каждое КТ-исследование было независимо интерпретировано шестью специалистами, после чего обнаруженные ими подозрительные структуры проходили арбитраж другим экспертом. Для каждого эксперта подсчитывали количество истинно положительных, ложноположительных, истинно отрицательных и ложноотрицательных находок, на основании которых проводили оценку диагностической точности рентгенологов. Для анализа согласованности между заключениями рентгенологов использовали метрику процентного показателя.

Результаты. Увеличение количества специалистов, проводящих независимую интерпретацию КТ-исследований, ведёт к росту правильности их оценок при снижении согласованности. Среди факторов, влияющих на согласованность заключений между парами исследователей, выделяется расхождение мнений по поводу наличия лёгочного очага в конкретном участке КТ-снимка.

Заключение. Увеличение числа независимых первичных интерпретаций способно повысить их комбинированную правильность при условии проведения арбитража, причём квалификация рентгенологов не имеет определяющего значения для качества анализа. Проведение первичной разметки силами четырёх рентгенологов является оптимальным с точки зрения сочетания правильности интерпретации и её стоимости.

Ключевые слова: компьютерная томография; набор данных; эталонные данные; согласованность между заключениями.

Как цитировать

Кульберг Н.С., Решетников Р.В., Новик В.П., Елизаров А.Б., Гусев М.А., Гомболевский В.А., Владзимирский А.В., Морозов С.П. Вариабельность заключений при интерпретации КТ-снимков: один за всех и все за одного // *Digital Diagnostics*. 2021. Т. 2, № 2. С. 105–118. DOI: https://doi.org/10.17816/DD60622

Рукопись получена: 11.02.2021 **Рукопись одобрена:** 07.07.2021 **Опубликована:** 13.07.2021

DOI: https://doi.org/10.17816/DD60622

Inter-observer variability between readers of CT images: all for one and one for all

Nikolas S. Kulberg[1, 2], Roman V. Reshetnikov[1, 3], Vladimir P. Novik[1], Alexey B. Elizarov[1], Maxim A. Gusev[1, 4], Victor A. Gombolevskiy[1], Anton V. Vladzymyrskyy[1], Sergey P. Morozov[1]

[1] Moscow Center for Diagnostics and Telemedicine, Moscow, Russian Federation
[2] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russian Federation
[3] Institute of Molecular Medicine, The First Sechenov Moscow State Medical University, Moscow, Russian Federation
[4] Moscow Polytechnic University, Moscow, Russian Federation

ABSTRACT

BACKGROUND: The markup of medical image datasets is based on the subjective interpretation of the observed entities by radiologists. There is currently no widely accepted protocol for determining ground truth based on radiologists' reports.

AIM: To assess the accuracy of radiologist interpretations and their agreement for the publicly available dataset "CTLungCa-500", as well as the relationship between these parameters and the number of independent readers of CT scans.

MATERIALS AND METHODS: Thirty-four radiologists took part in the dataset markup. The dataset included 536 patients who were at high risk of developing lung cancer. For each scan, six radiologists worked independently to create a report. After that, an arbitrator reviewed the lesions discovered by them. The number of true-positive, false-positive, true-negative, and false-negative findings was calculated for each reader to assess diagnostic accuracy. Further, the inter-observer variability was analyzed using the percentage agreement metric.

RESULTS: An increase in the number of independent readers providing CT scan interpretations leads to accuracy increase associated with a decrease in agreement. The majority of disagreements were associated with the presence of a lung nodule in a specific site of the CT scan.

CONCLUSION: If arbitration is provided, an increase in the number of independent initial readers can improve their combined accuracy. The experience and diagnostic accuracy of individual readers have no bearing on the quality of a crowd-tagging annotation. At four independent readings per CT scan, the optimal balance of markup accuracy and cost was achieved.

Keywords: X-ray computed tomography; datasets as topic; ground truth; observer variation.

To cite this article
Kulberg NS, Reshetnikov RV, Novik VP, Elizarov AB, Gusev MA, Gombolevskiy VA, Vladzymyrskyy AV, Morozov SP. Inter-observer variability between readers of CT images: all for one and one for all. *Digital Diagnostics*. 2021;2(2):105–118. DOI: https://doi.org/10.17816/DD60622

Received: 11.02.2021 **Accepted:** 07.07.2021 **Published:** 13.07.2021

DOI: https://doi.org/10.17816/DD60622

CT图像解释中结论的可变性：一个为所有和所有为一

Nikolas S. Kulberg[1, 2], Roman V. Reshetnikov[1, 3], Vladimir P. Novik[1], Alexey B. Elizarov[1], Maxim A. Gusev [1, 4], Victor A. Gombolevskiy[1], Anton V. Vladzymyrskyy[1], Sergey P. Morozov[1]

[1] Moscow Center for Diagnostics and Telemedicine, Moscow, Russian Federation
[2] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russian Federation
[3] Institute of Molecular Medicine, The First Sechenov Moscow State Medical University, Moscow, Russian Federation
[4] Moscow Polytechnic University, Moscow, Russian Federation

结构简评

理由： 医学图像集的标记在很大程度上依赖于观察到的可疑结构的主观解释。目前，没有推荐的协议用于根据医学描述确定参考数据（ground truth）。

目标： 评估参与编制公开数据集»CTLungCa-500»的放射科医生评估的正确性和一致性，以及确定这些指标与对CT研究进行独立解释的专家数量的关系。

方法： 该数据集包括有患肺癌风险的患者的536项CT研究，其中34名放射科医生参加了该研究。每项CT研究都由六位专家独立解释，之后他们发现的可疑结构由另一位专家进行仲裁。对于每位专家计算真阳性，假阳性，真阴性和假阴性结果的数量，在此基础上评估放射科医生的诊断准确性。为了分析放射科医生的结论之间的一致性，使用了百分比度量。

结果：对CT研究进行独立解释的专家数量的增加在一致性降低的情况下导致其评估的正确性增加。在影响成对研究人员之间结论一致性的因素中，关于CT图像的特定部分中存在肺焦点的观点不一致。

结论：独立的初级解释数量的增加使它们的组合正确性会升高，但需要仲裁，放射科医生的资格对分析的质量没有决定性的价值。从结合解释的正确性及其成本的角度来看，由四名放射科医生进行主要标记是最佳的。

关键词： 计算机断层扫描，数据集，参考数据，结论之间的一致性.

引用本文：
Kulberg NS, Reshetnikov RV, Novik VP, Elizarov AB, Gusev MA, Gombolevskiy VA, Vladzymyrskyy AV, Morozov SP. CT图像解释中结论的可变性：一个为所有和所有为一. *Digital Diagnostics*. 2021;2(2):105–118. DOI: https://doi.org/10.17816/DD60622

收到： 11.02.2021 **接受：** 07.07.2021 **发布日期：** 13.07.2021

ОБОСНОВАНИЕ

В 2017 г. С.П. Морозов и соавт. подготовили общедоступный набор данных «Тэгированные результаты компьютерных томографий лёгких», впоследствии получивший название «CTLung500-Ca» [1, 2]. Этот набор содержит 536 рентгеновских изображений органов грудной клетки пациентов из группы риска развития рака лёгкого, полученных методом компьютерно-томографического (КТ) исследования. Интерпретацию каждого исследования независимо проводили шестеро рентгенологов, находки впоследствии проходили проверку дополнительным экспертом. При разметке использовали подход со слабой аннотацией находок, т.е. указание ограниченного числа очагов на КТ-снимке, локализацию которых осуществляли заданием координат охватывающих сфер максимального диаметра с последующей их кластеризацией [2, 3]. С.П. Морозов и соавт. разработали такой протокол разметки и аннотации, поскольку интерпретации рентгенологов имеют склонность к субъективности и не защищены от ошибок. В условиях когда стоимость ложноположительных и ложноотрицательных находок одинаково высока, арбитраж первичных интерпретаций может увеличить правильность заключений [4]. Отметим, что такой арбитраж эффективен только в случае, если рентгенологи совершают разные ошибки. Согласно P.G. Herman и S.J. Hessel, вероятность того, что одну и ту же ложноположительную находку сделают два и более рентгенологов, невелика. Однако существенную долю ложноотрицательных ошибок, как правило, совершают два и более специалистов [5]. Таким образом, количество рентгенологов, проводящих независимую интерпретацию КТ-снимков, может существенно повлиять на точность разметки и аннотации.

Цель исследования. Первичной целью настоящего исследования являлось изучение взаимосвязи между числом независимых интерпретаций расположенных в базе данных CTLungCa-500 КТ-снимков и количеством и типом совершаемых ошибок, а также поиск протокола интерпретации КТ-исследований, способствующего оптимальной точности разметки. Вторичной целью исследования было изучение согласованности заключений рентгенологов, принимавших участие в подготовке набора данных.

МЕТОДЫ

Дизайн исследования

В настоящей работе анализировали данные ретроспективного многоцентрового обсервационного исследования, посвящённого изучению перспектив применения технологий компьютерного зрения в системе здравоохранения г. Москвы.

Критерии соответствия

Критерии включения: пациенты поликлиник г. Москвы в возрасте от 50 до 75 лет, которые проходили диагностическое КТ-исследование по направлению лечащего врача в связи с подозрением на рак лёгкого.

Условия проведения

Согласно критериям включения, из Единого радиологического информационного сервиса было выгружено 3897 КТ-исследований. Из этого количества случайным образом отобрали 550 КТ-исследований для создания набора данных «Тэгированные результаты компьютерных томографий лёгких». Из выборки исключили 14 КТ-снимков по причине несоответствия критериям включения либо протоколу медицинского вмешательства.

Продолжительность исследования

Набор данных включает результаты КТ-исследований, проводимых с 01 января 2015 г. по 31 декабря 2017 г.

Описание медицинского вмешательства

Рекомендуемые параметры сканирования для взрослых пациентов (рост 170 см, масса тела 70 кг): автоматическая модуляция тока на трубке при напряжении 120 кВ, FOV 350 мм, толщина среза ≤1,5 мм, расстояние между соседними срезами ≤ толщины среза. Сканирование проводили в положении пациента лежа на спине, направление сканирования — от диафрагмы до верхушек лёгких за одну задержку дыхания на вдохе. Ядра реконструкции были специфичны для конкретного производителя томографа: для аппаратов Toshiba — FC50, FC51, FC52, FC53, FC07 для лёгких и FC07, FC08, FC09, FC17, FC18 для мягких тканей; для аппаратов Siemens — B70, B75 и B80; для аппаратов Philips — Y-Sharp и LUNG для лёгких и SOFT для мягких тканей; для аппаратов GE (General Electrics) — LUNG для лёгких и SOFT для мягких тканей.

Основной исход исследования

В разметке и аннотировании исследований принимали участие две группы рентгенологов-волонтёров. Представители первой группы (первичные эксперты), состоящей из 15 специалистов с опытом от 2 до 10+ лет, выполняли первичную интерпретацию КТ-снимков. Согласно сформированной методологии, врачи искали на изображениях КТ лёгочные очаги размерами от 4 до 30 мм, сохраняя такую информацию о находках, как локализацию лёгочного очага (положение центра находки по двум измерениям на изображении и номеру среза); диаметр находки; тип лёгочного очага (солидный, полусолидный или очаг по типу матового стекла). Врачам рекомендовали не отмечать кальцинированные и перифиссуральные очаги в лёгких, а также не отмечать более пяти крупнейших лёгочных очагов на одном КТ-снимке. Для уменьшения вероятности пропусков потенциальных лёгочных очагов каждое исследование

независимо просматривали шестеро рентгенологов. Затем один из участников второй группы (арбитры), составленной из трёх рентгенологов с опытом 10+ лет, просматривал разметку, сделанную рентгенологами первой группы, оценивая достоверность каждой отметки. Арбитры также проводили оценку злокачественности обнаруженных очагов, относя их к категории «злокачественный» или «доброкачественный», руководствуясь рекомендациями Флейшнеровского общества [6].

Этическая экспертиза

Исследование, данные которого использовали для анализа в настоящей работе, получило одобрение независимого этического комитета Московского регионального отделения Российского общества рентгенологов и радиологов (протокол № 2 1-II-2020 от 20 февраля 2020 г.). Все процедуры, выполняемые пациентам в ходе исследования, соответствовали стандартам регионального и национального исследовательского комитета, а также Хельсинкской декларации и Декларации Тайбэя Всемирной медицинской ассоциации.

Статистический анализ

Для определения специфичности и чувствительности индивидуальных специалистов подсчитывали количество истинно положительных, ложноположительных, истинно отрицательных и ложноотрицательных находок для каждого рентгенолога, проводившего первичную интерпретацию. Истинно положительными (ИП) признавали случаи, в которых совпадали мнения рентгенолога и арбитра по поводу наличия и типа лёгочного очага (солидный, полусолидный или уплотнение по типу матового стекла) в конкретной области. К ложноположительным (ЛП) относили случаи, в которых арбитр признал оценку первичного эксперта ошибочной в отношении наличия или типа лёгочного очага в данном участке. Истинно отрицательными (ИО) признавали ситуации, в которых рентгенолог не отметил уплотнения, некорректно, по мнению арбитра, принятого за очаговое образование одним или более из пяти других первичных экспертов. Наконец, к ложноотрицательным (ЛО) относили случаи, в которых рентгенолог не распознавал лёгочный очаг, корректно, по мнению арбитра, идентифицированный одним или более из пяти других участников. При анализе данных исходили из предположения, что суждение арбитра всегда верное.

Чувствительность (sensitivity, Se) рассчитывали по формуле

$$Se = \frac{\text{ИП}}{(\text{ИП} + \text{ЛО})}. \quad (1)$$

Специфичность (specificity, Sp) вычисляли как

$$Sp = \frac{\text{ИО}}{\text{ИО} + \text{ЛП}}. \quad (2)$$

Для каждого участника определяли индекс Юдена (J):

$$J = Se + Sp - 1. \quad (3)$$

Для расчёта показателя правильности (accuracy, *Acc*) различных выборок первичных экспертов ИП результатом признавали случаи, в которых хотя бы один специалист из выборки корректно, по мнению арбитра, идентифицировал лёгочный очаг в конкретной области КТ-снимка. К ИО результатам относили случаи, в которых хотя бы один специалист из выборки не отмечал уплотнения, некорректно, по мнению арбитра, принятого за лёгочный очаг любым другим участником исследования. Правильность вычисляли как

$$Acc = \frac{(\text{ИП} + \text{ИО})}{(\text{П} + \text{О})} \times 100, \quad (4)$$

где П — количество корректных находок, О — количество некорректных находок.

Существует целый ряд метрик для оценки согласованности у одного или нескольких исследователей. В частности, O. Gerke и соавт. в своих рекомендациях по систематизации исследований согласованности предлагают использовать анализ Бланда–Альтмана [7]. Другими распространёнными метриками являются каппа Коэна [8] и Флейса [9]. Однако при всех преимуществах этих методов они сложны в интерпретации, поэтому авторы настоящей работы остановились на простейшем варианте — процентном показателе согласованности между исследователями (Inter-observer agreement, IOA), который не учитывает фактора случайных совпадений заключений рентгенологов, но при этом интуитивно понятен и достоверно отображает основные закономерности при условии проведения повторных экспериментов. Процентный показатель рассчитывали как долю очагов, для которых мнения экспертов (наличие, тип) совпадали в общем числе совместно размеченных очагов:

$$\text{IOA} = \frac{\text{Совпадения}}{\text{Совпадения} + \text{Несовпадения}} \times 100. \quad (5)$$

Статистический анализ проводили, используя пакеты dplyr [10], irr [11] и ggplot2 [12] для R 3.6.3 [13]. При подготовке данных использовали самостоятельно разработанные скрипты на языке Python 3.8.2 [14].

РЕЗУЛЬТАТЫ

Объекты исследования

Всего в первичной интерпретации КТ-снимков принял участие 31 рентгенолог. Каждый рентгенолог из исходной когорты, состоящей из 15 специалистов, в процессе исследования был заменён другим специалистом по причине отказа или невозможности продолжать исследование; один участник был заменён дважды.

Нагрузка на рентгенологов была распределена неравномерно. Каждый специалист из исходной когорты принял участие в разметке и аннотации в среднем 1050±140 подозрительных структур. Заменившие их рентгенологи разметили в среднем по 110±42 очаговых образования.

По результатам разметки набор данных включал 72 КТ-исследования, на которых рентгенологи не обнаружили лёгочных очагов от 4 до 30 мм, и 464 КТ-снимка с лёгочными очагами, содержащие в сумме 3151 подтверждённую арбитром находку. Из этого количества 1761 очаг эксперты отнесли к вероятно злокачественным образованиям, 445 — к доброкачественным, 945 уплотнений имели иную природу (содержали кальцинаты, жировую, фиброзную ткань либо жидкость).

Основные результаты исследования

Чувствительность и специфичность рентгенологов, принимавших участие в разметке

В процессе работы над набором данных каждому рентгенологу был присвоен трёхзначный идентификационный номер (ID). В случае замены специалиста новый участник наследовал его ID с дополнительным символом «+». Среднее значение чувствительности составило 34,9% (95% доверительный интервал [ДИ] 30,4–39,4), специфичности — 78,4% (95% ДИ 74,9–81,9), что заметно уступает минимальным показателям, продемонстрированным в схожем сценарии исследования рентгенологами в работе D. Ardila и соавт.: 62,5% (95% ДИ 54,4–70,7) и 95,3% (95% ДИ 94,0–96,6) соответственно [15].

Возможной причиной наблюдаемого различия являются условия разметки, руководствуясь которыми первичные эксперты размечали максимум пять очагов на снимке. Эта рекомендация основана на результатах исследования NELSON, согласно которым риск первичного рака повышается с ростом количества очагов до четырёх, но снижается для пациентов с пятью и более очагами [16]. В случаях множественных очагов (>5) такой подход способен искусственно занизить диагностическую точность первичных экспертов, поскольку привносит дополнительную степень свободы, связанную с конкретным набором очагов, который разметил каждый рентгенолог. Корректировать эту неопределённость можно введением альтернативной классификации находок, признавая за ИП случаи, в которых первичный эксперт разметил хотя бы один подтверждённый очаг на КТ-снимке. При подобной схеме оценки средняя чувствительность первичных экспертов составила 66,2% (95% ДИ 62,1–69,9), специфичность — 78,5% (95% ДИ 72,3–84,8). Однако целью разметки было создание набора данных, предназначенного для обучения алгоритмов искусственного интеллекта, вследствие чего интерес представляет каждая подозрительная структура на КТ-снимке. По этой причине в настоящей работе для оценки диагностической точности использовали критерии, изложенные в разделе «Методы». В соответствии с этими критериями по показателю индекса Юдена наибольшую эффективность продемонстрировал рентгенолог с ID 012+ (J=0,472), наименьшую (J=-0,188) — специалист с ID 008+ (табл. 1).

Влияние числа исследователей на правильность интерпретации

Интерпретация двумя первичными экспертами. В данном анализе рассматривали выборку из 97 КТ-исследований, в интерпретации которых принимал участие рентгенолог с ID 012+, продемонстрировавший наивысший индекс Юдена среди всех участников (см. табл. 1). При таком размере выборки все полученные оценки могут отличаться от средних для полного набора данных не более чем на 10% [17]. Размеченная этим специалистом выборка содержала 53 солидных лёгочных очага, 6 полусолидных и 5 уплотнений по типу матового стекла. Правильность оценок рентгенолога 012+ составила 65,98%: он корректно идентифицировал 28 солидных очагов и избежал 32 из 33 ложноположительных ошибок, совершённых другими специалистами на этих же исследованиях, неверно распознав при этом 2 солидных очага и 1 полусолидный и совершив 34 ложноотрицательных ошибки. Помимо него, в разметке всех 97 КТ-исследований в выборке участвовал также рентгенолог с ID 012, имеющий один из самых низких индексов Юдена (0,058, 24-е место, см. табл. 1). Этот специалист корректно распознал 32 солидных очага, 1 полусолидный, 1 уплотнение по типу матового стекла, избежал 18 ложноположительных ошибок. При согласованности между исследователями равной 59,8% комбинированная правильность их оценок составила 81,44%. Источниками несогласованности послужили расхождение мнений в паре по поводу наличия подозрительной структуры в конкретной области (92,3% случаев) и типа лёгочного очага (7,7% случаев).

Распределение КТ-исследований между специалистами проводили случайным образом. По этой причине получилось так, что интерпретацию всех 97 КТ-исследований в изучаемой выборке проводили только первичные эксперты 012 и 012+. Помимо них в разметке выборки приняли участие 17 рентгенологов (в скобках для каждого указано число размеченных очагов): 000(11), 002(54), 003(30), 004(27), 005(18), 006(40), 007(10), 008(16), 009(17), 010(32), 011(24), 013(30), 014(52), 004+(7), 005+(10), 011+(1) и 014+(9), что предоставило возможность сравнить ситуацию, в которой второе мнение по всем исследованиям в выборке выражает один специалист, с моделью многопользовательской разметки (crowd-tagging), в которой это мнение обеспечивает участник, отбираемый случайным образом из некоторой группы экспертов с вариабельными показателями специфичности и чувствительности.

Таблица 1. Диагностическая точность участников исследования

ID эксперта	Показатели по отдельным очагам			
	Se, %	Sp, %	Индекс Юдена	Число размеченных очагов*
000	39,52	73,17	0,127	1079
001	32,63	79,04	0,117	1068
002	28,25	80,19	0,084	1045
003	44,05	67,75	0,118	1094
004	31,37	68,75	0,001	844
005	33,08	72,76	0,058	1222
006	36,91	71,32	0,082	1085
007	37,31	73,43	0,107	884
008	42,01	68,00	0,100	1227
009	36,79	79,50	0,163	1265
010	38,62	71,16	0,098	1166
011	26,05	79,51	0,056	853
012	33,97	71,88	0,058	1045
013	38,52	77,40	0,159	1028
014	37,16	82,32	0,195	850
000+	31,63	79,17	0,108	194
001+	52,94	82,46	0,354	108
002+	62,50	57,14	0,196	46
003+	60,71	86,21	0,469	86
004+	27,78	86,49	0,143	110
005+	41,49	75,86	0,173	152
006+	31,34	74,14	0,055	125
007+	29,73	85,71	0,154	86
008+	18,99	62,16	-0,188	176
009+	25,76	85,11	0,109	113
010+	25,00	75,36	0,004	145
011+	31,58	93,33	0,249	68
012+	53,85	93,33	0,472	97
013+	34,29	85,71	0,170	77
014+	17,95	100,0	0,179	63
000++	0,00	94,87	-0,051	48

Примечание. * Учитываются все найденные очаги в КТ-исследованиях, в разметке которых принимал участие эксперт, вне зависимости от того, распознал он их или нет.

В первую группу попали 6 исследователей (табл. 2). Средний индекс Юдена в этой группе составил 0,078±0,045 (максимальное значение 0,127, минимальное — 0,001), что превышает показатель участника с ID 012 (0,058). Тем не менее согласованность оценок с экспертом 012+ составила всего лишь 40,2%, а комбинированная правильность оценок — 74,23%. Источником большинства несогласий в паре (97,4%) стало расхождение мнений по поводу наличия лёгочного очага.

В повторном аналогичном эксперименте анализировали группу с другим составом участников (табл. 3). Число и состав участников различались между группой 1 (см. табл. 2) и группой 2 (см. табл. 3); более того, распределение числа очагов, размеченных каждым из них, было неравномерным.

Таблица 2. Распределение размеченных подозрительных структур в группе 1

ID исследователя	000	002	003	004	005	006
Количество размеченных очагов	11	54	9	3	11	9

DOI: https://doi.org/10.17816/DD60622

Таблица 3. Распределение размеченных подозрительных структур в группе 2

ID исследователя	005+	010	003	004	005	006	008	009
Количество размеченных очагов	10	10	21	9	7	31	8	1

Средний индекс Юдена в группе 2 равнялся 0,099±0,055 (максимальное значение 0,173, минимальное — 0,01) и был выше, чем у участника 012 и в группе 1. Согласованность и комбинированная правильность оценок участников группы 2 и рентгенолога 012+ также были наивысшими из трёх рассмотренных вариантов интерпретации КТ-исследований двумя экспертами, составляя 71,1 и 83,50% соответственно. Несогласие между исследователями в 89,3% случаев было ассоциировано с наличием лёгочного очага в данном участке и в 10,7% — с его типом. Средняя правильность интерпретаций при первичной разметке двумя специалистами в любых сочетаниях составила 79,72±4,87%.

Интерпретация тремя и более исследователями. При анализе интерпретации тремя и более исследователями все группы включали исследователей 012 и 012+. При первичной разметке и аннотации тремя рентгенологами согласованность их оценок колебалась от 32,0 до 42,3%, средняя комбинированная правильность составила 89,18±5,10%. Согласованность оценок четырёх независимых специалистов упала до 16,5±5,7% при росте средней комбинированной правильности до 93,82±3,57%. Для пяти рентгенологов согласованность оценок продолжила снижаться до 9,8±8,1%, а правильность — повышаться до 97,94±0,14%. Наконец, комбинированная правильность шести экспертов составила 100% в условиях нашего эксперимента при согласованности 3,1% (рис. 1). Таким образом, наблюдается значительная обратная корреляция между правильностью и согласованностью оценок экспертов: r=-0,78, p <0,05.

В подтверждение выводов P.G. Herman и S.J. Hessel [5] в выборке из 97 исследований при интерпретации шестью специалистами 85,7% ложноположительных ошибок были совершены только одним экспертом, 11,4% — двумя, 2,9% — тремя одновременно. Все шесть экспертов корректно идентифицировали 8,1% положительных находок в выборке; 25,8% ложноотрицательных ошибок были совершены одним экспертом из шести, 8,1% — двумя, 8,1% — тремя, 19,3% — четырьмя, 30,6% — пятью (рис. 2).

Стоимость разметки

Для оценки оптимальной эффективности разметки с позиции рационального использования ресурсов необходимо учитывать стоимость задействования дополнительных экспертов в интерпретации КТ-снимков. Таким образом, можно будет сопоставить улучшение правильности с увеличением расходов на аннотацию исследований.

Поскольку в разметке набора данных принимали участие рентгенологи-волонтёры, их труд не оплачивался. Вследствие этого расчёт стоимости разметки целесообразно проводить в терминах затраченного экспертами времени. В среднем первичный эксперт затрачивал на интерпретацию одного КТ-снимка 12 мин, арбитр — 4 мин. В настоящем исследовании стоимость устранения ошибки *C* в изучаемой выборке из 97 КТ-снимков рассчитывали как разницу средней стоимости разметки заданным числом первичных экспертов с привлечением арбитра и стоимости разметки одним рентгенологом

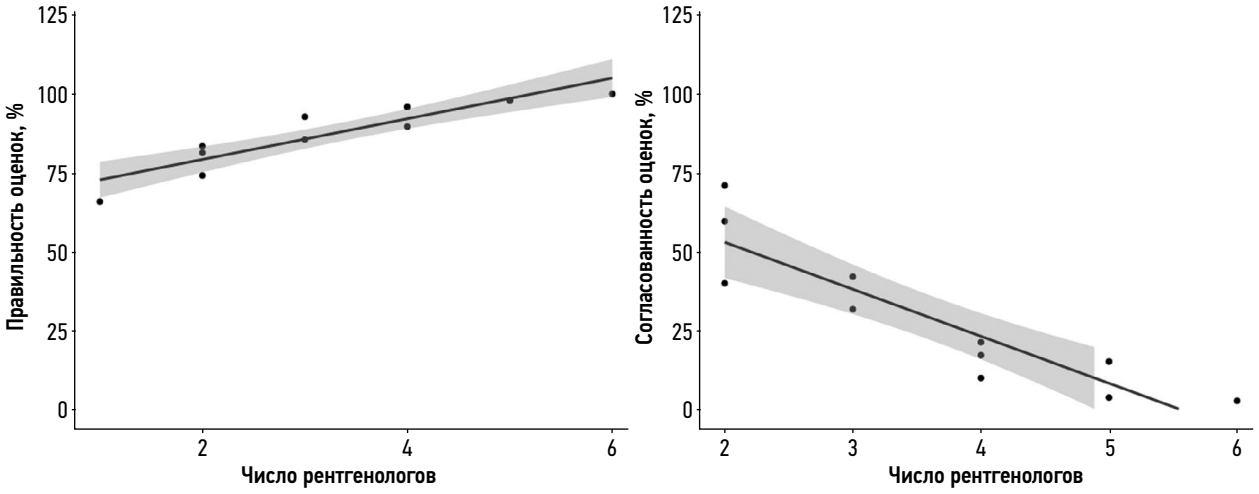

Рис. 1. Правильность и согласованность оценок как функция от числа рентгенологов, принимающих участие в первичной разметке. Серым цветом показан 95% доверительный интервал. Точки соответствуют разным выборкам первичных экспертов. Для экспериментов с двумя, тремя и четырьмя экспертами отбирали по три различающихся выборки из исходных шести рентгенологов; для пяти — по две.

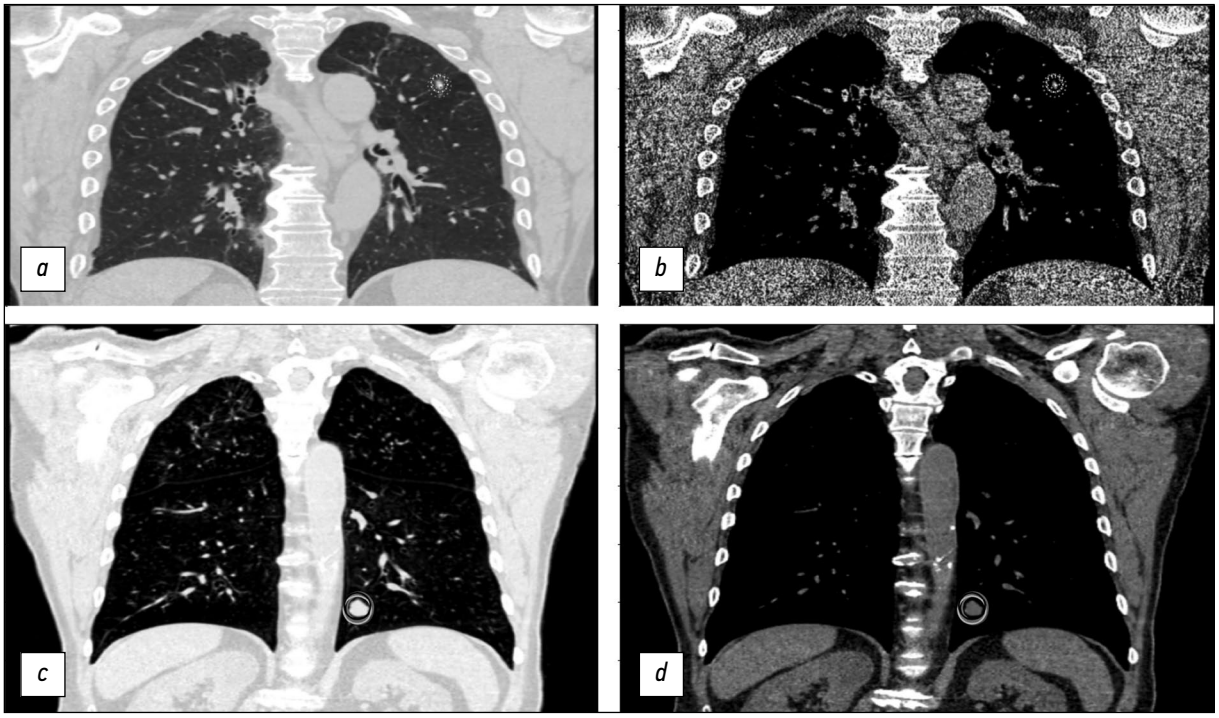

Рис. 2. Примеры КТ-исследований с существенным несогласием (*a*, *b*, CTLungCa-500 AN RLADD02000018919, ID RLSDD02000018855) и полным согласием (*c*, *d*, CTLungCa-500 AN RLAD42D007-25151, ID RLSD42D007-25151) между экспертами. Исследования приведены во фронтальной проекции в лёгочном (*a*, *c*) и мягкотканном (*b*, *d*) режимах. Отметки рентгенологов показаны различающимися цветами: *a*, *b* — очаг разметили пять первичных экспертов из шести, четверо присвоили ему солидный тип и один — полусолидный. Арбитр не согласился с их мнением, признав находку доброкачественным кальцинатом; *c*, *d* — все шестеро первичных экспертов и арбитр классифицировали очаг как потенциально злокачественный солидный.

без привлечения арбитра, поделённую на количество устранённых ошибок (N_{err}):

$$C = \frac{(n \times 12 \times 97 + n \times 4 \times 97) - 12 \times 97}{N_{err}}, \qquad (6)$$

где n — число первичных экспертов.

Эксперт 012+ допустил 33 ложноположительных и ложноотрицательных ошибки. Количество устранённых ошибок, достигаемое за счёт привлечения дополнительных экспертов и проведения арбитража, а также соответствующая стоимость устранения ошибки представлены в табл. 4. Существует закономерность, согласно которой каждый новый первичный эксперт увеличивает стоимость устранения ошибки на 42,5±10,7 мин, за исключением одной точки. Разметка набора данных силами четырёх первичных экспертов с последующим арбитражем сопровождалась резким повышением числа устранённых ошибок и, соответственно, снижением стоимости (см. табл. 4).

Таблица 4. Оценка стоимости устранения ошибки

Число первичных экспертов	Число устранённых ошибок	Стоимость, мин/ошибка
2	15	129,3
3	19	183,8
4	29	173,9
5	31	212,8
6	33	246,9

Дополнительные результаты исследования

Из-за особенностей дизайна исследования, в котором каждый эксперт интерпретировал индивидуальный КТ-снимок только по одному разу, в рамках настоящей работы не проводили оценку согласованности заключений у отдельных рентгенологов. Среднее значение согласованности оценок между парами специалистов составило 60,5±5,3%, с минимальным значением 53,1% и максимальным 73,0%.

Другим способом оценить согласие между первичными экспертами является анализ положительных находок каждого рентгенолога (рис. 3). Для каждого представителя исходной когорты максимальная доля выявленных очагов (37,6±5,4%) соответствовала уникальным находкам, не распознанным другими экспертами (см. рис. 3, *a*). Затем в порядке убывания следуют находки, с которыми был согласен один (21,4±2,8%), два (14,0±2,0%), четыре (9,5±2,3%), три (9,2±1,8%) и пять (8,1±3,1%) первичных экспертов. Только для четырёх рентгенологов из исходной когорты (ID 002, 004, 007 и 010) доля единогласно одобренных находок превышает 10%. Отметим,

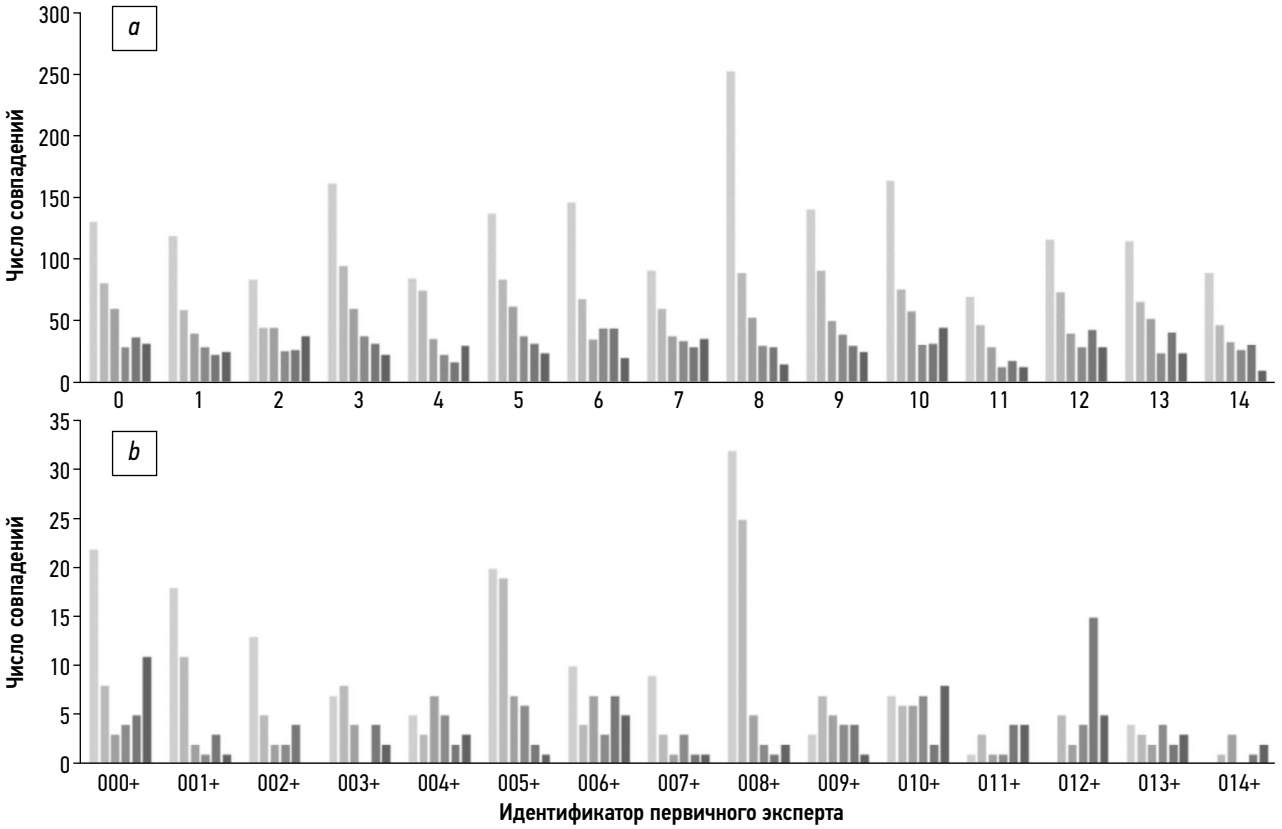

Рис. 3. Согласие между первичными экспертами: *a* — для представителей исходной когорты из 15 рентгенологов; *b* — для рентгенологов, пришедших им на замену. Данные для эксперта с ID 000++ не приведены по причине малого количества отмеченных очагов. Для каждого рентгенолога первый столбец соответствует числу очагов, уникально размеченных этим специалистом (ни один из пяти других экспертов не распознал данную находку). Далее следуют столбцы, соответствующие случаям, когда выявленный рентгенологом очаг отмечали один, два, три, четыре и пять других первичных экспертов. При построении графика не учитывали одобрение арбитра, а также расхождения во мнениях между рентгенологами по поводу типа очага.

что ни один из этих экспертов не входит в лидирующую группу по значению индекса Юдена, рассчитанного по предложенной в настоящей работе методике; более того, эксперт 004 является худшим в когорте по этому показателю (см. табл. 1). В то же время эксперт 014 с максимальным в когорте индексом Юдена (0,195) ничем не выделяется среди своих коллег по согласованности положительных находок (см. рис. 3, *a*).

В когорте рентгенологов, пришедших на замену исходным первичным экспертам, было иное распределение согласованности находок (см. рис. 3, *b*). Максимальная доля выявленных очагов (28,9±18,2%) была всё так же представлена уникальными находками. Затем следовали находки, выявленные одновременно двумя (23,3±11,0%), тремя (13,3±10,7%), пятью (13,2±11,9%), шестью (11,5±9,8%) и четырьмя (9,7±7,6%) экспертами. В этой когорте было уже восемь рентгенологов (ID 000+, 004+, 006+, 010+, 011+, 012+, 013+, 014+), для которых доля единогласно одобренных положительных находок превышала 10%, причём для четырёх из них (ID 000+, 010+, 011+, 014+) она была выше 20%. Тем не менее эти показатели могут быть обусловлены небольшим количеством положительных находок в данной когорте, о чём косвенно свидетельствует высокая вариация их согласованности, выраженная в отношении средних значений и стандартных отклонений. В качестве примера можно привести эксперта 014+, участвовавшего в интерпретации КТ-исследований, на которых другие эксперты выявили 63 подозрительные структуры (см. табл. 1). Этот эксперт разметил только семь очагов, из которых один был также выявлен одним другим экспертом, три — двумя, один — пятью и два — шестью (см. рис. 3, *b*). При этом эксперт совершил 32 ложноотрицательных ошибки, проигнорировав таким образом ~50% истинно положительных находок. Для этой когорты также не наблюдали корреляции между согласованностью положительных находок и индексом Юдена эксперта.

ОБСУЖДЕНИЕ

Резюме основного результата исследования

Наши результаты демонстрируют, что увеличение количества специалистов, проводящих независимую интерпретацию КТ-исследований, ведёт к росту правильности их оценок, причём уровень квалификации не оказывает существенного воздействия ни на согласованность мнений рентгенологов, ни на их комбинированную правильность. Среди факторов, влияющих

на согласованность заключений между парами исследователей, выделяется расхождение мнений по поводу наличия очага в конкретном участке КТ-снимка.

Обсуждение основного результата исследования

На настоящий момент не существует консенсусного мнения по поводу рекомендуемого количества рентгенологов, участвующих в первичной разметке и аннотации наборов данных медицинских изображений. Это значение, как правило, находится в границах от одного [18, 19] до четырёх [20]. Единственным известным нам исследованием, затрагивающим данный вопрос, является работа P.G. Herman и S.J. Hessel, согласно которой при увеличении числа специалистов, предоставляющих независимые интерпретации исследований, происходит постепенное снижение числа безошибочных описаний [5]. Хотя это, безусловно, интересное наблюдение, оно не представляет особой практической ценности, поскольку модель с арбитражем в принципе основана на предположении, что первичные интерпретации будут содержать ошибки. Более того, её эффективность возрастает при условии, что эти ошибки будут разными.

Последнее утверждение не всегда является верным. В частности, результаты настоящей работы показывают, что совершение рентгенологами разных ошибок не приводит автоматически к повышению комбинированной правильности их заключений. В эксперименте с двумя специалистами, проводившими первичную интерпретацию КТ-снимков, наибольший уровень несогласия наблюдали во второй паре (согласованность 40,2%), однако она же продемонстрировала и наименьшую правильность из трёх рассмотренных (74,2 против 81,4 и 83,5%). При этом самое высокое значение правильности с максимальной согласованностью (71,1%) показала третья пара. Тем не менее, согласно полученным в настоящей работе данным, существует значительная отрицательная корреляция между согласованностью оценок экспертов и их правильностью (r=-0,78). Так, при первичной интерпретации двумя рентгенологами наблюдали согласованность 57,0±15,6% при правильности 79,7±4,9%; для пяти рентгенологов эти показатели равнялись 9,8±8,1% и 97,9±0,1% соответственно, и эта зависимость сохранялась во всех рассмотренных вариантах разметки набора данных (см. рис. 1).

Согласно результатам настоящего исследования, оптимального сочетания правильности и стоимости разметки позволяет добиться подход с привлечением четырёх первичных экспертов и последующим арбитражем (см. табл. 4). Для него наблюдается резкое увеличение числа устранённых ошибок по сравнению с разметкой силами трёх рентгенологов, что сопровождается снижением времени, затрачиваемого на устранение одной ошибки (-9,9 мин). Задействование дополнительных первичных экспертов приводило к дальнейшему росту правильности интерпретаций, однако это происходило за счёт увеличения стоимости устранения ошибки в среднем на 42,5±10,7 мин.

В настоящей работе при отнесении оценок первичных экспертов к категориям ЛО, ИО, ЛП и ИП опирались на предположение, что на каждом КТ-снимке будут размечены все лёгочные очаги. Однако результаты исследования свидетельствуют о том, что участники исследования ограничивались пятью крупнейшими лёгочными очагами на КТ-снимке, выполняя выданные им рекомендации. Таким образом, существенная доля лёгочных очагов была проигнорирована индивидуальными рентгенологами, что сказалось на их показателях диагностической точности, а также значениях согласованности в парах экспертов. Тем не менее расхождения во мнениях первичных экспертов являются желательным исходом при использовании арбитража, поскольку расширяют каталог отмеченных подозрительных структур. Это снижает долю ложноотрицательных находок, даже в условиях искусственных ограничений на число размечаемых очагов. Один из главных выводов настоящей работы — то, что консенсус между несколькими рентгенологами не является необходимым условием для качественной разметки наборов данных. Основная ответственность лежит на арбитрах, которые должны корректно интерпретировать все отмеченные первичными экспертами подозрительные структуры (см. рис. 2, *a*, *b*).

Ограничения исследования

Основным ограничением настоящей работы является модель определения эталонных данных (ground truth) — тех находок, которые следует считать лёгочными очагами. При интерпретации КТ-снимков рентгенологи не имели доступа к клиническим, биологическим и геномным данным пациентов; более того, ни для одного из пациентов набор не содержал двух разнесённых во времени исследований, которые позволили бы оценить динамику развития подозрительных структур. Мы исходили также из предположения, что мнение арбитра всегда правильное, и трактовали несогласие между первичным мнением и мнением арбитра всегда в пользу последнего. Однако набор содержит ряд примеров, которые вызывают сомнение в надёжности такого подхода: в частности, 19 лёгочных очагов были отмечены арбитром одновременно как доброкачественные и злокачественные. Это согласуется с результатами S.J. Hessel и соавт., продемонстрировавших, что арбитры способны корректно разрешить лишь порядка 80% несогласий между первичными экспертами [4].

Другим ограничением работы является невозможность проведения оценки воспроизводимости заключений отдельных рентгенологов. Для достижения основных целей исследования использовали ограниченную выборку; для более достоверной статистики оптимальным подходом было бы использование метода размножения

выборок (bootstrap). Наконец, оценка диагностической точности первичных экспертов в настоящем исследовании опиралась на предположение, что они будут размечать все лёгочные очаги. В случае если количество очагов на КТ-снимке превышало пять, это предположение входило в конфликт с рекомендациями по разметке, что могло влиять на итоговые индивидуальные показатели чувствительности и специфичности. Для компенсации этого методологического ограничения авторы исследования провели попытку оценки согласованности по числу положительных находок каждого первичного эксперта, одобренных двумя, тремя, четырьмя и пятью другими рентгенологами (см. рис. 3). Однако такой анализ не учитывает ложноотрицательные ошибки, в связи с чем его результаты не коррелируют с полученными значениями индекса Юдена для каждого эксперта. В довершение, в настоящей работе изучали результаты интерпретации полнодозных КТ-исследований. Таким образом, её выводы могут не распространяться на данные, полученные в ходе скрининг-исследований, для которых характерно использование низкодозных и ультранизкодозных протоколов КТ.

ЗАКЛЮЧЕНИЕ

Несмотря на ограничения, настоящая работа убедительно демонстрирует, что увеличение числа независимых первичных интерпретаций способно повысить их правильность при условии проведения арбитража. При этом квалификация рентгенологов не имеет определяющего значения для качества анализа, поскольку, согласно полученным результатам, комбинированная правильность их оценок не зависела от индивидуальных индексов Юдена. Оптимальное сочетание правильности и стоимости разметки достигается при первичной независимой интерпретации КТ-исследований силами четырёх экспертов. Это наблюдение создаёт теоретическую базу для выработки требований к алгоритмам искусственного интеллекта, предназначенным для использования в диагностике заболеваний посредством разметки подозрительных структур на КТ-снимках и направления внимания врача-рентгенолога. Помимо этого, полученные в работе результаты позволяют обосновать модель проектов многопользовательской разметки наборов данных (crowd-tagging), при которых рост количества разметчиков повлечёт снижение согласованности и одновременный рост качества конечного продукта, обеспечиваемый за счёт арбитража.

ДОПОЛНИТЕЛЬНО

Источник финансирования. Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

Конфликт интересов. Авторы данной статьи подтвердили отсутствие конфликта интересов, о котором необходимо сообщить.

Вклад авторов. Все 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: Н.С. Кульберг — дизайн набора данных, концептуализация исследования, подготовка и редактирование текста статьи; Р.В. Решетников — статистический анализ, написание текста статьи; В.П. Новик — подготовка набора данных, написание скриптов для сбора данных, статистический анализ; А.Б. Елизаров — подготовка набора данных, написание скриптов для сбора данных; М.А. Гусев — подготовка набора данных, написание скриптов для сбора данных; В.А. Гомболевский — концептуализация исследования, дизайн набора данных; А.В. Владзимирский — концептуализация исследования, редактирование текста статьи; С.П. Морозов — дизайн набора данных, концептуализация и финансирование исследования.

Благодарности. Авторы выражают благодарность Черниной Валерии Юрьевне за методические консультации, а также всем врачам-рентгенологам, принимавшим участие в разметке набора данных.

Funding source. This study was not supported by any external sources of funding.

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. The largest contributions are as follows: N.S. Kulberg — dataset design, conceptualization of the study, preparation and editing of the text of the article; R.V. Reshetnikov — statistical analysis, writing of the manuscript; V.P. Novik – dataset preparation, software development for data processing, statistical analysis; A.B. Elizarov — dataset preparation, software development for data processing; M.A. Gusev — dataset preparation, software development for data processing; V.A. Gombolevskiy — conceptualization of the study, dataset design; A.V. Vladzymyrskyy — conceptualization of the study, editing of the text of the article; S.P. Morozov — dataset design, conceptualization and funding of the study.

Acknowledgments. The authors express their gratitude to Valeria Chernina for methodological consultations and all radiologists who participated in the dataset markup and annotation.

СПИСОК ЛИТЕРАТУРЫ

1. Морозов С.П., Кульберг Н.С., Гомболевский В.А., и др. Датасет радиологии Москвы CTLungCa-500. 2018. Режим доступа: https://mosmed.ai/datasets/ct_lungcancer_500/. Дата обращения: 11.02.2021.

2. Morozov S.P., Gombolevskiy V.A., Elizarov A.B., et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT Scans // Comput Methods Programs Biomed. 2021. Vol. 206. P. 106111. doi: 10.1016/j.cmpb.2021.106111

3. Kulberg N.S., Gusev M.A., Reshetnikov R.V., et al. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images // Heal Care Russ Fed. 2020. Vol. 64, N 6. P. 343–350. doi: 10.46563/0044-197X-2020-64-6-343-350
4. Hessel S.J., Herman P.G., Swensson R.G. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost // Radiology. 1978. Vol. 127, N 3. P. 589–594. doi: 10.1148/127.3.589
5. Herman P.G., Hessel S.J. Accuracy and its relationship to experience in the interpretation of chest radiographs // Invest Radiol. 1975. Vol. 10, N 1. P. 62–67. doi: 10.1097/00004424-197501000-00008
6. MacMahon H., Naidich D.P., Goo J.M., et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017 // Radiology. 2017. Vol. 284, N 1. P. 228–243. doi: 10.1148/radiol.2017161659
7. Gerke O., Vilstrup M.H., Segtnan E.A., et al. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation // BMC Med Imaging. 2016. Vol. 16, N 1. P. 54. doi: 10.1186/s12880-016-0159-3
8. Rasheed K., Rabinowitz Y.S., Remba D., Remba M.J. Interobserver and intraobserver reliability of a classification scheme for corneal topographic patterns // Br J Ophthalmol. 1998. Vol. 82, N 12. P. 1401–1406. doi: 10.1136/bjo.82.12.1401
9. Van Riel S.J., Sánchez C.I., Bankier A.A., et al. Observer variability for classification of pulmonary nodules on low-dose ct images and its effect on nodule management // Radiology. 2015. Vol. 277, N 3. P. 863–871. doi: 10.1148/radiol.2015142700
10. Wickham H., François R., Henry L., Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.4. 2021.
11. Gamer M, Lemon J, Fellows I, Singh P. irr: Various Coefficients of Interrater Reliability and Agreement. 2019.
12. Wickham H. ggplot2: elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. 260 p.
13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. Режим доступа: http://www.r-project.org/index.html. Дата обращения: 11.02.2021.
14. Van Rossum G., Drake F.L. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA; 2009.
15. Ardila D., Kiraly A.P., Bharadwaj S., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography // Nat Med. 2019. Vol. 25, N 6. P. 954–961. doi: 10.1038/s41591-019-0447-x
16. Peters R., Heuvelmans M., Brinkhof S., et al. Prevalence of pulmonary multi-nodularity in CT lung cancer screening. 2015.
17. Creative Research Systems. The survey systems: Sample size calculator. 2012.
18. Hugo G.D., Weiss E., Sleeman W.C., et al. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer // Med Phys. 2017. Vol. 44, N 2. P. 762–771. doi: 10.1002/mp.12059
19. Bakr S., Gevaert O., Echegaray S., et al. A radiogenomic dataset of non-small cell lung cancer // Sci Data. 2018. Vol. 5. P. 180202. doi: 10.1038/sdata.2018.202
20. Armato S.G., McLennan G., Bidaut L., et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans // Med Phys. 2011. Vol. 38, N 2. P. 915–931. doi: 10.1118/1.3528204

REFERENCES

1. Morozov SP, Kulberg NS, Gombolevsky VA, et al. Moscow Radiology Dataset CTLungCa-500. 2018. (In Russ). Available from: https://mosmed.ai/datasets/ct_lungcancer_500/
2. Morozov SP, Gombolevskiy VA, Elizarov AB, et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT Scans. *Comput Methods Programs Biomed*. 2021;206:106111. doi: 10.1016/j.cmpb.2021.106111
3. Kulberg NS, Gusev MA, Reshetnikov RV, et al. Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images. *Heal Care Russ Fed*. 2020;64(6):343–350. doi: 10.46563/0044-197X-2020-64-6-343-350
4. Hessel SJ, Herman PG, Swensson RG. Improving performance by multiple interpretations of chest radiographs: effectiveness and cost. *Radiology*. 1978;127(3):589–594. doi: 10.1148/127.3.589
5. Herman PG, Hessel SJ. Accuracy and its relationship to experience in the interpretation of chest radiographs. *Invest Radiol*. 1975;10(1):62–67. doi: 10.1097/00004424-197501000-00008
6. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. *Radiology*. 2017;284:228–243. doi: 10.1148/radiol.2017161659
7. Gerke O, Vilstrup MH, Segtnan EA, et al. How to assess intra- and inter-observer agreement with quantitative PET using variance component analysis: a proposal for standardisation. *BMC Med Imaging*. 2016;16(1):54. doi: 10.1186/s12880-016-0159-3
8. Rasheed K, Rabinowitz YS, Remba D, Remba MJ. Interobserver and intraobserver reliability of a classification scheme for corneal topographic patterns. *Br J Ophthalmol*. 1998;82(12):1401–1406. doi: 10.1136/bjo.82.12.1401
9. Van Riel SJ, Sánchez CI, Bankier AA, et al. Observer variability for classification of pulmonary nodules on low-dose ct images and its effect on nodule management. *Radiology*. 2015;277(3):863–871. doi: 10.1148/radiol.2015142700
10. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.4. 2021.
11. Gamer M, Lemon J, Fellows I, Singh P. irr: Various Coefficients of Interrater Reliability and Agreement. 2019.
12. Wickham H. ggplot2: elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. 260 p.
13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: http://www.r-project.org/index.html
14. Van Rossum G, Drake FL. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA; 2009.
15. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. 2019;25(6):954–961. doi: 10.1038/s41591-019-0447-x
16. Peters R, Heuvelmans M, Brinkhof S, et al. Prevalence of pulmonary multi-nodularity in CT lung cancer screening. 2015.

17. Creative Research Systems. The survey systems: Sample size calculator. 2012.
18. Hugo GD, Weiss E, Sleeman WC, et al. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Med Phys*. 2017;44(2):762–771. doi: 10.1002/mp.12059
19. Bakr S, Gevaert O, Echegaray S, et al. A radiogenomic dataset of non-small cell lung cancer. *Sci Data*. 2018;5:180202. doi: 10.1038/sdata.2018.202
20. Armato SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans. *Med Phys*. 2011;38(2):915–931. doi: 10.1118/1.3528204

ОБ АВТОРАХ

*** Кульберг Николай Сергеевич,** к.ф.-м.н.;
адрес: Россия, 127051, Москва, ул. Петровка, д. 24;
ORCID: https://orcid.org/0000-0001-7046-7157;
eLibrary SPIN: 2135-9543; e-mail: kulberg@npcmr.ru

Решетников Роман Владимирович, к.ф.-м.н.;
ORCID: https://orcid.org/0000-0002-9661-0254;
eLibrary SPIN: 8592-0558; e-mail: reshetnikov@fbb.msu.ru

Новик Владимир Петрович;
ORCID: https://orcid.org/0000-0002-6752-1375;
eLibrary SPIN: 2251-1016; e-mail: v.novik@npcmr.ru

Елизаров Алексей Борисович, к.ф.-м.н.;
ORCID: https://orcid.org/0000-0003-3786-4171;
eLibrary SPIN: 7025-1257; e-mail: a.elizarov@npcmr.ru

Гусев Максим Александрович;
ORCID: https://orcid.org/0000-0001-8864-8722;
eLibrary SPIN: 1526-1140; e-mail: m.gusev@npcmr.ru

Гомболевский Виктор Александрович, к.м.н.;
ORCID: https://orcid.org/0000-0003-1816-1315;
eLibrary SPIN: 6810-3279; e-mail: g_victor@mail.ru

Владзимирский Антон Вячеславович, д.м.н., профессор;
ORCID: https://orcid.org/0000-0002-2990-7736;
eLibrary SPIN: 3602-7120; e-mail: a.vladzimirsky@npcmr.ru

Морозов Сергей Павлович, д.м.н., профессор;
ORCID: https://orcid.org/0000-0001-6545-6170;
eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru

AUTHORS' INFO

*** Nikolas S. Kulberg,** Cand. Sci. (Phys.-Math.);
address: 24 Petrovka str., 109029, Moscow, Russia;
ORCID: https://orcid.org/0000-0001-7046-7157;
eLibrary SPIN: 2135-9543; e-mail: kulberg@npcmr.ru

Roman V. Reshetnikov, Cand. Sci. (Phys.-Math.);
ORCID: https://orcid.org/0000-0002-9661-0254;
eLibrary SPIN: 8592-0558; e-mail: reshetnikov@fbb.msu.ru

Vladimir P. Novik;
ORCID: https://orcid.org/0000-0002-6752-1375;
eLibrary SPIN: 2251-1016; e-mail: v.novik@npcmr.ru

Alexey B. Elizarov, Cand. Sci. (Phys.-Math.);
ORCID: https://orcid.org/0000-0003-3786-4171;
eLibrary SPIN: 7025-1257; e-mail: a.elizarov@npcmr.ru

Maxim A. Gusev;
ORCID: https://orcid.org/0000-0001-8864-8722;
eLibrary SPIN: 1526-1140; e-mail: m.gusev@npcmr.ru

Victor A. Gombolevskiy, MD, Cand. Sci. (Med.);
ORCID: https://orcid.org/0000-0003-1816-1315;
eLibrary SPIN: 6810-3279; e-mail: g_victor@mail.ru

Anton V. Vladzymyrskyy, MD, Dr. Sci. (Med.), Professor;
ORCID: https://orcid.org/0000-0002-2990-7736;
eLibrary SPIN: 3602-7120; e-mail: a.vladzimirsky@npcmr.ru

Sergey P. Morozov, MD, Dr. Sci. (Med.), Professor;
ORCID: https://orcid.org/0000-0001-6545-6170;
eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru

* Автор, ответственный за переписку / Corresponding author