

ЭТАЛОННЫЕ МЕДИЦИНСКИЕ ДАТАСЕТЫ (MosMedData) ДЛЯ НЕЗАВИСИМОЙ ВНЕШНЕЙ ОЦЕНКИ АЛГОРИТМОВ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ДИАГНОСТИКЕ

© Н.А. Павлов, А.Е. Андрейченко, А.В. Владзимирский, А.А. Ревазян, Ю.С. Кирпичев, С.П. Морозов

Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы, Москва, Российская Федерация

АННОТАЦИЯ

В статье описывается оригинальный подход к формированию аннотированных медицинских датасетов для проверки диагностических решений, основанных на технологиях искусственного интеллекта. Описаны 4 этапа формирования датасета — планирование, отбор исходных данных, разметка и верификация, документирование. Приведены примеры созданных по описанной методике датасетов. Методика является масштабируемой и универсальной, а значит, может быть использована в других областях медицины и здравоохранения, которые подлежат автоматизации и развитию с помощью технологий искусственного интеллекта и технологий больших данных.

Ключевые слова: искусственный интеллект; медицинские данные; датасет; разметка; машинное обучение; большие данные; верификация.

Как цитировать

Павлов Н.А., Андрейченко А.Е., Владзимирский А.В., Ревазян А.А., Кирпичев Ю.С., Морозов С.П. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике // *Digital Diagnostics*. 2021; 2(1):XX–XX. DOI: <https://doi.org/10.17816/DD60635>

REFERENCE MEDICAL DATASETS (MosMedData) FOR INDEPENDENT EXTERNAL VALIDATION OF ARTIFICIAL INTELLIGENCE ALGORITHMS IN MEDICAL DIAGNOSTICS

© Nikolay A. Pavlov, Anna E. Andreychenko, Anton V. Vladzimirskyy, Anush A. Revazyan, Yury S. Kirpichev, Sergey P. Morozov

Research and Practical Clinical Center of Diagnostics and Telemedicine Technologies, Department of Healthcare of Moscow, Moscow, Russian Federation

ABSTRACT

The article depicts an original approach to creation of annotated medical datasets for validation of diagnostic systems based on artificial intelligence. This approach consists of four steps: planning, selection of source data, annotation and verification, and documenting. Examples of medical imaging datasets created in accordance with this approach are presented. The methodology is believed to be universal and scalable, therefore it could be applied to other medical fields other than radiology in which artificial intelligence and big data technologies could potentially benefit medical specialists.

Keywords: artificial intelligence; medical data; dataset; annotation; machine learning; big data; validation.

To cite this article

Pavlov NA, Andreychenko AE, Vladzimirsky AV, Revazyan AA, Kirpichev YS, Morozov SP. Reference medical datasets (MosMedData) for independent external validation of artificial intelligence algorithms in medical diagnostics. *Digital Diagnostics*. 2021; 2(1):XX–XX. DOI: <https://doi.org/10.17816/DD60635>

Статья поступила 11.02.2021
Received: 11.02.2021

Принята к печати 23.03.2021
Accepted: 23.03.2021

Опубликована 30.03.2021
Published: 30.03.2021

Список сокращений

Гб, Мб, Тб	— объём цифрового хранилища: гигабайт, мегабайт, терабайт
Датасет (Dataset; Data set)	— структурированный набор информации, объединённой по определённым логическим принципам, пригодный для машинной обработки компьютерными методами анализа данных. Датасет — комплексное понятие, характеризующееся 4 основными этапами: наличие содержимого (наблюдения, значения, записи, файлы и др.); наличие цели (например, база знаний, использование для определённой задачи); наличие группировки (агрегация и организация содержимого в наборы, коллекции и др.); наличие связанности (отношение к субъекту, интегрированность, логическая коллекция содержимого и т.д.)
ЕМИАС	— Единая медицинская информационно-аналитическая система г. Москвы
ЕРИС	— Единый радиологический информационный сервис г. Москвы
ИИ (искусственный интеллект)	— наука и технология создания интеллектуальных компьютерных программ, способных выполнять задачи, для решения которых, как правило, требуется человеческий интеллект
КТ	— компьютерная томография
КТ 0–4	— классификация КТ-признаков COVID-19, разработанная Научно-практическим клиническим центром диагностики и телемедицинских технологий Департамента здравоохранения г. Москвы в 2020 г. КТ0 — норма и отсутствие КТ-признаков вирусной пневмонии. КТ1 — зоны уплотнения по типу матового стекла; вовлечение паренхимы лёгкого $\leq 25\%$. КТ2 — зоны уплотнения по типу матового стекла; вовлечение паренхимы лёгкого 25–50%. КТ3 — зоны уплотнения по типу матового стекла и консолидации; вовлечение паренхимы лёгкого 50–75%. КТ4 — диффузное уплотнение лёгочной ткани по типу матового стекла и консолидации в сочетании с ретикулярными изменениями; вовлечение паренхимы лёгкого $> 75\%$
МИС	— медицинская информационная система
ММГ	— маммография
НДКТ	— низкодозная компьютерная томография
ОГК	— органы грудной клетки
РГ	— рентгенография

ФЛГ	— флюорография
COVID-19	— инфекционное заболевание, вызванное вирусом SARS-CoV-2, распространение которого в 2020 г. было охарактеризовано Всемирной организацией здравоохранения как пандемия. Согласно Международной классификации болезней 10-го пересмотра, кодируется как U07.1 или U07.2 в зависимости от наличия/отсутствия лабораторной идентификации вируса соответственно
DICOM (Digital Imaging and Communications in Medicine)	— медицинский отраслевой стандарт создания, хранения, передачи и визуализации цифровых медицинских изображений и документов обследованных пациентов
MeSH (Medical Subject Headings)	— тезаурус, содержащий ключевые медицинские термины, используемый для индексации, каталогизации и поиска статей в англоязычной текстовой базе данных медицинских и биологических публикаций, созданной Национальным центром биотехнологической информации США (PubMed)
README	— <i>англ.</i> «прочитай меня», устоявшееся название документа, сопровождающего исполняемый код, базу данных или другие программные продукты, обычно содержащий основную информацию о файлах, находящихся в той же директории.
SARS-CoV-2	— оболочечный одноцепочечный (+)РНК-вирус рода <i>Betacoronavirus</i>

ОБОСНОВАНИЕ

Прогресс в технологиях искусственного интеллекта (ИИ) и их практическое применение в разных сферах, в частности в медицине, демонстрирует потенциал технологий в таких решениях, как автоматизированные системы диагностики, системы распознавания неструктурированных медицинских записей и понимание естественного языка, системы анализа и предсказания событий, системы автоматической классификации и сверки информации, а также автоматические чат-боты для поддержки пациентов [1]. В связи с бурным развитием глубокого машинного обучения и связанного с ним компьютерного распознавания образов и паттернов на изображениях значительное внимание среди всех областей применения автоматизированных систем диагностики уделяется анализу медицинских изображений, в частности лучевых исследований [2].

Для практического здравоохранения задача автоматизации процессов диагностики является одной из первоочередных в условиях старения населения, повышения доступности и, соответственно, количества проводимых диагностических процедур, которые не компенсируются ростом числа квалифицированного персонала, необходимого для обеспечения должной интерпретации результатов и, как следствие, оказания своевременной медицинской помощи. В особенности данная проблема остро стоит в лучевой диагностике [3], в основе которой лежит визуальный анализ изображений врачом. Для самых современных методов в лучевой диагностике количество двумерных изображений на пациента, подлежащих интерпретации, может достигать 1000 и более. В связи с этим именно лучевая диагностика в настоящий момент является областью активного развития технологий глубокого обучения, входящей в состав понятия ИИ, для создания систем компьютерного зрения, обеспечивающих автоматизацию интерпретации медицинских снимков. Отличительной особенностью глубокого обучения от других методов машинного обучения является то, что точность,

надёжность и практическая ценность создаваемых моделей напрямую зависит от количества и качества используемых данных в процессах обучения, валидации (тонкой настройки) и тестирования [4].

Именно поэтому одним из основных барьеров развития решений на основе ИИ в медицинской диагностике является отсутствие выверенных (свободных от неполных и ошибочных) и качественных (унифицированных, подготовленных для автоматической обработки машинным способом) наборов данных [5]. Аннотированные датасеты [6] необходимы не только для «обучения» ИИ, в частности для машинного обучения компьютерных нейронных сетей, но и для тестирования сетей, обученных на других данных.

Требования, предъявляемые к датасетам, не позволяют использовать простую выгрузку из медицинской информационной системы, а обязывают проводить целый ряд манипуляций с данными, прежде чем они станут аннотированным датасетом, пригодным для эффективного использования моделями ИИ. Отличие медицинских данных от данных в других областях, в которых активно используется машинное обучение (например, банковская сфера, сфера услуг), заключается в исторически сложившейся культуре ведения медицинских карт, отсутствии или минимальном структурировании и ограниченном сопоставлении различных исследований одного пациента между собой. На настоящий момент литература о подготовке медицинских датасетов представлена только небольшим количеством публикаций [7–9]. Данной публикацией авторы преследуют цель расширить понимание проблемы и особенностей подготовки датасетов на основе медицинских данных среди медицинских специалистов, касающихся или задействованных в процесс разработки или тестирования ИИ, а также программистов и специалистов по работе с данными (data science) для улучшения процесса независимой оценки алгоритмов на основе ИИ.

В настоящей статье представлен унифицированный подход (методика) к формированию датасетов для объективного (насколько представляется возможным в каждом конкретном случае) тестирования решений, использующих технологии ИИ, в области лучевой диагностики. По ходу описания этапов предлагаемой нами методики мы приводим практические примеры датасетов, разработанных нами в период с сентября 2019 по декабрь 2020 г., с использованием данных отделений лучевой диагностики медицинских амбулаторных и стационарных учреждений г. Москвы, депонированных в Едином радиологическом информационном сервисе (ЕРИС ЕМИАС) [10]. Основные принципы, описанные в статье, могут быть использованы для формирования медицинских датасетов в других областях медицины.

МЕТОДЫ И РЕЗУЛЬТАТЫ

Датасет (набор данных) отличается от простого сбора медицинских данных тем, что он наделён особыми свойствами — унификацией и структурированностью данных; отсутствием грубых неточностей, ошибочных исследований; наличием дополнительной информации (категории и значения признаков или характеристик элементов данных); наличием сопроводительной документации. В Российской Федерации датасет приравнивается к базе данных и подлежит добровольной государственной регистрации в качестве результата интеллектуальной деятельности. В зарубежной практике датасеты зачастую публикуются не только как доступные для скачивания наборы данных, но и как научные публикации в журналах. Каждый датасет уникален не только составом исследований, но также способом их классификации и подходами к разметке, а сам процесс создания датасета носит исследовательский характер. Даже при наличии структурированной методики формирования датасетов от неё возможны отхождения,

исключения и изменения определённых этапов в зависимости от назначения формируемого датасета.

Весь процесс можно разделить на 4 крупных этапа — планирование, отбор исходных данных, разметка и верификация, документирование (рис. 1).

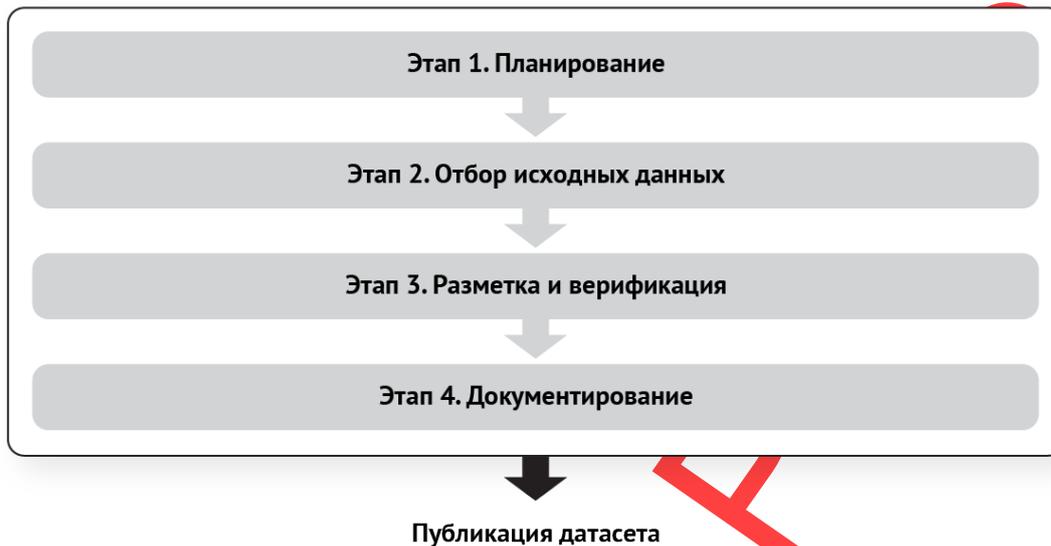


Рис. 1. Этапы формирования медицинского датасета.

1. Этап планирования

Подготовка датасета, как и научного исследования, начинается с этапа планирования:

- формулировка клинической и/или практической задачи в области медицины, которая (потенциально) подвержена автоматизации с помощью интеллектуальных систем;
- составление перечня признаков и/или характеристик исходных данных, информацию о которых планируется получать от интеллектуальной системы в процессе решения поставленной задачи и по которым возможно оценить корректность принятого системой решения;
- определение методики верификации значений выбранных признаков и/или характеристик у элементов формируемого набора данных;
- определение источников данных;
- описание планируемых шагов по обезличиванию данных;
- определение критериев включения, невключения и исключения исследования из датасета;
- определение значимых характеристик данных, необходимых для оценки не только точности, но и пределов надёжности и масштабируемости интеллектуальной системы.

Постановка клинической задачи — одна из наиболее ответственных задач, стоящих перед создателем датасета. Недостаточное внимание к ней приводит к внезапно всплывающим вопросам как в процессе подготовки датасета, так и уже при внедрении диагностического алгоритма на основе ИИ в клиническую практику (рис. 2).



Рис. 2. Взаимосвязь клинической задачи, датасета и успеха при внедрении решения на основе искусственного интеллекта (ИИ) в рутинную клиническую практику.

Для того чтобы задача соответствовала классу задач, в котором ИИ зарекомендовал себя как перспективная технология, и одновременно с этим имела важную социоэкономическую составляющую с точки зрения клинических специалистов, в определении задачи должна участвовать рабочая группа из профессионалов различных профилей, а именно врачей-клиницистов, специалистов по обработке медицинских данных, инженеров-исследователей (занимающихся машинным обучением или проводящих валидацию решений на основе ИИ), а также администраторов, обеспечивающих доступ и выгрузку исходных данных.

Клиническая задача должна позволить создателям датасета ответить на следующие вопросы:

- 1) какие модальности, какие процедуры, какую клиническую, демографическую и подобного рода информацию следует принимать на вход алгоритму для её решения, и что принимать за одну единицу данных;
- 2) какие признаки должны определяться с помощью технологий ИИ;
- 3) к каким нозологиям или группам нозологий относятся искомые признаки;
- 4) как решение задачи помогает клиническому специалисту;
- 5) какое количество единиц данных необходимо и достаточно для цели применения создаваемого датасета (валидация ИИ, машинное обучение и т.д.).

Важным критерием для подбора количества единиц данных и характеристик исследования является цель применения набора данных по отношению к ИИ. Можно привести следующую классификацию датасетов по их назначению:

1) общие наборы:

- селф-тест для проверки ИИ на техническое соответствие;
- клинический тест для оценки метрик точности работы и продуктивности ИИ;
- «дообучение» для донастройки уже обученной модели ИИ;
- машинное обучение для обучения новых моделей, лежащих в основе ИИ, и решения новых клинических задач;

2) специализированные наборы:

- динамические наборы для оценки изменений во времени (с привязкой нескольких элементов данных к одному субъекту);

- технологические дефекты для оценки стабильности и надёжности работы диагностических решений на основе ИИ при попытке анализа дефектного исследования.

Количество единиц исследований, требуемое для селф-теста, обычно исчисляется единичными примерами для каждого типа или модели диагностического устройства; количество единиц исследований в динамических наборах и датасетах для клинического теста обычно составляет несколько десятков; датасеты для обучения и «дообучения» могут содержать от нескольких сотен до нескольких десятков тысяч исследований. Указанные количества носят грубый оценочный характер и могут варьировать в широких рамках в зависимости от наличия исследований в источнике данных, сложности клинической задачи, детальности и трудоёмкости аннотирования данных и других факторов. Практические примеры различных датасетов, используемых авторским коллективом, с указанием количества исследований (в скобках) приведены на рис. 3.

После того как клиническая задача определена, из неё логическим образом вытекают критерии, по которым интеллектуальная система принимает решение, отнести ли то или иное исследование или найденную на изображении область к группе интереса (базовые диагностические требования к работе ИИ). Диагностические требования включают формальное описание искомых признаков на исследовании, а также составляют перечень признаков и/или характеристик, на основе которых в дальнейшем будет проводиться разметка данных в датасете. Данная информация позволяет разработчикам точнее настроить решения для определения требуемых признаков, а специалистам по подготовке датасета — составить инструкцию по разметке и верификации данных.

Ключевое значение для ценности и значимости полученного анализа систем на основе технологий искусственного интеллекта с применением датасета имеет баланс классов, а именно, в какой пропорции распределены исследования в датасете, относящиеся к различным признакам и/или характеристикам. В самом простом случае для оценки работы интеллектуальных диагностических систем, предоставляющих дихотомические ответы, применяется равное деление между двумя категориями (например, 50% исследований с признаками патологии согласно базовым диагностическим требованиям к работе ИИ и 50% исследований без признаков патологии). В более сложных случаях разделение между несколькими классами может быть неравным и зависеть от метода сравнения, который планируется применять впоследствии.

Исследования, разделённые на классы по значимому признаку, могут иметь и другие различия — как в клиническом (например, преобладание пациентов женского пола в категории с признаками патологии, обусловленное половозрастной картиной заболеваемости), так и техническом (например, искусственное смещение выборки за счёт предпочтения направлять пациентов с уже выявленной патологией на исследование, выполняемое на устройстве с большей разрешительной способностью) аспекте. Во избежание систематической ошибки необходимо выявить признаки, хоть и незначимые для решения клинической задачи, но влияющие на работу диагностической интеллектуальной системы, а при отборе исследований в датасет — стремиться представить различные примеры в каждом из классов. Вопрос по систематизации таких признаков и характеристик для широкого спектра клинических задач (то есть вопрос баланса классов в датасетах) остаётся открытым и активно исследуется в настоящее время [9].

В конце этапа планирования определяются *источники исходных данных, а также критерии включения, невключения и исключения* исследований из датасета.

Для создания наиболее репрезентативного датасета источники данных должны по возможности быть либо теми же, либо релевантными тем информационным системам,

в которых в дальнейшем планируется внедрение решений на основе ИИ. Для московского здравоохранения пример такого источника — ЕРИС ЕМИАС, объединяющий системы хранения отделений лучевой диагностики десятков амбулаторных и стационарных медицинских учреждений г. Москвы.

Критерии включения и невключения чаще определяются клинической и/или практической задачей, тогда как критерии исключения обычно дополняются в ходе работы с первичными данными, так как находятся те или иные критерии, негативно сказывающиеся на структурированности и унификации набора данных. Данные критерии могут носить как медицинский (например, возраст от 18 до 99 лет; наличие сохранной структуры искомого органа, и т.д.), так и технический (фильтр КТ — мягкотканый, ядро свёртки — FC51, и т.д.) характер. Унификация данных необходима для надёжной работы инструментов оценки работы решения на основе ИИ (см. п. 3 «Разметка и верификация»).

Для примера приведён датасет «MosMedData: результаты исследований ультранизкодозной компьютерной томографии с очагами в лёгких»¹. Цель создания базы данных — обеспечение возможности проверки готовности автоматизированных систем (в том числе использующих ИИ) к работе в ЕРИС ЕМИАС. Клиническая задача — поиск и определение лёгочных очагов в скрининге рака лёгкого. Для датасета были выбраны обезличенные исследования компьютерной томографии (КТ) в формате DICOM, проведённые в особом режиме ультранизкодозной томографии (эффективная доза лучевой нагрузки менее 1 мЗв при повышенном напряжении — 135 кВ). За одну единицу взято одно исследование КТ органов грудной клетки (ОГК), отвечающее указанным ниже критериям.

А. Критерии включения:

1. Возраст пациента старше 55 лет и моложе 75 лет.
2. Стаж курения более 30 пачка/лет (не менее 1 пачки в день в течение 30 лет или 2 пачек в день в течение 15 лет, и т.д.).
3. Курение в настоящий момент или отказ от курения не более 15 лет назад.
4. Исследование проведено в режиме ультранизкодозной КТ в первом раунде скрининга рака лёгкого.

Б. Критерии невключения:

1. Выявленный рак лёгкого в течение 2 лет после проведения первого раунда скрининга рака лёгкого с помощью ультранизкодозной КТ.
2. В анамнезе — рак лёгкого и/или операции на лёгком (не включая чрескожной биопсии лёгкого).
3. В анамнезе — онкологическое заболевание, выявленное менее 5 лет назад, за исключением рака кожи и рака шейки матки *in situ*.
4. Наличие выраженной патологии сердечно-сосудистой, иммунной, дыхательной, эндокринной системы, а также ожидаемый срок жизни менее 5 лет.
5. Острое заболевание дыхательной системы.
6. Лечение антибиотиками за последние 12 недель.
7. Наличие кровохаркания или потеря веса более 10 кг за последний год.

В. Критерии исключения:

1. Отсутствие лёгочных очагов в первом раунде московского скрининга рака лёгкого.

¹ Морозов С.П., Гончар А.П., Николаев А.Е. и др. MosMedData: результаты исследований ультранизкодозной компьютерной томографии с очагами в лёгких (база данных). Свидетельство о государственной регистрации базы данных № 2020622727 от 21.12.2020.

Целевое значение количества исследований в итоговом датасете — 300 — достаточно для тестирования автоматизированных систем диагностики на основе ИИ (итоговое количество — 312 единиц).

2. Этап отбора исходных данных

После получения доступа к источнику исходных данных наступает этап отбора исходных («сырых») данных. Подход к их получению (выгрузке) зависит от источника и способа хранения данных.

Медицинские данные могут накапливаться как в ходе рутинного диагностического процесса в медицинском учреждении (MeSH: Routinely Collected Health Data), так и сбора данных непосредственно пациентом и/или его родственниками и социальными работниками (MeSH: Patient Generated Health Data), а также в результате целенаправленного сбора данных, например в процессе клинического исследования. Данные, собранные рутинно, обычно имеют широкую вариативность в параметрах и позволяют создать наиболее репрезентативный датасет. При использовании данных, собранных в ходе клинического исследования, обращают на себя внимание (1) критерии включения, невключения и исключения субъектов в исследование, заданные его дизайном и ограничивающие возможности при подготовке датасета, а также (2) объём данных, ограниченный мощностью исследования.

Оцифровка документов, не являющихся первично электронными, практически не имеет смысла; документы, хранящиеся на внешних носителях, часто недостаточно структурированы, а оцифровка и/или перенос данных с иных носителей могут оказаться затратными (например, перенос базы данных лучевых исследований, хранящихся на компакт-дисках). Наличие медицинской информационной системы (МИС; MeSH: Health Information Systems) упрощает выгрузку, поскольку позволяет применять фильтры и отбирать необходимые исследования по таким признакам, как, например, наличие того или иного исследования или диагноза. Однако не для всех клинических задач необходимая информация содержится в электронных историях болезней: списки пациентов, подходящих под критерии клинической задачи, могут быть сформированы отдельно от МИС, и отбор исследований для пациентов из этих списков занимает значительное время.

Общие принципы отбора «сырых» данных заключаются:

- 1) в выборе как можно большего спектра исследований интересующей модальности и процедуры;
- 2) сохранении необходимого для решения клинической задачи количества сопроводительной информации (включая текстовые документы, описывающие результаты исследования, клинический диагноз пациента, которым закончился медицинский случай, и др.);
- 3) по возможности, обезличивании исследований «на месте», не покидая информационного контура учреждения, в котором происходит отбор данных.

На этапе отбора применяются также критерии включения и невключения исследования в будущий датасет. Эта операция может проводиться как непосредственно при отборе исследований в МИС, так и сразу после выгрузки (уже вне пределов информационного контура медицинской организации). Необходимо учитывать, что этот шаг может повлечь за собой уменьшение датасета в несколько десятков раз.

При отборе исследований следует также помнить о соблюдении баланса классов, определённом на этапе 1.

Например, для датасета «MosMedData: результаты исследований ультранизкодозной компьютерной томографии с очагами в лёгких», упомянутого в описании этапа 1, этап 2 может состоять из следующих шагов:

- 1) отбор пациентов в МИС, которым было проведено исследование низкодозной КТ ОГК с целью скрининга злокачественных образований лёгких;
- 2) анализ электронных медицинских карт отобранных пациентов (анамнез жизни, анамнез предыдущих заболеваний, данные предыдущих исследований) для выбора пациентов в соответствии с критериями включения и исключения, сформированными на этапе 1;
- 3) принятие решения о включении исследований в датасет в соответствии с желаемым балансом классов.

3. Этап разметки и верификации

Разметка — процесс определения значения признаков или характеристик у элемента данных в датасете. На основе разметки становится возможным провести классификацию элементов, отнести их к той или иной группе. Для разметки может использоваться как уже имеющаяся на момент отбора исходных данных информация (ретроспективная разметка), так и разметка, сделанная специалистом с медицинским образованием и/или опытом работы уже после этапа отбора (проспективная разметка) [9].

Для ретроспективной разметки могут быть использованы данные сопроводительных документов (такими, например, являются тексты заключений для результатов инструментальных исследований), МИС, электронных медицинских карт и др. Примером могут служить и метаданные, генерируемые автоматически прибором при проведении исследования и хранящиеся в исходных данных. Очевидное преимущество ретроспективной разметки заключается в том, что на неё требуется значительно меньше времени со стороны медицинских специалистов, так как большую часть подготовительной работы выполняет специалист по работе с данными.

Проспективная разметка предполагает активное вовлечение медицинских специалистов в процесс «насыщения» датасета дополнительной информацией, например, позволяющей эффективно разделить элементы датасета на классы и категории. В лучевой диагностике под разметкой чаще всего понимают классификацию исследований по классам (наличие или отсутствие рентгенологических признаков выбранного заболевания), а также графическое обозначение области интереса, соответствующей искомым признакам (например, очаги демиелинизации при рассеянном склерозе на МР-изображениях головного мозга). Степень вовлечения может быть разделена на более и менее затратную: в первом случае экспертам предлагается обвести контур области интереса, во втором — обозначить её координаты простой геометрической фигурой.

В тех случаях, когда экспертное мнение является наиболее значимым фактором при определении значений признаков или характеристик данных, разумным решением будет проведение одновременного чтения исследования двумя независимыми экспертами. В случае несогласованности между двумя экспертами спорные исследования направляются третьему, более квалифицированному эксперту (по практическому опыту, наличию степени либо другим критериям). Исследования, оставшиеся спорными после прочтения тремя экспертами, могут быть признаны спорными и исключены из датасета. Из нашей практики подготовки датасета, состоящего из 100 исследований КТ ОГК с признаками различных патологий дыхательной системы, до 1/4 исследований могут оказаться спорными после двух независимых чтений; до 4% исследований могут остаться спорными после чтения третьим, более квалифицированным экспертом (медицинский стаж которого составляет более 5 лет).

Прежде чем приступать к проспективной разметке, необходимо определить объём исследований на одного специалиста, критерии размечаемых признаков, программное обеспечение, позволяющее осуществить текстовое, графическое или иное обозначение

искомых признаков, подготовить инструкцию врача-разметчика. В процесс подготовки такой инструкции по возможности должна быть вовлечена та же рабочая группа, которая определяла клиническую задачу на этапе планирования.

Верификация разметки обеспечивает степень «доверия» разметке со стороны разработчиков или специалистов по оценке работы интеллектуальных систем. Верификацию разметки можно разделить:

- на низкую (факт наличия находки) — на основе документации;
- среднюю (классификация находок) — на основе экспертного мнения;
- высокую (подтверждённый диагноз) — на основе результатов более чувствительного метода исследования либо динамического наблюдения (повторного проведения того же метода через определённый временной промежуток).

Классификация типов разметки представлена на рис. 4. Часть датасета может иметь один класс, тогда как другая — другой класс; в одном датасете допускается сочетание ретро- и проспективной разметки. Важной составляющей процесса разметки является её правильное документирование в сопроводительной документации (см. п. 4 «Этап документирования»).

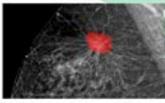
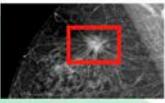
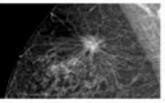
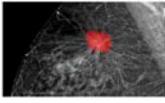
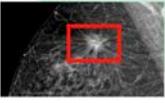
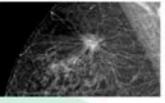
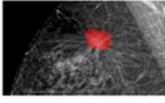
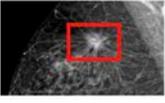
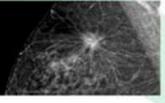
Классы разметки		A		B		C	
		ПРОСПЕКТИВНАЯ		ПРОСПЕКТИВНАЯ		РЕТРОСПЕКТИВНАЯ	
ЦЕННОСТЬ		Пиксельная маска		Координаты области		Метаданные	
1	Подтвержденный диагноз		РМЖ (ДАННЫЕ ГИСТОЛОГИИ)		РМЖ (ДАННЫЕ ГИСТОЛОГИИ)		РМЖ (ДАННЫЕ ГИСТОЛОГИИ)
2	Классификация находок		BI-RADS 2		BI-RADS 2		BI-RADS 2
3	Наличие находок		'НЕТ ОЧАГОВ'/ 'ЕСТЬ ОЧАГИ'		'НЕТ ОЧАГОВ'/ 'ЕСТЬ ОЧАГИ'		'НЕТ ОЧАГОВ'/ 'ЕСТЬ ОЧАГИ'

Рис. 4. Классификация разметки по трездатратам и степени верификации.

Как для ретро-, так и для проспективной разметки могут использоваться различные средства автоматизации работы с данными (например, просмотр результатов медицинской визуализации и создание бинарных масок, анализ баз данных) с применением различных технологий и языков программирования (C/C++, Python, Kotlin, Java и др.) [11].

4. Этап документирования

После того как датасет прошёл все предыдущие стадии и готов к передаче третьим лицам, он считается «готовым к публикации». Публикация датасета сопровождается выпуском первой мажорной версии (1.0.0), а также подготовкой и публикацией сопроводительной документации (README файл).

В процессе подготовки датасета неизбежно упускаются те или иные критерии, всплывающие при непосредственной работе с датасетом конечными пользователями (специалистами по проведению валидации решений на основе ИИ или исследователями,

использующими машинное обучение). Внесение корректировок в датасет должно быть прозрачным для всех участников процесса и пользователей. Ведение версионирования датасета позволяет отслеживать подобные изменения.

Нами предложен следующий оригинальный подход к решению описанной проблемы как вариация семантического версионирования [12]:

1. Мажорная версия (Major): увеличивается при изменении значимых параметров датасета, связанных с клинической задачей, целью, принципами разметки и верификации данных.
2. Минорная версия (Minor): увеличивается при замене, добавлении или удалении единиц данных в датасете без изменения других значимых параметров датасета; при этом алгоритмы обучения или валидации могут использовать новую минорную версию без изменения кода. При выпуске новой мажорной версии минорная версия устанавливается равной 0.
3. Патч-версия (Patch): увеличивается при внесении корректировок в сопроводительную документацию, исправлении опечаток и иных ошибок в файлах разметки, при этом не меняется количество и качество единиц данных в датасете. При выпуске новой мажорной и/или минорной версии патч-версия устанавливается равной 0.

В целях удобства использования датасета в корневую директорию помещаются файл с названием README.md в формате Markdown и сгенерированный на его основе README.pdf в формате Adobe PDF. Единый подход к структуре README файла позволит в будущем организовать удобный поиск и фильтрацию по всем опубликованным датасетам. Базовая структура README файла представлена на рис. 5, однако при необходимости в файл могут быть добавлены и другие разделы.

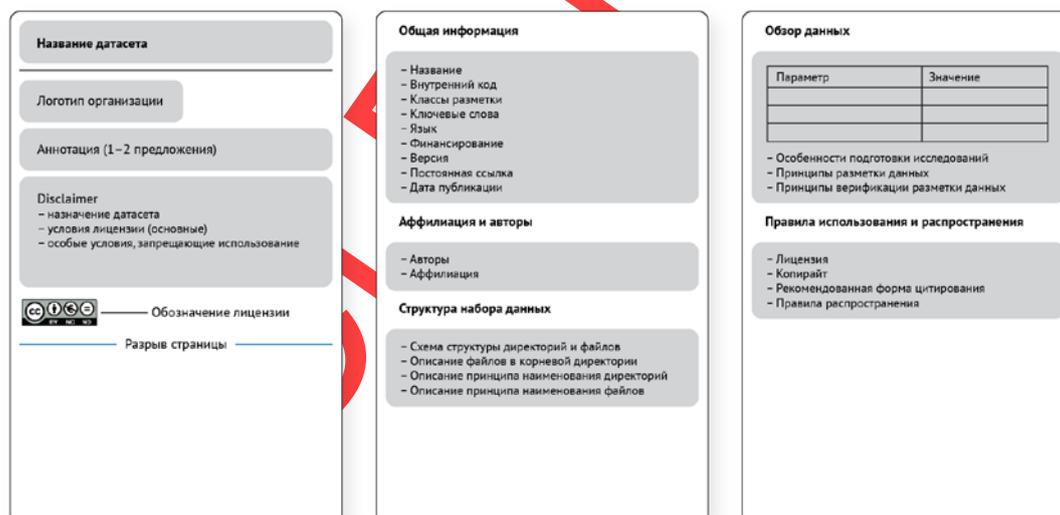


Рис. 5. Базовая структура README файла.

С точки зрения удобства ведения отчётности практическую ценность имеет единый реестр подготовленных наборов данных, пример которого приведён в таблице [13].

Минимальный набор рекомендуемых полей реестра:

1. Порядковый номер записи в реестре.
2. Внутренний код, уникальный для датасета в текущем реестре и/или учреждении.
3. Назначение датасета, область применения.
4. Модальность/процедура (характеристики исследований, пригодные для их поиска и выбора в МИС).

5. Искомые признаки и/или целевая патология (при возможности с указанием кода Международной классификации болезней).
6. Определение единицы данных.
7. Количество единиц данных (при возможности с указанием выходного объёма данных в Мб, Гб или Тб).
8. Классы разметки с указанием количества записей в каждом классе.

ОБСУЖДЕНИЕ

В данной работе представлен экспериментальный подход к формированию наборов медицинских данных (датасетов) для использования при разработке и оценке медицинских диагностических интеллектуальных систем, использующих технологии ИИ.

Использование масштабной МИС (ЕРИС ЕМИАС) в качестве источника данных для датасета является определённой гарантией его репрезентативности. Параметры эффективности работы алгоритма ИИ после внедрения в диагностический процесс с большой вероятностью будут соответствовать параметрам, полученным во время валидации на таком датасете. При этом необходимо учитывать вариативность парка диагностических аппаратов, а также вариации физических параметров проводимых исследований, стараясь представить наиболее широкий спектр исследований в датасете. Ценность вариативности представленных в датасетах аппаратов различных производителей может иметь практическую значимость для тонкой настройки порога срабатывания систем ИИ с целью обеспечения их надёжной работы [14].

Другим преимуществом работы с ЕРИС ЕМИАС является практически неограниченный доступ к сотням тысяч лучевых исследований различных модальностей, что позволяет создавать датасеты с чрезвычайно разными наборами технических, демографических и клинических характеристик. Такие вариации гарантируют ценность формируемых наборов данных не только для оценки точности, но и масштабируемости и надёжности разрабатываемых и испытываемых систем ИИ.

Предлагаемый подход был разработан и апробирован при создании 25 датасетов по 7 направлениям в лучевой диагностике общим числом более 1400 единиц данных (исследований), в том числе при реализации московского эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения г. Москвы [15] (см. рис. 3). Полный перечень датасетов приведён в таблице. Положения, описанные в настоящей статье, согласуются с критериями эталонных датасетов, включёнными в методические рекомендации по проведению клинических испытаний программного обеспечения на основе интеллектуальных технологий в лучевой диагностике [16].

В ходе московского эксперимента независимая внешняя оценка алгоритмов ИИ обеспечивается в два этапа (функциональное и калибровочное тестирование соответственно): на первом используются относительно небольшие датасеты (до 5 единиц данных) для проверки технической возможности чтения и обработки исследований; на втором этапе используются датасеты среднего размера (в среднем от 100 до 200 единиц данных) для сравнения результатов обработки исследований ИИ с верифицированной разметкой. В случаях, когда в результате первичного тестирования разработчик решения на основе ИИ получает рекомендации для доработки своего решения, предусмотрена возможность повторного тестирования на другом датасете.



Рис. 3. Датасеты московского эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения г. Москвы, подготовленные по настоящей методике.

Важной частью жизненного цикла датасета как набора данных после его публикации является научное представление результатов работы в соответствующих публикациях и рукописях. Одним из порталов, поддерживающих бесплатное размещение информации по общедоступным датасетам, является medRxiv² — сервис препринтов по биомедицинской тематике. Достоинством сервиса является отсутствие внешнего рецензирования публикаций, что позволяет осуществить информирование сообщества о результатах своей работы в кратчайшие сроки. Пример публикации о датасете на портале medRxiv представлен в [17].

Необходимо отметить, что датасеты, сформированные по настоящей методике, успешно используются отечественными и зарубежными научными коллективами, о чём свидетельствуют последние публикации [18, 19]. Использование результатов работы на практике подтверждает своевременность и адекватность сформулированных подходов и методологии.

При внесении необходимых изменений методика может быть полностью или частично использована не только для других направлений в лучевой диагностике, но и вне её, в других областях практической медицины, в которых в ходе медицинской деятельности происходит накопление первично-электронной информации (электроэнцефалограммы, электрокардиограммы и другие записи физиологических сигналов, записи с прикроватных реанимационных мониторов, лог-записи современного лабораторного оборудования, например химических анализаторов, и др.). В частности, принципы формулирования клинической и/или практической задачи, работы с МИС по выгрузке исходных данных, общие принципы разметки и документирования в экспериментальном режиме были успешно опробованы при формировании набора данных электрокардиограмм с признаками сердечно-сосудистых заболеваний. В перспективе настоящая методика может быть включена в государственный стандарт, обеспечив тем самым преемственность и унификацию медицинских датасетов для обучения и тестирования технологий ИИ на национальном уровне.

² Режим доступа: <https://medrxiv.org>. Дата обращения: 15.01.2021.

Остро обсуждаемым вопросом является проблема обезличивания медицинских данных, особенно результатов лучевых исследований. В настоящее время нет общепринятого стандарта обезличивания медицинских изображений. Специалисты, работающие с подобного рода данными, должны следовать здравой логике недопущения разглашения медицинской тайны и персональных данных пациента. Необходимо помнить, что результаты лучевого исследования сами по себе могут служить источником персональных данных: например, по срезам головы возможно восстановить объемное изображение мягких тканей лицевого черепа, что в свою очередь в достаточной степени позволяет идентифицировать человека. Несмотря на отсутствие явных законодательных норм либо стандартов обезличивания в таких ситуациях, автору датасета следует принимать решение об удалении записи мягких тканей головы из исследований, отталкиваясь от поставленной клинической и/или практической задачи, а также от назначения датасета.

Для поддержания темпов роста рынка технологий ИИ в медицине следует по возможности рассматривать обеспечение свободного доступа к датасетам при соблюдении всех описанных выше условий обезличивания. Для публикации статей с описанием датасетов используют такие порталы, как arXiv (<https://arxiv.org>), medRxiv (<https://medrxiv.org>) и Zenodo (<https://zenodo.org>). Существует большое количество публичных репозиторий открытых датасетов, а также интегральный поиск по ним, например Google's Dataset Search³. Одним из путей не только обеспечения легального доступа к датасетам, но и привлечения внимания со стороны сообщества разработчиков ИИ является проведение онлайн-соревнований среди разработчиков ИИ на таких платформах, как Kaggle⁴.

Перспективным направлением развития является использование «цифровых двойников заболевания» — обширных наборов информации о пациентах различного профиля (социальной, демографической, поведенческой и пр.) для формирования статистических признаков, свойственных пациентам, страдающим конкретным заболеванием. Использование такой информации может позволить создавать более репрезентативные медицинские датасеты, включающие наиболее широкий спектр признаков и факторов заболевания, значимых для клинической и/или практической задачи. Основой для создания «цифрового двойника заболевания» являются в первую очередь анализ и переработка обезличенной информации, полученной из «цифровых двойников пациентов», содержащих максимально широкий набор разноплановой информации о пациенте.

Представленный в настоящей статье подход позволяет систематизировать и стандартизировать подготовку датасетов и их жизненный цикл для последующего использования при проведении испытаний интеллектуальных систем (в том числе на основе ИИ) и регистрации испытуемых систем в целях их дальнейшего применения в сфере здравоохранения. Такая поэтапная и детализированная методология формирования датасетов позволит разработчикам объективно оценивать свои продукты, а регуляторам — удостовериться в объективности и прозрачности процесса оценки с использованием датасетов, созданных на базе предлагаемой методологии.

ЗАКЛЮЧЕНИЕ

Задача формирования наборов медицинских данных для обучения и валидации диагностических систем на основе технологий искусственного интеллекта становится особенно актуальной в связи с активным развитием этой области. Оригинальные

³ Режим доступа: <https://datasetsearch.research.google.com>. Дата обращения: 15.01.2021.

⁴ Режим доступа: <https://kaggle.com>. Дата обращения: 15.01.2021.

подходы, описанные в статье, могут послужить отправной точкой для создания полноценной методологии подготовки и стандартизации медицинских датасетов различных модальностей и типов данных, а также определить необходимые условия и факторы для успешного практического применения этой методологии.

ДОПОЛНИТЕЛЬНО

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

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

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.А. Павлов — формирование структуры публикации, написание текста публикации, разработка подходов подготовки датасетов, подготовка примеров датасетов; С.П. Морозов — общая концепция исследования; А.Е. Андрейченко — разработка дизайна исследования, курация написания текста публикации, редактирование текста; А.В. Владзимирский — научное обоснование подходов к формированию датасетов; А.А. Ревазян — подготовка обзора литературы, участие в подготовке примеров датасетов; Ю.С. Кирпичев — участие в подготовке примеров датасетов.

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

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

Author contribution. N.A. Pavlov — manuscript design and writing, development of an approach for dataset formation, formation of dataset examples; S.P. Morozov — concept of research; A.E. Andreychenko — study design, manuscript curation and editing; A.V. Vladzimirsky — scientific rationale for dataset formation; A.A. Revazyan — literature review, formation of dataset examples; Yu.S. Kirpichev — formation of dataset examples. 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.

СПИСОК ЛИТЕРАТУРЫ

1. Гусев А.В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения // Врач и информационные технологии. 2017. № 3. С. 92–105.
2. Ranschaert E.R., Morozov S., Algra P.R., eds. Artificial intelligence in medical imaging. Cham: Springer International Publishing; 2019. doi: 10.1007/978-3-319-94878-2
3. Griffith B., Kadom N., Straus C.M. Radiology Education in the 21st Century: Threats and Opportunities // J Am Coll Radiol. 2019. Vol. 16, N 10. P. 1482–1487. doi: 10.1016/j.jacr.2019.04.003
4. Savadjiev P., Chong J., Dohan A., et al., Demystification of AI-driven medical image interpretation: past, present and future // European Radiology. 2019. Vol. 29. N 3, P. 1616–1624. doi: 10.1007/s00330-018-5674-x

5. Ng A. What artificial intelligence can and can't do right now. *Harvard Business Review*; 2016. Available from: <https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now>
6. Renear H., Sacchi S., Wickett K.M. Definitions of dataset in the scientific and technical literature // *Proceedings of the American Society for Information Science and Technology*. 2010. Vol. 47, N 1. P. 1–4. doi: 10.1002/meet.14504701240
7. Tan S.L., Gao G., Koch S. Big data and analytics in healthcare // *Methods Inf Med*. 2015. Vol. 54, N 6. P. 546–547. doi: 10.3414/ME15-06-1001
8. Kohli M.D., Summers R.M., Geis J.R. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session // *J Digit Imaging*. 2017. Vol. 30, N 4. P. 392–399. doi: 10.1007/s10278-017-9976-3
9. Willemink M.J., Koszek W.A., Hardell C., et al. Preparing medical imaging data for machine learning // *Radiology*. 2020. Vol. 295, N 1. P. 4–15. doi: 10.1148/radiol.2020192224
10. Морозов С.П., Шелехов П.В., Владзимирский А.В. Современные стандартизованные подходы к совершенствованию службы лучевой диагностики // *Проблемы стандартизации в здравоохранении*. 2019. № 5-6. С. 30–34. doi: 10.26347/1607-2502201905-06030-034
11. 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 // *Health Care Russian Federation*. 2021. Vol. 64, N 6. P. 343–350. doi: 10.46563/0044-197x-2020-64-6-343-350
12. Preston-Werner T. *Semantic Versioning 2.0.0* [Internet]. Available from: <https://semver.org>
13. Морозов С.П., Проценко Д.Н., Сметанина С.В. и др. Лучевая диагностика коронавирусной болезни (COVID-19): организация, методология, интерпретация результатов : препринт № ЦДТ — 2020 — II. Версия 2 от 17.04.2020. Серия «Лучшие практики лучевой и инструментальной диагностики». Вып. 65. Москва : ГБУЗ НПКЦ ДиТ ДЗМ, 2020. 80 с. Режим доступа: <https://tele-med.ai/biblioteka-dokumentov/luchevaya-diagnostika-koronavirusnoj-bolezni-covid-19-organizaciya-metodologiya-interpretaciya-rezultatov>. Дата обращения: 15.01.2021.
14. Pavlov N. et al. *ECR 2021: Value of technical stratification of medical datasets for AI services*. Moscow, 2021. [Internet]. Available from: <https://connect.myesr.org/course/ai-in-breast-imaging/>
15. Морозов С.П., Владзимирский А.В., Ледихова Н.В. и др. Московский эксперимент по применению компьютерного зрения в лучевой диагностике: вовлеченность врачей-рентгенологов // *Врач и информационные технологии*. 2020. № 4. С. 14–23. doi: 10.37690/1811-0193-2020-4-14-23
16. Morozov S.P., Vladzimirskyu A.V., Klyashtorny V.G., et al. Clinical acceptance of software based on artificial intelligence technologies (radiology). Series "Best practices in medical imaging". Issue 57. Moscow; 2019. 45 p.
17. Morozov S.P., Andreychenko A.E., Pavlov N.A., et al. *MosMedData: Chest CT scans with COVID-19 related findings dataset* // *medRxiv*. 2020. doi: 10.1101/2020.05.20.20100362
18. Sushentsev N., Bura V., Kotniket M., et al. A head-to-head comparison of the intra- and interobserver agreement of COVID-RADS and CO-RADS grading systems in a population with high estimated prevalence of COVID-19 // *BJR Open*. 2020. Vol. 2, N 1. P. 20200053. doi: 10.1259/bjro.20200053
19. Jin C., Chen W., Caoet Y., et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis // *Nat Commun*. 2020. Vol. 11, N 1. P. 5088. doi: 10.1038/s41467-020-18685-1

REFERENCES

1. Gusev AV. Prospects for neural networks and deep machine learning in creating health solutions (Complex medical information system, Russian). *Vrach i Informatsionnye Tekhnologii*. 2017;(3):92–105. (In Russ).
2. Ranschaert ER, Morozov S, Algra PR, eds. Artificial intelligence in medical imaging. Cham: Springer International Publishing; 2019. doi: 10.1007/978-3-319-94878-2
3. Griffith B, Kadom N, Straus CM. Radiology Education in the 21st Century: Threats and Opportunities. *J Am Coll Radiol*. 2019;16(10):1482–1487. doi: 10.1016/j.jacr.2019.04.003
4. Savadjiev P, Chong J, Dohan A, et al. Demystification of AI-driven medical image interpretation: past, present and future. *Eur Radiol*. 2019;29(3):1616–1624. doi: 10.1007/s00330-018-5674-x
5. Ng A. What artificial intelligence can and can't do right now. *Harvard Business Review*; 2016. Available from: <https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now>
6. Renear H, Sacchi S, Wickett KM. Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*. 2010;47(1):1-4. doi: 10.1002/meet.14504701240
7. Tan SL, Gao G, Koch S. Big data and analytics in healthcare. *Methods Inf Med*. 2015;54(6):546–547. doi: 10.3414/ME15-06-1001
8. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. *J Digit Imaging*. 2017;30(4):392–399. doi: 10.1007/s10278-017-9976-3
9. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295(1):4–15. doi: 10.1148/radiol.2020192224
10. Morozov SP, Shelekhov PV, Vladzimirsky AV. Modern approaches to the radiology service improvement. *Health Care Standardization Problems*. 2019;(5-6):30–34. (In Russ). doi: 10.26347/1607-2502201905-06030-034
11. 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. *Health Care Russian Federation*. 2021;64(6):343–350. doi: 10.46563/0044-197x-2020-64-6-343-350
12. Preston-Werner T. Semantic Versioning 2.0.0 [Internet]. Available from: <https://semver.org>
13. Morozov SP, Protsenko DN, Smetanina SV, et al. Radiation diagnostics of coronavirus disease (COVID-19): organization, methodology, interpretation of results: Preprint No.CDT — 2020 — II. Version 2 from 17.04.2020. The series "Best practices of radiation and instrumental diagnostics". Issue 65. Moscow : Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Department of Health; 2020. 80 p. (In Russ). Available from: <https://tele-med.ai/biblioteka-dokumentov/luchevaya-diagnostika-koronavirusnoj-bolezni-covid-19-organizaciya-metodologiya-interpretaciya-rezultatov>
14. Pavlov N. et al. ECR 2021: Value of technical stratification of medical datasets for AI services. Moscow, 2021. [Internet]. Available from: <https://connect.mysr.org/course/ai-in-breast-imaging/>
15. Morozov SP, Vladzimirskyy A, Andreychenko A, et al. Moscow experiment on computer vision in radiology: involvement and participation of radiologists. *Vrach i informacionnye tehnologii*. 2020;(4):14–23. doi: 10.37690/1811-0193-2020-4-14-23

16. Morozov SP, Vladzimirskyy AV, Klyashtornyy VG, et al. Clinical acceptance of software based on artificial intelligence technologies (radiology). Series "Best practices in medical imaging". Issue 57. Moscow; 2019. 45 p.
17. Morozov SP, Andreychenko AE, Pavlov NA, et al. MosMedData: Chest CT scans with COVID-19 related findings dataset. *medRxiv*. 2020. doi: 10.1101/2020.05.20.20100362
18. Sushentsev N, Bura V, Kotniket M, et al. A head-to-head comparison of the intra- and interobserver agreement of COVID-RADS and CO-RADS grading systems in a population with high estimated prevalence of COVID-19. *BJR Open*. 2020;2(1):20200053. doi: 10.1259/bjro.20200053
19. Jin C, Chen W, Caoet Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nat Commun*. 2020;11(1):5088. doi: 10.1038/s41467-020-18685-1

ОБ АВТОРАХ	AUTHORS INFO
* Павлов Николай Александрович ; адрес: Россия, 109029, Москва, ул. Средняя Калитниковская, д. 28, стр.1; тел.: +7 (495) 276-04-36; ORCID: https://orcid.org/0000-0002-4309-1868 ; eLibrary SPIN: 9960-4160; e-mail: n.pavlov@npcmr.ru	* Nikolay A. Pavlov , MD; address: 28-1, Srednyaya Kalitnikovskaya street, 109029, Moscow, Russia; tel.: +7 (495) 276-04-36; ORCID: https://orcid.org/0000-0002-4309-1868 ; eLibrary SPIN: 9960-4160; e-mail: n.pavlov@npcmr.ru
Андрейченко Анна Евгеньевна , к.ф.-м.н.; ORCID: https://orcid.org/0000-0001-6359-0763 ; eLibrary SPIN: 6625-4186; e-mail: a.andreychenko@npcmr.ru	Anna E. Andreychenko , Cand. Sci. (Physics & Mathematics); ORCID: https://orcid.org/0000-0001-6359-0763 ; eLibrary SPIN: 6625-4186; e-mail: a.andreychenko@npcmr.ru
Владимирский Антон Вячеславович , д.м.н.; ORCID: https://orcid.org/0000-0002-2990-7736 ; eLibrary SPIN: 3602-7120; e-mail: a.vladimirsky@npcmr.ru	Anton V. Vladzimirskyy , MD, Dr. Sci. (Med.); ORCID: https://orcid.org/0000-0002-2990-7736 ; eLibrary SPIN: 3602-7120; e-mail: a.vladimirsky@npcmr.ru
Ревазян Ануш Артуровна ; ORCID: http://orcid.org/0000-0003-1589-2382 ; e-mail: anushrevazyan@gmail.com	Anush A. Revazyan ; ORCID: http://orcid.org/0000-0003-1589-2382 ; e-mail: anushrevazyan@gmail.com
Кирпичев Юрий Сергеевич ; ORCID: http://orcid.org/0000-0002-9583-5187 ; eLibrary SPIN: 3362-3428; e-mail: y.kirpichev@npcmr.ru	Yury S. Kirpichev ; ORCID: http://orcid.org/0000-0002-9583-5187 ; eLibrary SPIN: 3362-3428; e-mail: y.kirpichev@npcmr.ru
Морозов Сергей Павлович , д.м.н., профессор; ORCID: http://orcid.org/0000-0001-6545-6170 ; eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru	Sergey P. Morozov , MD, Dr. Sci. (Med.), Professor; ORCID: http://orcid.org/0000-0001-6545-6170 ; eLibrary SPIN: 8542-1720; e-mail: morozov@npcmr.ru

Таблица. Перечень медицинских датасетов, разработанных с применением описываемого экспериментального подхода

№	Внутренний код	Назначение	Модальность / процедура	Искомые признаки и/или целевая нозология	Единица данных	Количество единиц данных	Классы разметки (кол-во ед. данных)
1.	DS_FT-I_CT_OGK_CANCER	ФТ-I	КТ ОГК	Рак лёгкого	Исследование КТ	5	Без признаков патологии (2), с признаками патологии (2), с техническим дефектом (1)
2.	DS_FT-I_CT_OGK_COVID	ФТ-I	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	5	
3.	DS_FT-II_CT_OGK_COVID	ФТ-II	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	4	Без признаков патологии (2), с признаками патологии (2)
4.	DS_FT-I_LDCT_OGK_CANCER	ФТ-I	Низкодозная КТ ОГК	Рак лёгкого	Исследование КТ	5	Без признаков патологии (2), с признаками патологии (2), с техническим дефектом (1)
5.	DS_FT-I_MMГ_CANCER	ФТ-I	ММГ	Рак молочной железы	Исследование ММГ	5	
6.	DS_FT-II_MMГ_CANCER	ФТ-II	ММГ	Рак молочной железы	Исследование ММГ	4	Без признаков патологии (2), с признаками патологии (2)
7.	DS_FT-I_DX_OGK_PAT	ФТ-I	РГ ОГК	Патологии органов дыхания	Исследование РГ	5	Без признаков патологии (2), с признаками

№	Внутренний код	Назначение	Модальность / процедура	Искомые признаки и/или целевая нозология	Единица данных	Количество единиц данных	Классы разметки (кол-во ед. данных)
8.	DS_FT-II_DX_OGK_PAT	ФТ-II	РГ ОГК	Патологии органов дыхания	Исследование РГ	5	патологии (2), с техническим дефектом (1)
9.	DS_FT-I_DX_OGK_COVID	ФТ-I	РГ ОГК	Вирусная пневмония (COVID-19)	Исследование РГ	4	Без признаков патологии (2), с признаками патологии (2)
10.	DS_FT-II_DX_OGK_COVID	ФТ-II	РГ ОГК	Вирусная пневмония (COVID-19)	Исследование РГ	4	патологии (2)
11.	DS_FT-I_FLG_OGK_PAT	ФТ-I	ФЛГ ОГК	Патологии органов дыхания	Исследование ФЛГ	4	
12.	DS_FT-II_FLG2_OGK_PAT	ФТ-II	ФЛГ ОГК	Патологии органов дыхания	Исследование ФЛГ	4	
13.	DS_CT-I_CT_OGK_CANCER	КлТ-I	КТ ОГК	Рак лёгкого	Исследование КТ	100	Без признаков патологии (50), с признаками патологии (50)
14.	DS_CT-I_CT_OGK_COVID	КлТ-I	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	100	Без признаков патологии (50), с признаками патологии (50)
15.	DS_CT-II_CT_OGK_COVID	КлТ-II	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	100	Без признаков патологии (50), с признаками патологии (50)
16.	DS_CT-II_CT_OGK_COVID_2	КлТ-I	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	125	КТ0 (25), КТ1 (25), КТ2 (25),

№	Внутренний код	Назначение	Модальность / процедура	Искомые признаки и/или целевая нозология	Единица данных	Количество единиц данных	Классы разметки (кол-во ед. данных)
							КТ3 (25), КТ4 (25)
17.	DS_CT-II_CT_OGK_COVID_3	КлТ-I	КТ ОГК	Вирусная пневмония (COVID-19)	Исследование КТ	200	КТ0 (100), КТ1 (25), КТ2 (25), КТ3 (25), КТ4 (25)
18.	DS_CT-I_LDCT_OGK_CANCER	КлТ-I	Низкодозная КТ ОГК	Рак лёгкого	Исследование КТ	100	Без признаков патологии (50), с признаками патологии (50)
19.	DS_CT-I_MMG_CANCER	КлТ-I	ММГ	Рак молочной железы	Исследование ММГ	100	
20.	DS_CT-II_MMG_CANCER	КлТ-II	ММГ	Рак молочной железы	Исследование ММГ	100	
21.	DS_CT-I_DX_OGK_CANCER	КлТ-I	РГ ОГК	Патологии органов дыхания	Исследование РГ	100	
22.	DS_CT-II_DX_OGK_CANCER	КлТ-II	РГ ОГК	Патологии органов дыхания	Исследование РГ	100	
23.	DS_CT-I_DX_OGK_COVID	КлТ-I	РГ ОГК	Вирусная пневмония (COVID-19)	Исследование РГ	100	
24.	DS_CT-I_FLG_OGK_CANCER	КлТ-I	ФЛГ ОГК	Патологии органов дыхания	Исследование ФЛГ	100	

№	Внутренний код	Назначение	Модальность / процедура	Искомые признаки и/или целевая нозология	Единица данных	Количество единиц данных	Классы разметки (кол-во ед. данных)
25.	DS_CT-II_FLG_OGK_CANCER	КлТ-II	ФЛГ ОГК	Патологии органов дыхания	Исследование ФЛГ	100	

Примечание. ФТ-I — первичное функциональное тестирование; ФТ-II — повторное функциональное тестирование; КлТ-I — первичное калибровочное тестирование; КлТ-II — повторное калибровочное тестирование; КТ — компьютерная томография; ММГ — маммография; РГ — рентгенография; ФЛГ — флюорография; ОГК — органы грудной клетки; КТ1 — зоны уплотнения по типу матового стекла. Вовлечение паренхимы лёгкого $\leq 25\%$; КТ2 — зоны уплотнения по типу матового стекла. Вовлечение паренхимы лёгкого 25–50%; КТ3 — зоны уплотнения по типу матового стекла и консолидации. Вовлечение паренхимы лёгкого 50–75%; КТ4 — диффузное уплотнение лёгочной ткани по типу матового стекла и консолидации в сочетании с ретикулярными изменениями. Вовлечение паренхимы лёгкого $>75\%$ [13].