Государственное бюджетное учреждение здравоохранения города Москвы
«Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»

На правах рукописи

Бобровская Татьяна Михайловна

МЕТОДОЛОГИЯ ФОРМИРОВАНИЯ НАБОРОВ ДАННЫХ И ИХ ИСПОЛЬЗОВАНИЕ ДЛЯ ОЦЕНКИ ДИАГНОСТИЧЕСКОЙ ТОЧНОСТИ СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЛУЧЕВОЙ ДИАГНОСТИКЕ

3.3.9. – Медицинская информатика (медицинские науки)

Диссертация
на соискание ученой степени
кандидата медицинских наук

Научный руководитель:
Арзамасов Кирилл Михайлович,
доктор медицинских наук

Москва – 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность и степень разработанности темы исследования

Технологии искусственного интеллекта (ТИИ) находят все более широкое применение во всех сферах нашей жизни, включая здравоохранение. Начавшаяся в 2011 г. масштабная цифровизация [1] способствовала росту количества медицинских данных, что, в свою очередь, продиктовало необходимость их централизации для обработки и хранения. Наличие большого количества цифровых данных дало импульс для развития ТИИ в медицине. Кроме того, Национальная стратегия развития искусственного интеллекта на период до 2030 года (далее – Национальная стратегия), вступившая в силу в конце 2019 г. [2], также поддерживает создание и развитие ТИИ в РФ: ожидается, что внедрение таких решений будет способствовать росту мировой экономики, а в социальной сфере – созданию условий для улучшения уровня жизни населения, в том числе за счет повышения качества услуг в сфере здравоохранения. Благодаря этому в последние годы появилось множество инструментов, способствующих повышению качества и скорости оказания медицинских услуг: начиная от электронных медицинских карт, чат-ботов поддержки пациентов, систем поддержки принятия врачебных решений и заканчивая автоматическими системами диагностики заболеваний [3]. ТИИ – это совокупность технологий, включающая в себя компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы искусственного интеллекта (ИИ) (методы, направленные на создание принципиально новой научно-технической продукции, в том числе в целях разработки универсального (сильного) ИИ) [2]. ТИИ применяются практически во всех направлениях медицины: радиологии, онкологии, офтальмологии, хирургии, фармацевтике, генетике, неврологии, психиатрии и т. д. [4, 5]. Система ИИ (СИИ) – техническая система, в которой используются ТИИ. Применение СИИ уменьшает время ожидания результатов исследований, улучшает приверженность лечению, помогает

подобрать дозировки препаратов и интерпретировать диагностические исследования [4].

Одним из самых популярных направлений для внедрения ТИИ является медицинская диагностика, в частности лучевая. Алгоритмы анализа рентгенологических изображений стали одними из первых зарубежных разработок программного обеспечения (ПО) на основе ТИИ для медицинской диагностики, одобренного FDA (Food and Drug Administration) [4]. В нашей стране активное внедрение ТИИ в медицину также началось в лучевой диагностике, и одним из наиболее успешных и наглядных проектов стал «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения этих технологий в системе здравоохранения» (далее – Эксперимент) [6]. Эксперимент был начат в период пандемии COVID-19 в 2020 г. и показал, насколько важно наличие цифрового помощника в условиях острой нехватки персонала. Так, применение ТИИ позволило снизить время ожидания результатов диагностического исследования, а также производить их сортировку по степени тяжести для первоочередной маршрутизации более тяжелых пациентов с целью оказания своевременной помощи [6]. В рутинной практике ТИИ также могут способствовать повышению качества оказания медицинской помощи и снижению трудозатрат. Одним из примеров является применение ТИИ в профилактической маммографии (ММГ), для которой обязательно двойное чтение с целью минимизации пропуска патологии. Многочисленные исследования показали, что СИИ демонстрируют значения критериев диагностической точности [7–9], сопоставимые с врачами-рентгенологами, и могут использоваться в качестве «второго чтения» [10, 11]. Потребность в СИИ возрастает с каждым днем, о чем говорит динамика увеличения количества так называемых ИИ-сервисов (СИИ в Эксперименте) [6, 12], а также их внедрение в систему обязательного медицинского страхования в Москве (код процедуры – 001601 «Описание и интерпретация данных маммографического исследования с использованием искусственного интеллекта») [11, 13].

Однако, несмотря на все преимущества применения ТИИ, имеется ряд технических, этических и правовых ограничений [4, 14]. Оценивая зрелость СИИ для лучевой диагностики [15, 16], можно сделать вывод об отсутствии решений, имеющих точность, превышающую 95,0 %. Одной из главных проблем является отсутствие качественных данных для обучения моделей ИИ. На сегодняшний день одним из наиболее популярных направлений в сфере ИИ является машинное обучение, которое имеет ряд особенностей [2]:

– для поиска вычислительной системой непредвзятого решения требуется ввести репрезентативный, релевантный и корректно размеченный набор данных (НД);

– алгоритмы работы нейронных сетей крайне сложны для интерпретации и, следовательно, результаты их работы могут быть подвергнуты сомнению и отменены человеком. Отсутствие понимания того, как ИИ достигает результатов, является одной из причин низкого уровня доверия к современным ТИИ и может стать препятствием для их развития.

Медицинские данные характеризуются большим объемом, сложностью и беспорядочностью, и для того, чтобы их можно было использовать в машинном обучении, они должны быть соответствующим образом обработаны, стандартизированы и размечены, так как точность модели в значительной степени зависит от качества НД, его репрезентативности и релевантности [14]. Национальная стратегия [2] ставит ряд задач, которые требуют создания таких НД:

– разработка и развитие СИИ (необходимы НД для обучения, дообучения, тестирования и мониторинга работы СИИ);

– повышение доступности и качества данных для СИИ (создание открытых библиотек НД, принципов их стандартизации и инструментов контроля качества);

– поддержка научных исследований в целях обеспечения опережающего развития ИИ (создание НД с целью изучения ТИИ, поиска новых направлений развития).

НД требуются не только для обучения СИИ. Для допуска систем в практическую деятельность необходимо проведение независимого тестирования

с определением критериев диагностической точности на НД, который не использовался при обучении. Согласно обзору [17], проведенному в 2021 г., существуют большие различия в методологиях выполнения тестирований, в создании НД, эталонных стандартах определения заболеваний, а также интерпретируемости результатов и терминологии. В отдельных исследованиях в той или иной степени представлены этапы создания НД [18–20], однако они носят описательный характер, не имеют четкой структуры и зачастую освещают не все аспекты процесса создания НД. Больше объективности и системности вносят специализированные чек-листы [21, 22], но они учитывают лишь малую часть параметров создания НД и представляют собой памятку по описанию НД, а не алгоритм его формирования. Поэтому представляется актуальным создание стандартизированной методологии формирования НД, а также инструментов управления, автоматизации и контроля качества, что отражено в следующих направлениях по повышению доступности и качества данных Национальной стратегии:

– разработка унифицированных и обновляемых методологий описания, сбора и разметки данных, а также механизма контроля за соблюдением указанных методологий;

– создание и развитие информационно-коммуникационной инфраструктуры для обеспечения доступа к НД посредством создания (модернизации) общедоступных платформ для хранения НД, соответствующих методологиям описания, сбора и разметки данных.

Цели и задачи

Цель исследования – создание методологии формирования наборов данных для обеспечения качества систем искусственного интеллекта в лучевой диагностике.

Задачи исследования:

1. Оценить управляемость, надежность и устойчивость процессов

формирования НД, применяемых при разработке и тестировании программных средств анализа медицинских изображений в лучевой диагностике.

2. Обосновать принципы систематизации НД в лучевой диагностике и разработать концепцию выбора и применения глоссария и тезауруса для описания процессов, связанных с созданием и использованием НД.

3. Разработать подход к определению минимального размера НД для тестирования СИИ в лучевой диагностике.

4. Создать, внедрить и оценить эффективность методов стандартизации и оптимизации процессов формирования НД.

Научная новизна

1. Впервые определены пути оптимизации подготовки НД.

2. Впервые разработаны принципы стандартизации и систематизации НД в лучевой диагностике.

3. Впервые разработаны эмпирические принципы расчета минимального размера НД на основании критериев диагностической точности для независимой оценки СИИ.

4. Разработана новый инструмент для контроля качества подготовки и применения наборов медицинских данных.

5. Впервые внедрены методы стандартизации и оптимизации процессов формирования НД.

Теоретическая и практическая значимость работы

1. Сформулированы основные способы повышения качества процесса создания НД.

2. Разработаны и внедрены практические рекомендации по созданию НД для лучевой диагностики.

3. Создана инфраструктура с целью обеспечения доступа к НД и информации о них: реестр и библиотеки.

4. Предложены практические рекомендации по оценке СИИ: минимальный объем выборки и баланс классов.

5. Сформирован задел для создания методологии расчета минимального объема выборки для оценки различных показателей СИИ.

Методология и методы исследования

Методы исследования:

– аналитические (анализ, синтез, индукция, дедукция);

– оценка диагностической точности СИИ (построение и анализ характеристической кривой ROC);

– статистические.

Положения, выносимые на защиту

1. Отсутствие единой методологии создания НД, регламентирующей все этапы, а также единых принципов классификации НД и систематизации информации о них приводит к появлению большого числа ошибок, затрудняет и замедляет процесс формирования НД, что в конечном счете при использовании согласно назначению может привести к некорректным результатам.

2. Использование реестра, аккумулирующего и систематизирующего основную информацию о НД, позволяет снизить время выполнения отдельных этапов создания НД на 97 %, оперативно получать справочную и отчетную информацию, на основании которой принимаются управленческие решения, а также способствует централизации хранения, автоматизации и контролю качества НД.

3. Минимальный объем НД при проведении оценки диагностической точности СИИ с бинарным исходом с помощью ROC-анализа составляет 80 исследований при балансе классов 0,5 (50 % представленность каждого класса)

и 0,2 (20 % представленность целевого признака), 120 – при 0,4 (40 % представленность целевого признака), 150 – при 0,3 (30 % представленность целевого признака), 190 – при 0,1 (10 % представленность целевого признака).

4. Единая методология формирования НД способствует повышению качества НД, эффективному использованию ресурсов, ускорению и автоматизации создания НД, позволяет создавать специализированные платформы подготовки НД.

Степень достоверности и апробация результатов

Внедрение результатов исследования:

– Внедрение методологии подготовки НД в практическую деятельность ГБУЗ «НПКЦ ДиТ ДЗМ» – по разработанной методологии было подготовлено 40 НД.

– Внедрение принципов систематизации и стандартизации данных: реестр наборов медицинских данных ГБУЗ «НПКЦ ДиТ ДЗМ».

– Полученные результаты также были внедрены в практическое здравоохранение в виде учебно-методического пособия «Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта» и одноименного учебного пособия.

– Результаты работы внедрены в педагогический процесс ФГБОУ ВО «МИРЭА – Российский технологический университет».

– Результаты диссертационного исследования стали основой национального стандарта Российской Федерации: ГОСТ Р 59921.5-2022 «Системы искусственного интеллекта в клинической медицине. Часть 5. Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов».

Личный вклад автора. Автор определил актуальность, сформулировал тему, цель и задачи диссертационной работы, сформировал дизайн и этапы исследования, установил необходимые методы. Автором был выполнен поиск

литературы по теме диссертационного исследования, изучение НД, находящихся в открытом доступе и созданных в ГБУЗ «НПКЦ ДиТ ДЗМ». Автор принимал участие в процессах создания и использования НД для тестирования СИИ в рамках Эксперимента, создании реестра НД, его актуализации, а также формировании сопроводительной документации, участвовал в формировании и внедрении методики создания НД, разработал схему эксперимента по изучению объема выборки, реализовал его в виде программного кода. Также были проведены анализ и интерпретация полученных результатов.

Степень достоверности результатов. Разработанная автором методика была апробирована и успешно внедрена в Эксперимент (40 НД по рентгенографии/флюорографии (РГ/ФЛГ), ММГ, КТ (компьютерной томографии) и МРТ (магнитно-резонансной томографии)). Анализ данных проводился с применением современных подходов и методов. Статистическая обработка данных, анализ зависимостей, оценка распределений, ROC-анализ выполнялись с использованием языков программирования R и Python.

Апробация результатов исследования. Основные результаты работы были представлены и обсуждены на конференциях международного, всероссийского и регионального уровней:

1. VI Форум «Онлайн диагностика 3.0», 14–16 апреля 2022 г., г. Москва;
2. XIV международный конгресс «Невский радиологический форум – 2023», 7–8 апреля 2023 г., г. Санкт-Петербург;
3. III Российский диагностический саммит, 4–6 октября 2023 г., г. Москва;
4. Научно-практическая конференция по медицинской визуализации «Онлайн-диагностика 24», 28–30 марта 2024 г., г. Москва;
5. «Искусственный интеллект и Радиомика: от диагностики к лечению», 17 мая 2024 г., г. Москва;
6. IX Всероссийская научно-практическая конференция по Искусственному интеллекту в здравоохранении и системам поддержки принятия врачебных решений ITM-AI, 6–7 февраля 2025 г., г. Москва.

Соответствие паспорту научной специальности. Цели и задачи данной работы соответствуют следующим пунктам специальности 3.3.9. «Медицинская информатика»:

– п. 1. Информационное, математическое и компьютерное моделирование в медицине.

– п. 3. Разработка компьютерных методов, баз данных и программных средств для получения, накопления, обработки, передачи и систематизации медицинских и экологических данных с целью использования в лечебно-диагностическом, реабилитационном, профилактическом, образовательном процессах.

– п. 7. Информатизация клинической практики. Элементы деятельности медицинского работника как объект информатизации. Структуризация и формализация медицинской информации.

– п. 9. Инженерия медицинских знаний в области извлечения информации, концептуализации, визуализации и формализации знаний. Разработка баз знаний для использования в лечебно-диагностическом и образовательном процессах.

– п. 12. Системы управления медицинскими данными и знаниями в исследовательской и клинической деятельности, в медицинском образовании.

– п. 16. Разработка методов, алгоритмов и информационных технологий для управления здравоохранением. Создание моделей, алгоритмов и информационных технологий для построения регистров по направлениям медицины.

Публикации. По материалам диссертационного исследования опубликовано 10 печатных работ в отечественных и зарубежных изданиях, из них 3 – в изданиях, рекомендованных ВАК при Минобрнауки России по специальности 3.3.9. Медицинская информатика, 6 – в изданиях, входящих в международные базы данных Web of Science и Scopus, 1 – иные статьи в изданиях, входящие в перечень ВАК при Минобрнауки России. Получено 42 патента на базы данных.

Структура и объем работы. Текст диссертации изложен на 138 страницах, состоит из введения, главы с обзором литературы, главы о материалах и методах исследования, главы о результатах исследования, заключения, выводов, практических рекомендаций, списка цитируемой литературы (124 источника) и 1 приложения. Диссертация включает 6 таблиц, 27 рисунков.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Наборы данных как основа создания и оценки технологий искусственного интеллекта

Технологический прогресс и развитие информационно-коммуникационных технологий в конце XX века привели к лавинообразному нарастанию данных, так называемому информационному взрыву, и, как следствие, возникновению новых инструментов для работы с большими данными и даже появлению новых профессий, таких, например, как исследователь, аналитик и инженер данных. Не стала исключением и медицина – помимо использования различных информационных систем и систем поддержки принятия врачебных решений, в последние годы в нее внедряются ТИИ, способные описывать исследования и прогнозировать течение заболеваний. ИИ – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые как минимум с результатами интеллектуальной деятельности человека [2]. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, ПО (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений [2]. ТИИ – технологии, основанные на использовании ИИ, включая компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы ИИ [2]. Частным случаем ТИИ является машинное обучение, процесс автоматического обучения и совершенствования поведения системы ИИ на основе обработки массива обучающих данных без явного программирования [23], которое требует введения репрезентативного, релевантного и корректно размеченного НД [2]. Кроме того, алгоритмы нейронных сетей, создаваемые в машинном обучении, представляют собой «черный ящик» и крайне сложны для интерпретации, что может послужить

причиной низкого доверия к ним и препятствовать их развитию. Поэтому одним из важнейших принципов использования ТИИ является прозрачность – объяснимость их работы и процесса достижения ими результатов, недискриминационный доступ пользователей к информации о применяемых алгоритмах [2].

Одним из наиболее успешных и перспективных направлений, достигшим определенных успехов в этой области, является медицинская диагностика, в частности – визуализационные методы, такие как патогистологические и лучевые [12, 24]. В качестве примера успешного внедрения ТИИ в лучевую диагностику можно привести Эксперимент [6]: за 3 года было проанализировано ИИ-сервисами (ИИ-сервис – СИИ, участвующая в Эксперименте) более 11 млн исследований по 29 направлениям (под направлением понимается модальность лучевого исследования в сочетании с целевой патологией). В дальнейшем это позволило внедрить СИИ в систему обязательного медицинского страхования [11]. Стоит отметить, что реализация такого масштабного проекта стала возможной благодаря созданию ЕРИС ЕМИАС (Единый радиологический информационный сервис Единой медицинской информационно-аналитической системы). ЕРИС – это сервис хранения и обработки медицинских изображений с возможностью получения врачебной интерпретации исследований [25]. С 2020 г. ЕРИС является частью ЕМИАС, предназначенной для автоматизации учета и анализа медицинских данных, а также обеспечения оказания медицинских услуг населению г. Москвы. В ЕРИС хранятся все диагностические лучевые исследования, выполненные в медицинских организациях Департамента здравоохранения г. Москвы. К сервису подключены рабочие места врачей-рентгенологов и рентгенолаборантов, а также имеются дополнительные инструменты для решения аналитических и управленческих задач [26].

После интеграции ЕРИС в ЕМИАС данные лучевых исследований стали доступны населению в электронной медицинской карте. ЕРИС ЕМИАС является важнейшим инструментом цифровизации здравоохранения не только как

централизованное хранилище данных и автоматизированная среда для работы врача-рентгенолога, но и управленческий инструмент, позволяющий проводить анализ и аудит текущего состояния службы лучевой диагностики и принимать организационные решения для оптимизации распределения ресурсов. Кроме этого, централизованность, стандартизация и цифровизация, которые обеспечил ЕРИС ЕМИАС, позволили проводить научные исследования и разработки и оперативно внедрять их в практическую деятельность МО (медицинских организаций). Так, был реализован Эксперимент, что позволило в том числе за счет внедрения ТИИ решать приоритетные задачи здравоохранения [27].

Успешное внедрение ТИИ в практическую деятельность зависит от достижения диагностической точности, не уступающей врачам. Кроме того, необходимо учитывать также вопросы эффективности, стоимости, доступности и медицинской этики. Однако к решению этих задач можно приступать тогда, когда диагностическая точность СИИ будет сопоставима с точностью врача. Для этого требуется критическая и независимая оценка, которая должна основываться на стандартизированной методологии, учитывающей большое количество параметров, а также на единых принципах интерпретации и терминологии. Такая оценка может быть обеспечена так называемой внешней (независимой) валидацией – это тестирование СИИ на группе новых пациентов с целью определения ее критериев диагностической точности [28]. К сожалению, часто при разработке СИИ такая оценка проводится не всегда или характеризуется плохим дизайном, неправильной обработкой, отсутствием калибровки, что приводит к завышению показателей диагностической точности [28, 29]. Примером качественной оценки СИИ является Эксперимент: прежде чем внедрить ИИ-сервис в практическую деятельность, проводится многоэтапное тестирование для оценки функциональных возможностей и критериев диагностической точности на независимых наборах данных, созданных в ГБУЗ «НПКЦ ДиТ ДЗМ» [6]. На первом этапе проводится функциональное тестирование с целью оценки возможности работы ИИ-сервиса: корректного отображения всех требуемых и заявленных параметров.

При успешном прохождении функционального тестирования ИИ-сервис допускается к калибровочному, в процессе которого происходит оценка критериев диагностической точности, сопоставление с заявленными показателями, вычисление оптимального порога срабатывания (порог cut-off – такая вероятность наличия целевого признака, при достижении которой принимается решение о наличии этого признака) с последующим заключением о возможности допуска ИИ-сервиса в практическую деятельность. То есть валидационным тестированием в контексте Эксперимента является калибровочное. Кроме того, в дальнейшем проводится регулярный технологический (проверка на наличие различных типов дефектов) и клинический (проверка корректности описания исследования и заключения) мониторинг его работы [16, 30].

Одним из основных принципов развития ТИИ является прозрачность – объяснимость работы ИИ и процесса достижения им результатов [2]. Для реализации этого принципа необходимы понятные, легко интерпретируемые критерии оценки. Именно поэтому для проведения калибровочных тестирований в Эксперименте был выбран метод ROC-анализа (англ. receiver operating characteristic – рабочая характеристика приемника). ROC-анализ используется для оценки способности классификационной модели различать между собой классы. Как правило, в медицинских исследованиях это наличие или отсутствие патологии, признака (бинарная классификация) или степень тяжести заболевания, классификация признака (мультиклассовая классификация). Любая задача мультиклассовой классификации может быть сведена к бинарной. ROC-кривая – это график, на котором отображается соотношение между чувствительностью (вероятность правильно классифицировать наличие целевого признака) и 1 – специфичностью (вероятность правильно классифицировать отсутствие целевого признака) СИИ при различных порогах. Пример ROC-кривой представлен на рисунке 1 (построена с помощью открытого инструмента построения ROC-кривых [31]): площадь под ROC-кривой = 0,942, доверительный интервал

рассчитан двумя способами: с помощью бутстрэппинга (bootstrapping) и по методу ДеЛонга (DeLong).

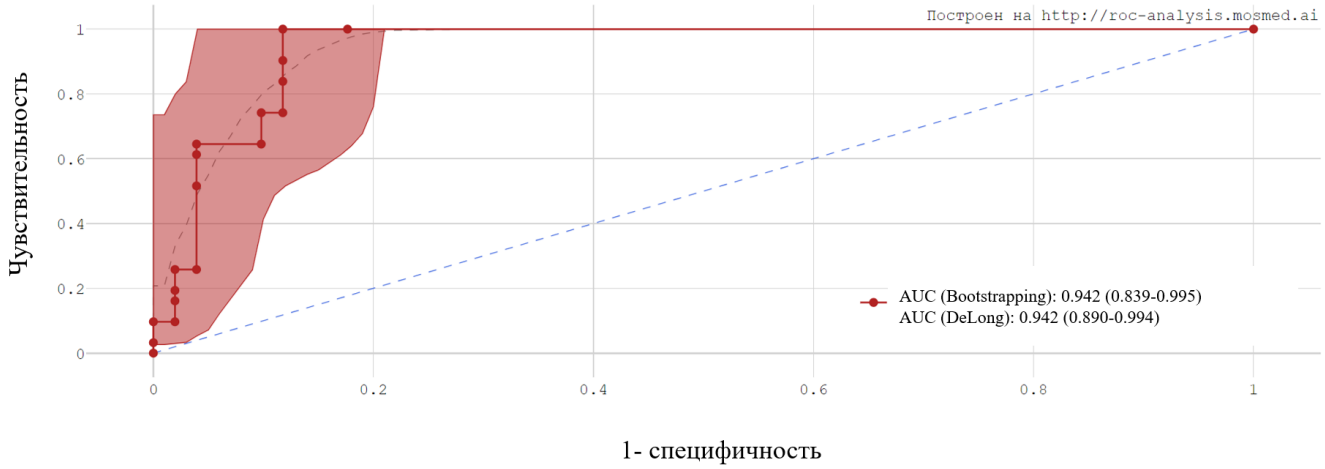


Рисунок 1 – Пример ROC-кривой: красная линия – ROC-кривая, точки на линии – значение чувствительности и соответствующей ей специфичности при заданной вероятности, полученной от СИИ, пунктирная линия – аппроксимация ROC-кривой, диагональ – наихудшее значение

ROC-анализ позволяет выбрать оптимальный порог (cut-off) для классификации. При этом происходит построение матрицы ошибок, которая обеспечивает расчет не только чувствительности и специфичности, но и других критериев оценки СИИ: точность, прецизионность, полнота, F-мера, прогностическая ценность положительного и отрицательного результата.

Еще одним способом оценки СИИ является их сравнение, однако различия в эталонных стандартах, возможностях тестировщиков, диагностике заболеваний и пороговых значениях очень затрудняют прямое сравнение исследований и алгоритмов [17]. Это возможно улучшить только с помощью хорошо спланированных исследований, в которых четко рассматриваются вопросы, касающиеся прозрачности, воспроизводимости, этики и эффективности [32], а также конкретных стандартов отчетности для исследований ИИ. Такие стандарты отражены в чек-листах [21, 22], где особое внимание уделяется описанию НД. Тем не менее ROC-анализ также может использоваться для сравнительной оценки СИИ. Например, при реализации алгоритмов симуляции выборки, таких как

бутстрэп, а также с помощью метода ДеЛонга [33] можно рассчитать доверительный интервал для каждой точки ROC-кривой или проводить статистическое сравнение двух кривых с помощью перестановочного теста [34]. Такие методы реализованы и представлены в веб-инструменте на сайте https://roc-analysis.mosmed.ai/ [35].

Исходя из вышесказанного, НД являются основой не только для создания, но и тестирования, а также изучения ТИИ. Трудно переоценить их значение для развития ТИИ. На гистограмме распределения публикаций реферативной базы данных PubMed по запросу «artificial intelligent radiology» (рисунок 2) наблюдается резкий рост количества публикаций после 2017 г. Возможно, он связан в том числе и с появлением первых размещенных в открытом доступе НД по лучевой диагностике [36]. Открытые НД предоставляют исследователям и разработчикам доступ к большим объемам реальных медицинских данных, что позволяет создавать и улучшать алгоритмы машинного обучения и модели ИИ. В результате этого возникает острая необходимость не только в методике оценки их диагностической точности, но и в новых НД, предназначенных для этой оценки.

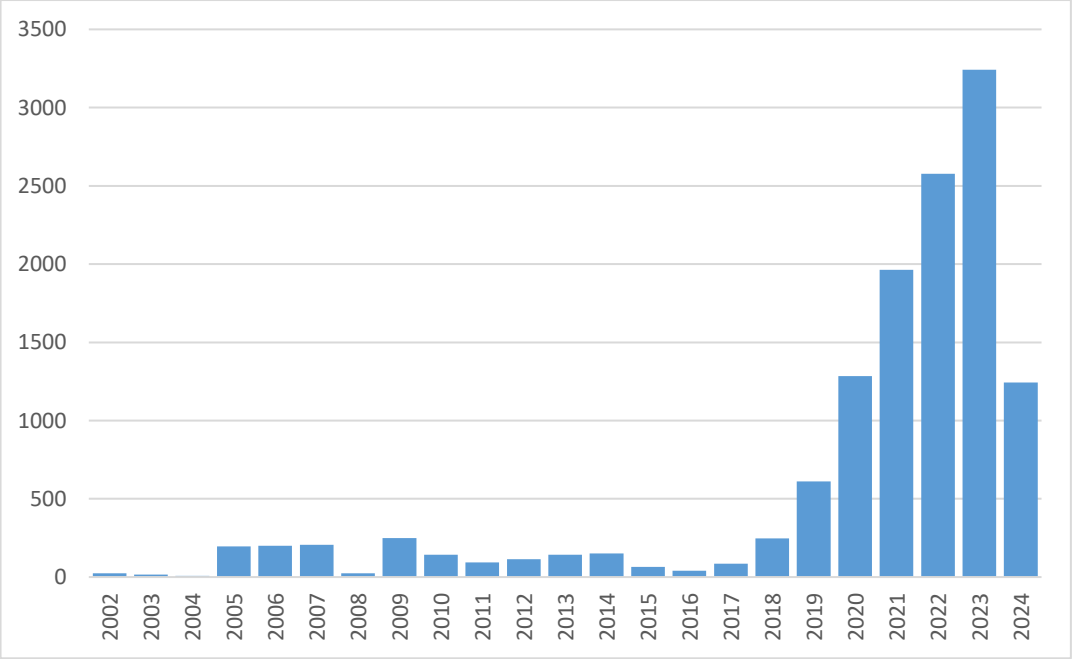


Рисунок 2 – Гистограмма распределения по годам публикаций по запросу «artificial intelligent radiology» («искусственный интеллект в радиологии») реферативной базы данных PubMed

1.2 Проблемы создания и использования наборов данных в лучевой диагностике

Одним из самых сложных и важных вопросов при проведении валидационного тестирования или исследований по оценке и сравнению СИИ является создание качественного, репрезентативного, релевантного НД. НД – состав данных, которые структурированы или сгруппированы по определенным признакам, соответствуют требованиям законодательства Российской Федерации и необходимы для разработки программ для электронных вычислительных машин на основе ИИ [2]. Данные – информация, представленная в формализованном виде, пригодном для ее передачи, интерпретации и обработки с участием человека или автоматическими средствами [38]. Также необходимо отметить, что зачастую используется термин «база данных», который имеет множество различных определений в зависимости от области использования [39], тем не менее в контексте машинного обучения общепринятым является термин «набор данных» (или англицизм «датасет», от английского «dataset»), который и используется в рамках данной диссертационной работы.

В обзоре R. Aggarwal и соавторов было отмечено, что проблемы с качеством и объемом НД – самый распространенный недостаток при анализе диагностической точности СИИ [17]. Создание НД для лучевой диагностики – сложный, многоступенчатый процесс, требующий учета множества аспектов и вовлечения большого количества специалистов из разных областей. Это обусловлено следующими причинами:

1. Отсутствие структурированности медицинской информации, ее сложность и неоднозначность [40].

Успешное применение ТИИ основывается на медицинских понятиях, требующих стандартизации и нормализации [41]. Существуют различные подходы к стандартизации медицинской информации. Наиболее известным является международная классификация болезней (МКБ). Однако есть и другие

справочники и классификаторы для удобства представления данных и обеспечения электронного обмена медицинской информацией:

– Современная процедурная терминология (Current Procedural Terminology) – набор кодов, описаний и руководств, предназначенных для описания процедур и услуг, выполняемых врачами и другими поставщиками медицинских услуг [42].

– Систематизированная номенклатура медицинских клинических терминов SNOMED – комплексная многоязычная клиническая терминология здравоохранения, обеспечивающая единообразное представление клинического содержания в электронных медицинских картах [43].

– Имена и коды логических идентификаторов наблюдений LOINC – терминологический стандарт для медицинских измерений, наблюдений и документов [44].

– Словарь радиологии RadLex – структурированный словарь радиологических терминов, включая соответствующую анатомию, заболевания и результаты визуализации [45].

Тем не менее они имеют ряд ограничений [46], редко используются и не решают полностью проблему стандартизации медицинской текстовой информации.

Данные медицинской визуализации являются более структурированными, что обеспечивается стандартом формата данных DICOM (Digital Imaging and COmmunications in Medicine). DICOM-файлы имеют объектно-ориентированную структуру с теговой организацией, где представлены серии изображений и сопровождающая информация – метаданные. Такая стандартизация хранения данных открывает возможности использования СИИ, например для оценки качества проведенных исследований, а также для автоматизации процесса отбора данных [47, 48].

2. Сложность и высокая стоимость разметки исследований.

Разметка (аннотация) – этап обработки структурированных и

неструктурированных данных, в процессе которого данным (в том числе текстовым документам, фото- и видеоизображениям) присваиваются идентификаторы, отражающие тип данных (классификация данных), и (или) осуществляется интерпретация данных для решения конкретной задачи, в том числе с использованием СИИ [49]. Необходимо отметить, что в некоторых зарубежных работах понятия разметки и аннотации разделяют, относя первую к размещению какой-либо информации (в том числе текстовой) на изображении, а вторую определяют как пояснительную или описательную информацию, непосредственно связанную с изображением [50]. Разметку текстовой информации можно доверить алгоритмам обработки естественного языка (например, MedLabel [51]), а для разметки диагностических изображений (сегментация, локализация) существует специальное ПО (3D Slicer, ITK Snap и другие [52]). Кроме того, разрабатывается ПО, способное упростить и ускорить процесс разметки специалистами. Например, в работе [53] представлен инструмент сегментации изображений, который обучается на первых размеченных изображениях и предлагает уже готовую разметку новых. Такой алгоритм позволил сократить время разметки на 80 %. Тем не менее результаты работы такого ПО должны быть перепроверены и скорректированы врачами-специалистами. В работе [54] подчеркивается важность выбора стратегии разметки исследований, при этом отмечается, что заключения врача-рентгенолога в качестве разметки необходимо использовать с осторожностью, так как они субъективны и могут основываться не только на данных целевого исследования, но и на дополнительных клинических параметрах. Кроме того, в зависимости от метода верификации и класса разметки могут потребоваться данные медицинской карты [55].

Необходимо также отметить, что для качественной разметки недостаточно мнения одного специалиста. В ряде работ, где описывается процесс создания НД, мало внимания уделяется описанию процесса разметки: иногда количество разметчиков на одно исследование и их квалификация вовсе не указываются [19], или это один специалист [56], что приводит к большому количеству ошибок и

снижению качества НД. Для того чтобы этого избежать, следует привлекать к разметке нескольких квалифицированных специалистов [57], что также увеличивает стоимость исследования. Кроме того, нужно учитывать требования к этим специалистам: образование, навыки, опыт работы, прохождение специализированного обучения [49]. Так, в работе [58] описана стратегия аннотации эндоскопических изображений при болезни Крона: 4 специалиста производили разметку в 3 этапа, в работе [18] разметка изображения одним специалистом проверялась двумя экспертами, а в работе [59] результаты офтальмоскопии аннотировали 22 эксперта. Вопрос поиска квалифицированных специалистов и определения стратегии разметки стоит настолько остро, что существуют специальные платформы, предоставляющие экспертов для решения задач в области разработки СИИ в медицине [60].

При разметке медицинских данных также необходимо помнить о сложности структуры медицинских терминов и их семантических связях [46], чтобы аннотации не превратились в данные, которые придется дополнительно структурировать. Для этого в процессе создания НД необходимо тщательно продумывать не только форму представления итоговых данных, так чтобы она имела машиночитаемый вид, но и промежуточные данные на этапах разметки. С этой целью создается специальное ПО (например, [50]), которое позволяет автоматизировать процесс разметки, тем самым упрощая и ускоряя создание НД и повышая его качество.

3. Этический вопрос и защита персональных данных.

Сбор медицинских данных следует осуществлять с согласия пациентов, которые должны быть надлежащим образом проинформированы об использовании их данных для исследовательских или аналитических целей во избежание злоупотребления, включая использование в коммерческих целях, или неправомерного доступа к информации о здоровье пациентов.

Сбор, хранение и использование этих данных должны сопровождаться мерами безопасности для защиты конфиденциальности и приватности пациентов.

Нарушение конфиденциальности и утечка данных могут привести к серьезным последствиям для пациентов и нарушению их прав, поэтому в соответствии с ФЗ о защите информации персональные данные должны быть обезличены [37]. Кроме того, согласно Национальной стратегии необходимо создание общедоступных баз данных, так называемых библиотек наборов данных. Для осуществления этого процесса необходимо тщательно изучить этические и законодательные аспекты с целью защиты персональных данных при публикации [61].

Обезличивание (псевдонимизация, деперсонализация) персональных данных – действия, в результате которых становится невозможным определить принадлежность персональных данных конкретному субъекту персональных данных без получения дополнительной информации [37]. Согласно рекомендациям Роскомнадзора [62] существуют 4 способа обезличивания:

– введение идентификаторов;
– изменение состава персональных данных или семантики;
– декомпозиция;
– перемешивание.

Анонимизация – действия, в результате которых происходит безвозвратная утрата способности данных быть связанными с конкретным субъектом, даже если будет использована какая-то дополнительная информация [63].

Этический вопрос и защита персональных данных являются чрезвычайно важными аспектами создания НД, особенно учитывая тенденции к увеличению количества данных в открытых и закрытых источниках, распространению технологий, использующих эти данные, и задачи Национальной стратегии, требующие формирования публичных НД для поддержания конкуренции и развития ТИИ. Все это обуславливает потребность в специализированных анонимизаторах – ПО, позволяющем удалять или заменять персональные данные. Кроме того, в работе [61] регламентируются процессы обработки НД с целью их публикации и предлагается создание специализированного «паспорта НД»,

в котором будут указываться способы и уровни доступа к данным в зависимости от их содержания в соответствии с законодательством.

4. Отсутствие стандартизированных методологий формирования НД.

На сегодняшний день создано уже большое количество НД в лучевой диагностике, однако крайне мало работ, посвященных обобщению и структуризации информации, необходимой для создания НД. Отсутствие четких стандартов и описания этапов приводит к снижению качества НД, скорости его создания и повышению стоимости. Кроме того, формирование алгоритма действий и стандартов способствует автоматизации процесса, что является одной из приоритетных задач Национальной стратегии [2].

Среди попыток стандартизировать описание НД можно отметить чек-лист оценки медицинских изделий на основе ИИ [22], который включает в себя следующие пункты, касающиеся создания НД:

– качество данных;

– данные демографии (пол, возраст, эпидемиология, этническая принадлежность, социоэкономический статус);

– параметры разметки (количество разметчиков, опыт, специализация, инструкция разметки, критерии согласия);

– описание данных (тип, статистическое описание);

– кодирование признаков (использование стандартов представления медицинской информации);

– анализ выбросов и пропусков;

– предобработка;

– балансировка;

– данные о доступе.

В работе [64] предлагается ряд вопросов, на которые необходимо ответить в процессе создания НД. Вопросы сгруппированы по разделам согласно жизненному циклу НД:

– композиция (планирование);

– сбор данных;

– предварительная обработка/очистка/маркировка;

– использование;

– распространение;

– поддержка.

В работе [65] освещены основные принципы создания НД в патогистологических исследованиях, которые также разделены на следующие группы:

– определение целевых показателей выборки;

– сбор данных (включая аннотирование, проверку качества, генерацию синтетических данных);

– определение размера НД (объема выборки);

– определение репрезентативности НД (стратификация, смещение данных);

– разделение (на обучающую и тестовую подвыборки);

– составление отчетов;

– нормативные требования.

Вышеперечисленные проблемы также могут являться причинами создания некачественных НД, что является серьезным вызовом, который может влиять на их достоверность и полезность для научных и медицинских исследований, а также для разработки СИИ. Качество данных можно определить как совокупность свойств и характеристик НД, которые влияют на его способность удовлетворять потребностям, возникающим в результате предполагаемого использования данных [66]. Литература по качеству данных охватывает ряд аспектов, среди которых наиболее часто упоминаются точность, полнота, согласованность и актуальность [67].

1. Полнота – степень, в которой все необходимые данные, которые могли бы быть зарегистрированы, действительно были зарегистрированы [66]. Некоторые данные могут отсутствовать из-за недостаточной записи, утери, программных

ошибок при извлечении или неполноты сведений, что может снизить ценность итогового НД.

2. Точность – степень, в которой зарегистрированные данные соответствуют истине [66]. Они могут содержать ошибки, опечатки или неточности из-за неправильного ввода информации или неверной интерпретации медицинской документации.

3. Актуальность – степень, в которой данные отвечают текущим и потенциальным потребностям пользователей [68]. В некоторых случаях они могут устаревать, и их актуальность для современной медицинской практики или исследований может быть снижена.

4. Согласованность – степень, в которой данные могут быть успешно сопоставлены как между собой, так и с другими источниками данных [68]. Создание единых стандартов и общих методологий способствует согласованности данных, в том числе и с течением времени.

Кроме того, важнейшим аспектом оценки качества НД является контекст его использования, то есть цель, с которой он создается [69]. Одни и те же данные могут использоваться по-разному – эта контекстуальная изменчивость усложняет процесс оценки и требует детального понимания конкретных обстоятельств, связанных с использованием данных.

Для преодоления проблем качества данных в медицинских наборах можно использовать методы очистки данных, валидации, нормализации, стандартизации, а также проверять данные на наличие ошибок и создавать инструменты контроля и повышения качества данных.

Одним из ключевых моментов также является наличие четкой документации и метаданных для облегчения интерпретации и использования НД. Так, в работе [70] авторы приводят пример необъективной разметки данных, связанной с индивидуальной интерпретацией врачами понятий «норма» и «патология» и отсутствием четких указаний в сопроводительной документации (техническом задании) относительно критериев отнесения к классам. Такие неточности привели

к занижению показателей дегенеративных заболеваний суставов, что в итоге отразилось на возможности использовать этот НД для обучения алгоритмов распознавания этой группы заболеваний.

1.3 Подходы к определению размера (объема выборки) набора данных и баланса классов

Проведение любого исследования начинается с постановки цели и планирования эксперимента. Создание набора данных – не исключение. Наряду с основными параметрами, такими как модальность исследования, целевая патология, анатомическая область, на первых этапах необходимо определить объем выборки, то есть количество исследований в планируемом НД. Анализ мировой литературы показал, насколько это сложная и неоднозначная задача [71]. Кроме того, зачастую в исследованиях вовсе не производится расчет необходимого объема выборки: в обзоре [17] отмечается, что из 279 работ по оценке диагностической точности СИИ в медицинской визуализации практически ни в одном исследовании не был рассчитан требуемый для проведения статистического анализа объем НД. Чтобы свести к минимуму вероятность ложных корреляций между целевым и дополнительными параметрами, наборы тестовых данных должны быть большими и разнообразными [72], но в то же время необходимо найти оптимальный баланс с реалистичными затратами и усилиями, что является сложной задачей [65].

Ключевыми параметрами, характеризующими качество НД, являются корректность данных, входящих в его состав (в т. ч. корректность разметки), и репрезентативность, т. е. способность отражать свойства целевой популяции. Репрезентативность, в свою очередь, обеспечивается наличием характерных для целевой популяции признаков и их соотношением. Безусловно, учесть все признаки популяции невозможно ввиду их огромного количества, а также с учетом того, что популяция – это живая система, которая постоянно меняется. Тем не

менее при решении конкретных задач необходимо выделить ключевые параметры популяции, требующие обоснования представленности в выборке, и в дальнейшем формировать ее с учетом этих параметров. В зарубежной литературе большое внимание уделяют качественному составу выборки: полу, возрасту, расе, географии распределения, этническим группам [73–76]. Также для обучающих выборок необходимо учитывать тип алгоритма, его архитектуру, количество параметров, желаемый уровень производительности, качество данных, распределение функций и шум. Зачастую при определении размера выборки используют так называемое «правило 10»: размер выборки должен превышать количество параметров в 10 раз [77]. Однако искусственная балансировка выборки по множеству параметров может привести к повышению предвзятости, смещению данных и другим систематическим ошибкам [78, 79]. Стратификация выборки – это очень сложный процесс, и он не всегда осуществим на практике, а увеличение количества параметров приводит к так называемому «проклятию размерности». Кроме того, прежде чем переходить к более сложным задачам, необходимо решить вопросы, возникающие в простейших случаях. Простейшим случаем в данном контексте является выборка, балансируемая по одному параметру – целевая патология/признак, при этом остальные параметры распределяются в ней случайным образом. Два основных вопроса, возникающих при формировании этой выборки для оценки диагностической точности: «Сколько необходимо единиц выборки?» и «Какой должен быть ее баланс?».

Существует множество методик расчета минимального объема выборки в различных статистических исследованиях, начиная от использования эмпирических таблиц [80, 81], формул [82] и номограмм [83] и заканчивая различными методиками математического моделирования [84]. Такой огромный разброс методов и их неоднозначность обуславливаются отсутствием единого стандарта и большим разнообразием медицинских задач, поэтому каждый раз перед исследователем встает вопрос, какой конкретно метод нужно применить для решения текущей задачи.

При проведении исследований по сравнению изучаемых признаков с целью оценки диагностической точности (например, сравнение процента поражения легких на КТ при COVID-19, рассчитанного врачом и СИИ [85]) в двух и более группах существуют методы расчета, основанные на заданном уровне статистической значимости (ошибка первого рода) и мощности (ошибка второго рода) (формулы 1–5).

$$n = \frac{(t_\alpha * \sqrt{2 * p * q} + t_\beta \sqrt{p_1 * q_1 + p_2 * q_2})^2}{(p_1 - p_2)}, \quad (1)$$

где n – количество исследований, необходимое в каждой группе,

p_1 – доля изучаемого признака в группе № 1,

p_2 – доля изучаемого признака в группе № 2,

q_1 и q_2 – обратные p_1 и p_2 величины,

p и q – соответственно:

$$p = \frac{p_1 + p_2}{2} \quad (2)$$

$$q = 1 - p, \quad (3)$$

где t_α – коэффициент Стьюдента, соответствующий уровню статистической значимости,

t_β – коэффициент Стьюдента, соответствующий значению мощности.

Можно также провести расчет на основе известных данных о дисперсии целевого признака, полученных из литературных источников или пилотного исследования:

$$n = \frac{2 * s^2 * (t_\alpha + t_\beta)^2}{(\bar{x}_1 - \bar{x}_2)^2}, \quad (4)$$

где $\bar{x}_1$ – среднее значение исследуемой величины в первой группе,

$\bar{x}_2$ – среднее значение исследуемой величины во второй группе,

s^2 – величина, определяемая по уравнению:

$$s^2 = \frac{(n_1-1)*s_1^2+(n_2-1)*s_2^2}{n_1+n_2-2}, \quad (5)$$

где n_1 и n_2 – объем выборки в каждой из групп пилотного исследования,

s_1 и s_2 – дисперсии в каждой из групп пилотного исследования.

Гораздо сложнее обстоит вопрос с размером НД при проведении ROC-анализа, который является одним из основных методов оценки диагностической точности СИИ.

Наиболее часто в литературе ссылаются на работы, где предлагаются фиксированные объемы выборки для прогностических моделей: от 200 до 400 исследований [86–88]. В работах [86, 88] в основе определения количества исследований лежит достижение 80 % мощности при установлении параметров калибровки модели, а также рассмотрено поведение калибровочных кривых при различных объемах выборок.

Калибровочные кривые – зависимость между предсказанными вероятностями и фактической наблюдаемой частотой положительного класса в задаче бинарной классификации. На рисунке 3 [89] представлены примеры калибровочных кривых в координатах: наблюдаемая доля значений с положительным целевым признаком от прогнозируемого результата. Калибровку, т. е. построение калибровочных кривых с определением коэффициентов перехвата (intercept) и наклона (slope), производят для оценки качества и настройки прогностических моделей. Идеальная модель стремится к главной диагонали. В вышеперечисленных работах были также определены различные уровни калибровки от «слабого» до «сильного», и объем выборки 400 исследований является минимальным для «слабой» калибровки [86, 88].

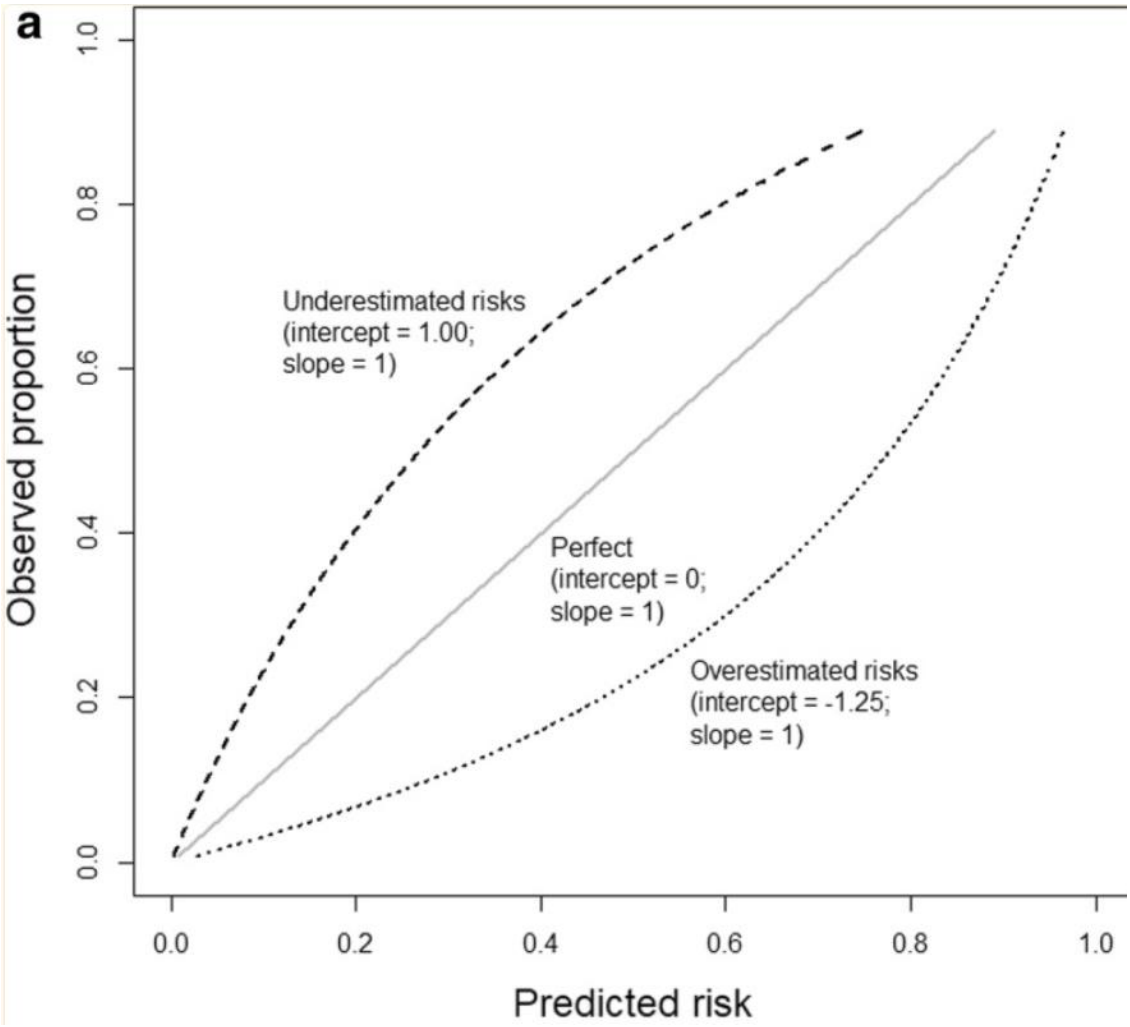


Рисунок 3 – Построение калибровочной кривой «Зависимость наблюдаемой доли значений с положительным целевым признаком (observed proportion) от прогнозируемого результата (predicted risk)» [89]: сплошная линия – идеальная модель (perfect), крупный пунктир – недооценивающая (underestimated) модель (предсказывает больше ложноотрицательных значений), мелкий пунктир – переоценивающая (overestimated) модель (предсказывает больше ложноположительных значений), intercept – коэффициент перехвата, slope – наклон

В работе [87] также изучались параметры калибровки, но объем выборки определялся исходя из коэффициента охвата и ширины доверительного интервала для 10 вариантов объема выборки (от 185 до 17 528) с балансом классов генеральной совокупности. Ожидаемо наилучшие результаты показал НД с наибольшим количеством исследований. Тем не менее авторы пришли к выводу, что наименьшим объемом выборки для валидационных тестирований является

100 единиц НД с целевым признаком при сохранении баланса классов генеральной совокупности.

Однако многие авторы были не согласны с данными работами и предлагали свои методики определения объема выборки, основанные на более сложных расчетах и других показателях. Следует отметить, что одним из наиболее удобных и наглядных методов анализа данных при тестировании является ROC-анализ [90]. Это важно для простой и применимой на практике интерпретации качества моделей, понимания принципов настройки порога принятия решения и анализа ошибок модели для предоставления данных с целью дальнейшей ее доработки. Поэтому интерес представляют работы, основанные на данном методе. Например, R. D. Riley и соавторы предложили несколько методик расчета объема выборки исходя не только из показателей калибровочных кривых, но и площади под характеристической кривой (ROC AUC) и чистой выгоды [91]. Данные способы расчета основаны на заданных целевых показателях точности. В частности, в примере, приведенном в статье, задается доля ожидаемых результатов и доверительный интервал, а авторы отмечают, что они определяются исследователями при планировании эксперимента и зависят от контекста.

M. Pavlou и соавторы в своем исследовании выводят ряд формул для расчета минимального объема выборки на основе заданной точности и мощности для показателей калибровки, AUC ROC и прецизионности (precision) и апробируют их на данных имитационного моделирования [84].

Еще одним немаловажным вопросом при проведении внешней валидации является баланс классов (норма и патология) в выборке, по которой происходит валидация [91]. Существует 2 наиболее популярных подхода: использовать в валидационной выборке такой же баланс классов, как в популяции, или применять соотношение 1:1. Первый вариант моделирует реальную картину распределения признака, однако частота встречаемости какого-либо заболевания может быть неизвестна, неправильно оценена или быть настолько низкой, что объем выборки для валидационного исследования будет очень большим, чтобы

туда попало хотя бы одно исследование с целевым признаком и, как следствие, качество модели на такой выборке объективно оценить будет невозможно. Кроме того, остро стоит вопрос определения понятия «генеральная совокупность». Если под ним понимается все мировое сообщество, то определить частоту встречаемости признака не представляется возможным. Если же выделяется определенная популяция (население страны, региона, города), то частота встречаемости признака будет зависеть и меняться от многих факторов, таких как миграция, эпидемиологическая обстановка, сезонность и т. д. Ситуация, когда выборка сбалансирована в соотношении 1:1, дает более объективные метрики на сравнительно небольших объемах выборки, однако не отражает реальное распределение признака.

1.4 Хранение, использование и публикация наборов данных

Несмотря на отсутствие единой методологии создания НД, существуют рекомендации и чек-листы по их описанию [21, 22]. В отдельных исследованиях согласно рекомендациям и требованиям научных издательств описывается процесс создания наборов данных под конкретные, узкие задачи [18, 19]. При этом вопрос стандартизированной методологии формирования НД остается актуальным и нераскрытым [20]. Еще меньше внимания уделяется процессу хранения и использования НД. Многие авторы отмечают не только неоднородность самих медицинских данных, но и сложность и отсутствие единообразия в организации НД [92–94]. Как отмечалось выше, разметка одного исследования – это сложный, дорогостоящий и трудозатратный процесс; кроме того, к медицинским данным имеет доступ ограниченный круг лиц. В связи с этим необходима рационализация использования имеющихся данных, которая обеспечивается за счет организации процессов хранения и использования НД, что также соответствует принципу разумной бережливости Национальной стратегии [2]. Данные могут использоваться многократно для решения однотипных (тестирование СИИ) и

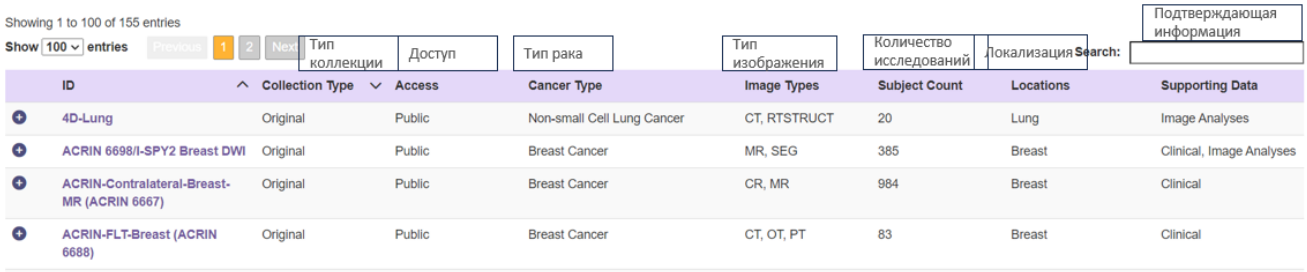
принципиально новых научных задач и задач разработки. Для этого необходимо создание управленческих инструментов, позволяющих не только хранить систематизированную информацию о НД, но и осуществлять процессы контроля качества на всех этапах его жизненного цикла. Обеспечить прозрачность процессов хранения и использования НД призвана в том числе и представленная в работе [95] система смены версионности при внесении изменений в НД. Кроме того, необходимо обеспечить долгосрочное, безопасное, централизованное хранение НД.

Тем не менее зачастую различные МО имеют локальные схемы кодирования, затрудняющие обмен данными за пределами учреждения [96]. Для обеспечения обмена данными между организациями с целью развития научных исследований и коммерческих разработок в области ТИИ в РФ и поддержки конкуренции необходимо создание библиотек НД. Они не только предоставляют доступ к НД, но и позволяют наглядно организовать их хранение в виде карточек. Несмотря на это, большое количество данных по-прежнему хранится разрозненно, а многие библиотеки представляют собой неструктурированное описание НД [94], как, например, одна из самых популярных библиотек kaggle.com [97] (рисунок 4).

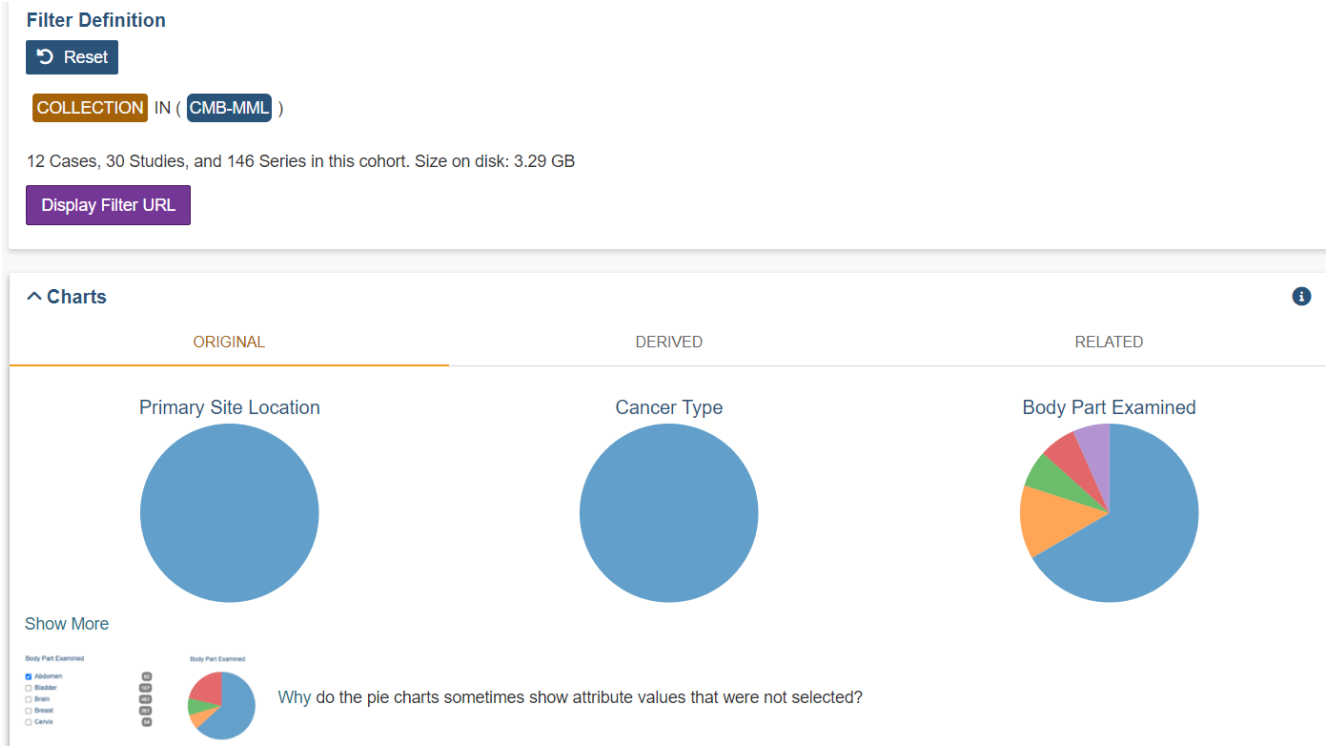
В библиотеке kaggle.com отсутствуют специализированные фильтры для поиска необходимой информации среди сотен различных наборов данных. Сами карточки представляют набор неструктурированной информации, важная для разработчиков информация зачастую отсутствует, встречаются также ошибки в описании.

Рисунок 4 – Карточка набора данных в библиотеке kaggle.com [97]

Более удобной и наглядной является библиотека Национального института рака (США) [98]: благодаря структурированному каталогу (рисунок 5а) исследователь или разработчик может быстрее найти интересующий его НД, а изучив карточку (рисунок 5б), оценить применимость НД в его исследовании или разработке. Однако в процессе изучения библиотеки обращает на себя внимание низкая полнота заполняемости данных в карточках. Для контроля точности и полноты заполняемости карточек библиотек НД также необходимы инструменты, позволяющие систематизировать и автоматизировать этот процесс.



а



б

Рисунок 5 – Библиотека наборов данных Национального института рака (США) [98]: а – каталог библиотеки, б – карточка набора данных

Резюмируя вышесказанное, можно выделить следующие основные направления для совершенствования процессов создания, хранения и использования НД (таблица 1).

Таблица 1 – Проблемы, возникающие при создании, хранении и использовании НД и пути их решения

Проблема	Что сделано	Что не сделано
Сложность, неструктурированность, неоднозначность медицинских данных	Введение МИС*. Введение справочников, классификаторов. Алгоритмы обработки естественного языка.	Взаимосвязь между справочниками и классификаторами. Не все данные классифицированы и стандартизированы.
Сложность и высокая стоимость разметки данных	Создание специализированного ПО, упрощающего процесс разметки. Создание синтетических данных, аугментация данных.	Единая методология, упрощающая процесс разметки и способствующая автоматизации части рутинных действий при аннотации, а также созданию инструментов контроля качества данных. Обоснование минимального размера и оптимального баланса классов НД с целью оптимизации ресурсов по их разметке без потери качества НД.
Этические аспекты и безопасность данных	Законодательные акты, регламентирующие защиту персональных данных. Инструменты анонимизации данных.	Работа с НД, обогащенными клинической информацией, а также предоставление доступа к данным.
Отсутствие унифицированных методологий и стандартов создания НД	Чек-листы и рекомендации по созданию наборов данных. Единичные алгоритмы под конкретные задачи.	Создание стандартизированной методологии формирования НД с целью дальнейшей автоматизации процессов и создания платформ подготовки НД.
Ограниченное количество данных для публичного использования, неструктурированность и низкое качество данных	Создание библиотек открытых наборов данных.	Унификация информации о НД (в т.ч. сопроводительной документации) и создание инструментов контроля качества и систематизации данных, в том числе с целью дальнейшей автоматизации процессов сбора сопроводительной информации и формирования карточек библиотек НД.

*МИС – медицинские информационные системы.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Общий ход и этапы исследования

Работа включала 4 этапа, соответствующих задачам исследования (рисунок 6), которые представили собой аналитическую (этапы 1, 2, 4) и экспериментальную (этап 3) части.

Этап 1. Был изучен российский и зарубежный (научные публикации, НД в открытом доступе, библиотеки НД) опыт формирования и использования НД. Также был изучен опыт создания НД на базе ГБУЗ «НПКЦ ДиТ ДЗМ», в том числе в рамках Эксперимента. Проводился сбор информации о процессах планирования, сбора данных, анонимизации, разметки, использования, хранения НД, а также изучение сопроводительной информации, документации и публикаций в открытом доступе.

Этап 2. Полученная на первом этапе информация была проанализирована, систематизирована и стандартизирована. Также был проведен анализ российских и зарубежных справочников и медицинских номенклатур. По результатам анализа был сформирован жизненный цикл НД, принципы классификации НД, перечень справочников и номенклатур, актуальный для создания и использования НД в области лучевой диагностики, реестр НД, а также разработан алгоритм формирования НД.

Этап 3. Был разработан дизайн эксперимента по изучению AUC ROC в зависимости от объема выборки на результатах работы СИИ в Эксперименте. На основании полученных данных и последующего их анализа была разработана методика определения объема выборки для валидационных тестирований.

Этап 4. Результаты предыдущих этапов сформированы в единую методологию создания, хранения и использования НД и внедрены в практическую деятельность ГБУЗ «НПКЦ ДиТ ДЗМ» в рамках Эксперимента, а также для научных исследований в области ТИИ.

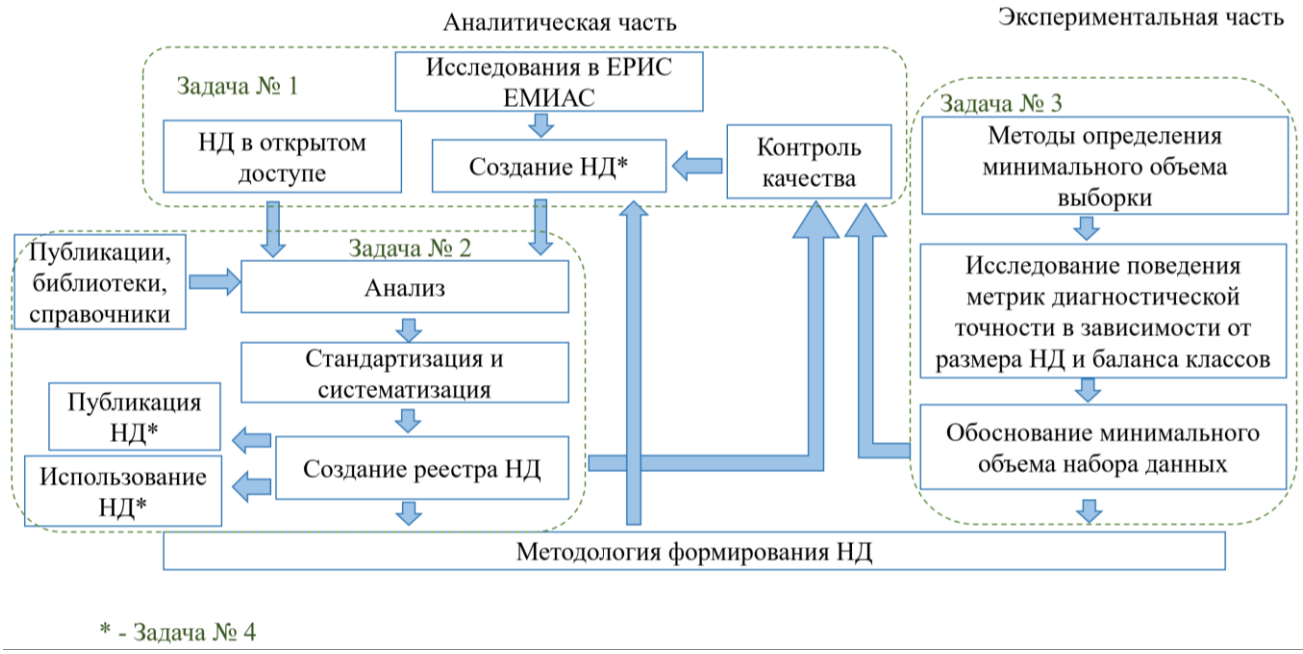


Рисунок 6 – Схема исследования

2.2 Материалы

Источники данных

На первых этапах настоящего исследования проводился поиск и анализ следующей информации:

1. Научные публикации по ключевым словам: «наборы данных», «база данных», «медицинские данные», «датасет», «реестр», «dataset», «medicine database», «registry», размещенные в реферативных базах данных РИНЦ, Scopus, Web of Science с 2017 по 2022 г.

2. Медицинские справочники, приказы и ГОСТы:

– Федеральный справочник инструментальных диагностических исследований [113];

– Федеральный справочник анатомических локализаций [114];

– Тезаурус радиологических терминов RadLex [45];

– Систематизированная машинно-обрабатываемая медицинская номенклатура SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [43];

– База данных для идентификации медицинских врачебных и лабораторных наблюдений LOINC [44];

– МКБ-10, алфавитный указатель к международной статистической классификации болезней и проблем, связанных со здоровьем;

– Справочник услуг ЕРИС;

– Стандарт DICOM [99];

– Номенклатура медицинских услуг;

– Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» [2];

– Федеральный закон от 27.07.2006 № 149-ФЗ «Об информации, информационных технологиях и о защите информации» [37];

– Указ Президента Российской Федерации от 07.05.2018 № 204 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года» [27];

– ГОСТ 34.320-96 «Информационные технологии. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной базы» [38].

3. Библиотеки медицинских НД, репозитории, базы данных и НД, находящиеся в открытом доступе, их наименования, сопроводительная информация и организация поиска, хранения и представления информации о них:

– https://paperswithcode.com/dataset/luna – набор изображений компьютерной томографии для проверки алгоритмов автоматического обнаружения легочных узлов;

– https://medpix.nlm.nih.gov/home – Национальная медицинская библиотека MedPix;

– https://portal.imaging.datacommons.cancer.gov/collections/ – база данных Национального института рака США;

– https://stanfordmlgroup.github.io/competitions/mura/ – база данных скелетно-мышечных рентгенологических исследований;

– http://www.oasis-brains.org/ – открытая библиотека серий изображений МРТ;

– http://imaging.cancer.gov/programsandresources/informationsystems/lidc – база данных компьютерной томографии легких;

– http://academictorrents.com/details/557481faacd824c83fbf57dcf7b6da9383b3235a – набор цифровых рентгенограмм грудной клетки;

– http://www.cancerimagingarchive.net/ – база данных различных типов рака с различными методами визуализации;

– http://braintumorsegmentation.org/ – база данных изображений магнитно-резонансной томографии для сегментации опухолей головного мозга;

– https://github.com/ – портал для разработчиков IT-проектов;

– https://www.kaggle.com/ – портал для специалистов по обработке данных и машинному обучению.

Изучалась структура библиотек и карточек НД (структурированное/ неструктурированное представление информации), а также параметры описания НД на предмет их применимости и целесообразности использования в ГБУЗ «НПКЦ ДиТ ДЗМ» в рамках задач по созданию и использованию НД лучевой диагностики.

4. Опыт ГБУЗ «НПКЦ ДиТ ДЗМ» по созданию и использованию НД [6, 95], сопроводительная информация (технические задания, требования, readme-файлы, карточки НД и прочие документы), номенклатура наименований НД. Предмет исследования – организация процесса подготовки НД, формирование НД. Объект исследования – НД.

Для создания НД на базе ГБУЗ «НПКЦ ДиТ ДЗМ» использовались лучевые исследования из ЕРИС ЕМИАС, который представляет собой систему, состоящую из диагностических устройств, PACS (Picture Archiving and Communication System – система архивирования и передачи) и автоматизированных рабочих мест (АРМ) врачей-рентгенологов. Результаты лучевых исследований, полученные рентгенолаборантами на диагностических устройствах, в стандартизированном

формате DICOM отправляются в PACS, откуда они доступны на АРМ врача-рентгенолога. Международный стандарт хранения и обмена медицинскими изображениями DICOM позволяет стандартизировать и обеспечить совместимость медицинских изображений и соответствующей информации с различных диагностических устройств, что необходимо для хранения и передачи данных между МО. Это обеспечивает эффективное совместное использование медицинской информации, улучшает диагностику, обеспечивает целостность и безопасность медицинских данных и способствует развитию и внедрению телемедицины и ТИИ. DICOM-файлы имеют объектно-ориентированную структуру с теговой организацией. В тегах представлены кадры или серии изображений, а также метаданные – сопроводительная информация (пол, возраст, вид исследования, диагностическое устройство, МО и т. д.) [99]. Врач описывает исследование, формирует заключение в виде текстового протокола и также загружает в PACS в формате DICOM. Таким образом, ЕРИС ЕМИАС позволяет хранить результаты диагностических исследований с их описанием, сформированным врачом-рентгенологом, в стандартизированном формате. Необходимо также отметить, что все каналы передачи и хранения информации защищены специальными протоколами передачи данных.

Системы искусственного интеллекта

В качестве СИИ в данной работе использовались ИИ-сервисы, участвующие в Эксперименте: «Цельс» (ООО «Медицинские скрининг системы», имеет регистрацию как медицинское изделие: РЗН 2021/14449) и «Трио ДМ» (АО «МТЛ») по направлениям: анализ маммографических исследований с целью выявления различных образований и классификации их по BI-RADS, автоматический анализ рентгеновских снимков органов грудной клетки с целью выявления признаков различных заболеваний.

НД, создаваемые по разработанной методологии, использовались для тестирования ИИ-сервисов при допуске в Эксперимент.

Набор данных

Для решения задач третьего этапа были созданы три НД. Данные (исследования с текстовыми протоколами врачей-рентгенологов, результаты работы ИИ-сервисов в виде бинарной разметки о наличии или отсутствии патологии) были получены из ЕРИС ЕМИАС. Период, за который проводились исследования, и объем выборки обусловлены временем работы выбранного ИИ-сервиса. Верификация проводилась по текстовым протоколам заключений врачей-рентгенологов с помощью алгоритма естественной обработки языка (MedLabel [51]). Общими критериями включения исследования для всех НД были: наличие ответа от ИИ-сервиса, наличие описания заключения от врача-рентгенолога. Перед использованием данные были предварительно обезличены.

НД1: 123 301 маммографическое исследование и результаты работы одного ИИ-сервиса. Данные были получены за период с 01.09.2021 по 27.12.2021. Маммографические исследования классифицировались по наличию (условно «патология») и отсутствию (условно «норма») признаков злокачественных новообразований молочной железы. При верификации НД по текстовым протоколам (согласно классификации по методу верификации – таблица 3) анализировались выставленные значения по шкале BI-RADS [100]: 0 – в случае диагностирования врачом 1-го или 2-го класса BI-RADS («норма») и 1 – BI-RADS 3, 4, 5 («патология»). Изначально баланс классов в исследовании составлял: «норма» 89,3 % / «патология» 10,7 %. Критерием исключения являлось отсутствие классификации по BI-RADS в тексте заключения.

НД2: 143 710 маммографических исследований с результатами работы ИИ-сервиса (отличного от первого НД), полученных за период с 01.02.2022 по 31.10.2022. Критерии исключения и принципы классификации аналогичны НД1. Изначально баланс классов в исследовании составлял: «норма» 88,8 % / «патология» 11,2 %.

НД3: 62 142 рентгенологических исследования с результатами работы ИИ-сервиса. Данные получены за период с 25.10.2023 по 21.11.2023.

Рентгенологические исследования классифицировались по наличию и отсутствию следующих патологических признаков органов грудной клетки: плевральный выпот, пневмоторакс, очаг затемнения, инфильтрация, консолидация, диссеминация, полость, ателектаз, кальцинат, расширение средостения, кардиомегалия, нарушение целостности кортикального слоя. Изначально баланс классов в исследовании составлял: «норма» 88,4 % / «патология» 11,6 %.

2.3 Методы

Дизайн исследования:

– для этапа 3 (обоснование объема выборки НД для тестирования СИИ): обсервационное ретроспективное когортное исследование; методы: статистические, ROC-анализ, анализ типа распределения, Фурье-анализ;

– для остальных этапов (1, 2, 4, 5): аналитическое исследование; методы: анализ, синтез, индукция, дедукция.

Обезличивание

Разметка данных производилась специалистами, имеющими доступ к ЕРИС ЕМИАС, в защищенном контуре. Обезличивание данных осуществлялось специалистами, также имеющими доступ к ЕРИС ЕМИАС, на двух виртуальных машинах с помощью специально разработанного инструмента, который удалял персональные данные из DICOM-тегов согласно таблице уровня атрибутов конфиденциальности [101], а также присваивал исследованиям новые уникальные идентификаторы, чтобы удалить связь между идентификатором исследования и персональными данными пациента. Инструмент реализован на языке Python.

Анализ текстовых протоколов

На первых этапах создания НД отбор исследований проводился вручную специалистом (врачом-рентгенологом) путем вычитки заключений в ЕРИС ЕМИАС в ходе описания исследований. В дальнейшем процесс был автоматизирован с помощью инструмента обработки естественного языка.

Инструмент реализован на языке Python и зарегистрирован в качестве программы для ЭВМ [51].

Программное обеспечение

Предобработка, анализ и структурирование данных, формирование сопроводительного текстового файла (readme) при создании НД, исследование поведения AUC ROC в зависимости от объема выборки проводились на языке Python версии 3.8.8 в среде Jupyter Notebook с помощью библиотек Pandas, Erislib, Numpy, Scikit-learn, Statmodels. Определение типа распределения, Фурье-анализ, а также построение графиков осуществлялись с помощью библиотек fitdistrplus, ggplot2, actuar языка R 4.3.3 в программной среде R-studio v.2023.06.1. Реестр НД реализован в виде таблицы Microsoft Excel.

Исследование поведения площади под характеристической кривой в зависимости от объема выборки и баланса классов

Из каждого НД, описанного в п. 2.2, формировались выборки с заданным балансом классов (отношение числа исследований с «патологией» к исследованиям с «нормой») и объемом (количество исследований). Изучались следующие балансы классов: 0,5 (50 % «нормы» и 50 % «патологии»), 0,4 (60 % «нормы» и 40 % «патологии»), 0,3 (70 % «нормы» и 30 % «патологии»), 0,2 (80 % «нормы» и 20 % «патологии»), 0,1 (60 % «нормы» и 10 % «патологии»). Необходимо отметить, что баланс 0,1 сопоставим с долей патологии исходных НД (10,7 %, 12,6 %, 13,1 % для НД1, НД2 и НД3 соответственно). Изучались следующие объемы выборки: от 30 до 25 000 с шагом в 10 (верхняя граница обусловлена ограничением вычислительных мощностей).

Каждая выборка с заданным размером и балансом классов формировалась 10 000 раз (бутстрэп [102]) из исходного НД. Для каждой выборки рассчитывалась площадь под характеристической кривой и усреднялась для одного баланса и размера. На выходе получался массив данных зависимости усредненной AUC ROC для каждого баланса классов. Подробная схема исследования представлена на рисунке 7.

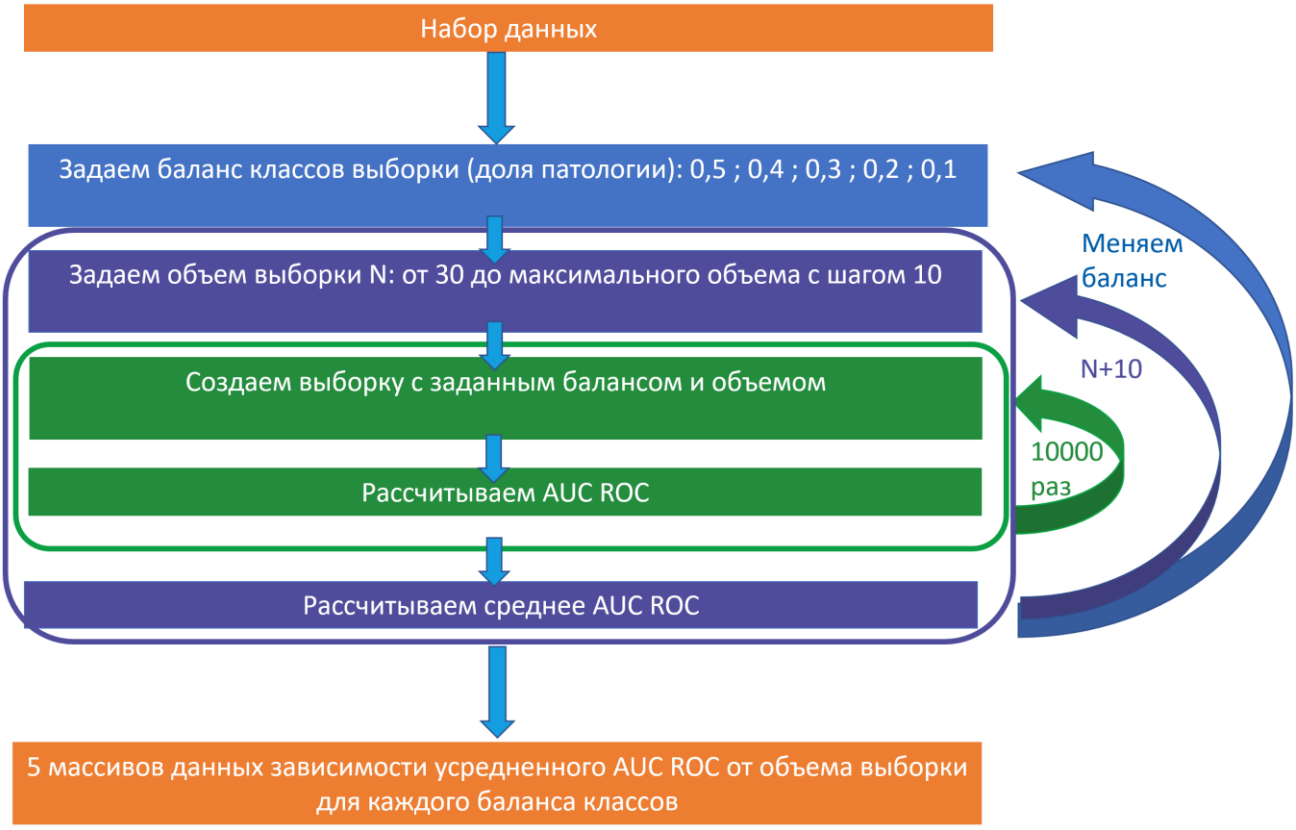


Рисунок 7 – Схема исследования по изучению зависимости площади под характеристической кривой ROC от объема и баланса классов выборки

Далее проводился Фурье-анализ значений AUC ROC в зависимости от количества данных для каждого баланса классов, затем – анализ наиболее близкого распределения полученных данных. Были рассмотрены 10 различных распределений:

1. Нормальное (Гауссово):

$$f(AUC) = \frac{1}{sd*\sqrt{2\pi}}\exp\left(-\left(\frac{(\mathrm{AUC}-\mu)^2}{2*sd^2}\right)\right), \quad (6)$$

где AUC – полученные значения AUC ROC,

μ – среднее значение AUC ROC,

sd – среднее квадратичное отклонение AUC ROC .

2. Логарифмически нормальное (логнормальное):

$$f(AUC) = \frac{1}{\mathrm{AUC}\cdot sd\cdot\sqrt{2\cdot\pi}}\exp\left(-\frac{(ln(AUC)-\mu)^2}{2\cdot sd^2}\right) \quad (7)$$

3. Экспоненциальное:

$$f(AUC) = \lambda \cdot exp(-\lambda \cdot AUC), \quad (8)$$

где λ= 1/μ – обратное математическое ожидание.

4. Пуассона:

$$f(AUC) = \frac{\mu^k}{k!} \cdot exp(-\mu), \quad (9)$$

где k – количество событий (исследований).

5. Коши:

$$f(AUC) = \frac{1}{\pi \cdot s \cdot \left[1+\left(\frac{\mathrm{AUC}-x_0}{s}\right)^2\right]}, \quad (10)$$

где s – масштабный коэффициент,

x_0 – коэффициент сдвига.

6. Гамма:

$$f(AUC) = \frac{1}{sd^{\alpha} \cdot \Gamma(\alpha)} \cdot AUC^{\alpha-1} \cdot exp\left(-\frac{AUC}{sd}\right), \quad (11)$$

где Γ (α) – гамма функция Эйлера,

α– параметр гамма функции.

7. Логистическое:

$$f(AUC) = \frac{1}{s} \cdot \frac{exp\left(-\frac{(AUC-\mu)}{s}\right)}{\left(1+exp\left(-\frac{(AUC-\mu)}{s}\right)\right)^2}, \quad (12)$$

где s − дисперсия логистического распределения.

8. Биноминальное:

$$f(AUC) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \quad (13)$$

где n – общее количество исследований,

k – количество исследований с данными AUC,

p – вероятность данного значения AUC.

9. Геометрическое:

$$f(k) = p \cdot (1-p)^k \quad (14)$$

10. Вейбулла:

$$f(\sigma) = \frac{a}{b} \cdot \left(\frac{AUC}{b}\right)^{(a-1)} \cdot exp\left(-\left(\frac{AUC}{b}\right)^a\right), \quad (15)$$

где a – коэффициент формы распределения Вейбулла,

b – коэффициент масштаба распределения Вейбулла.

Параметры каждого из распределений вычислялись методом максимального правдоподобия [103]. Далее для оценки однородности значений AUC ROC был проведен анализ коэффициента вариации в зависимости от количества исследований. В случае распределения Коши коэффициент вариации рассчитывался по уравнению:

$$K = \frac{\gamma}{x_0}, \tag{16}$$

где γ – масштабный параметр в распределении Коши,

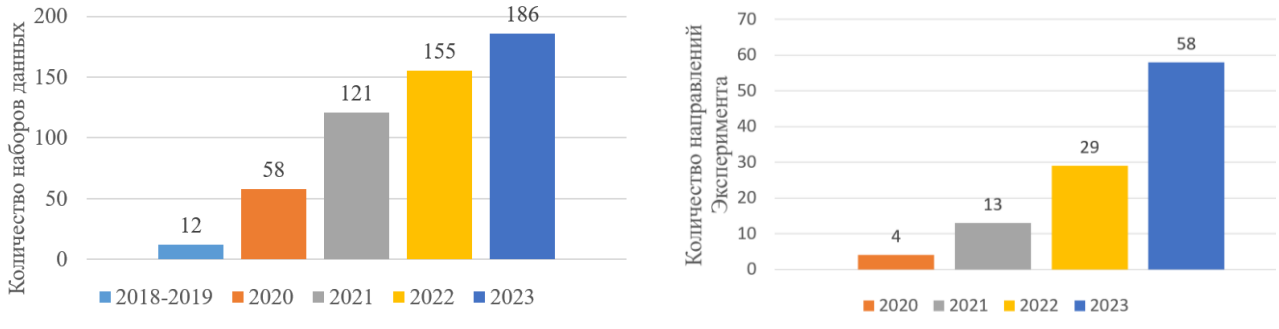
x_0 – параметр сдвига в распределении Коши.

ГЛАВА 3. РЕЗУЛЬТАТЫ

3.1 Управляемость, надежность и устойчивость процессов формирования наборов данных

На основании изученных публикаций, библиотек и НД были сформулированы основные проблемы, возникающие при создании и использовании НД. Были выявлены их причины и поставлены задачи для дальнейшей оптимизации процесса формирования НД.

Основной вклад в поиск причин возникновения ошибок внес собственный опыт создания и использования НД в ГБУЗ «НПКЦ ДиТ ДЗМ». До 2020 г. НД формировались преимущественно для научных задач в небольших количествах, однако с началом Эксперимента в 2020 г. потребовалось создавать больше НД с целью тестирования СИИ. В условиях роста количества направлений (модальностей и исследуемых патологий) и задач становилось труднее организовывать процессы управления и контроля качества НД. На рисунке 8а представлен график роста числа НД в ГБУЗ «НПКЦ ДиТ ДЗМ»: ежегодно количество созданных НД (включая смену версии) увеличивается минимум на 20 %, что обусловлено появлением новых направлений в Эксперименте (рисунок 8б), требующих НД для проведения валидационных тестирований.



а б

Рисунок 8 – Динамика количества и разнообразия данных в ГБУЗ «НПКЦ ДиТ ДЗМ»: а – количество наборов данных, созданных с 2018 по 2023 г., б – количество направлений Эксперимента с 2020 по 2023 г.

На первых этапах процесс создания НД в ГБУЗ «НПКЦ ДиТ ДЗМ» отличался невысокой устойчивостью к изменениям и недостаточной надежностью. Так, отсутствие опыта и структурированного алгоритма формирования НД приводило к большому количеству вопросов у исполнителей и ошибок, а следовательно, значительному числу итераций в процессе создания и использования (в т. ч. публикации для НД с открытым доступом). При появлении новых направлений или внесении изменений в готовые НД требовалось включение до 20 специалистов (инженеров, научных сотрудников, врачей, экспертов, руководителей и т. д.), тогда как в дальнейшем их количество сокращалось до 13. Отсутствие этапности и четких протоколов сбора данных приводило к постоянному пересмотру хода работы, возвращению на более ранние итерации, возникновению ошибок, сдвигу сроков работ и увеличению числа участников, в частности руководителей, для обеспечения процессами управления, так как ход работы над созданием НД был крайне неустойчив к возникающим изменениям (под изменениями понимаются изменения и уточнения, возникающие как в ходе создания НД, так и при работе над новым направлением). Отбор исследований осуществлялся врачом вручную при выполнении рутинных описаний исследований. При таком способе количество анализируемых исследований на этапе отбора зависело от частоты встречаемости целевого патологического признака: в случае если частота встречаемости составляла 10 %, то для получения 100 исследований с наличием этого признака был необходим анализ 1000 исследований, что требовало существенных затрат времени и ресурсов. В дальнейшем был разработан инструмент анализа текстовых протоколов, который значительно сократил процесс сбора данных и положил начало автоматизации отдельных этапов создания НД. Так, точность алгоритма анализа текстовых протоколов для поиска признаков, характерных для COVID-19, составила 99,8 %, а для рака молочной железы – 100 % [104].

Кроме того, важным этапом стало создание библиотеки открытых НД mosmed.ai. Тем не менее при публикации НД также возникали ошибки (рисунок 9),

а сам процесс публикации (в частности, внесение информации в карточку каталога) занимал много времени.

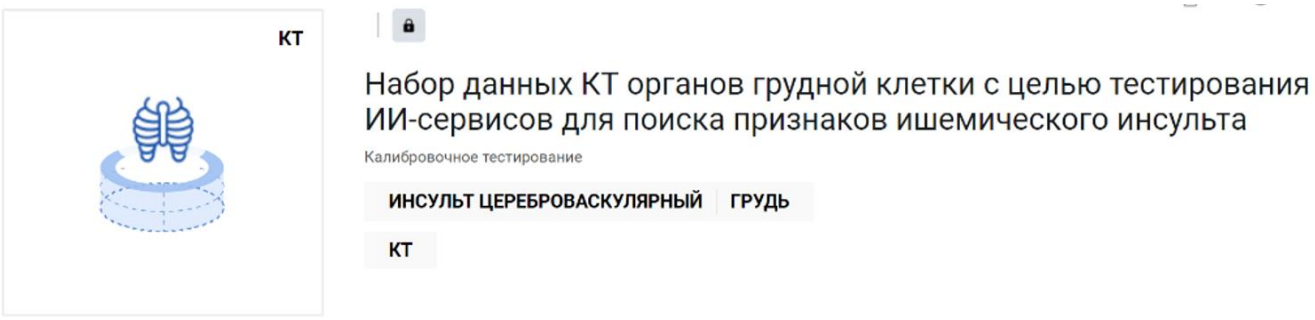
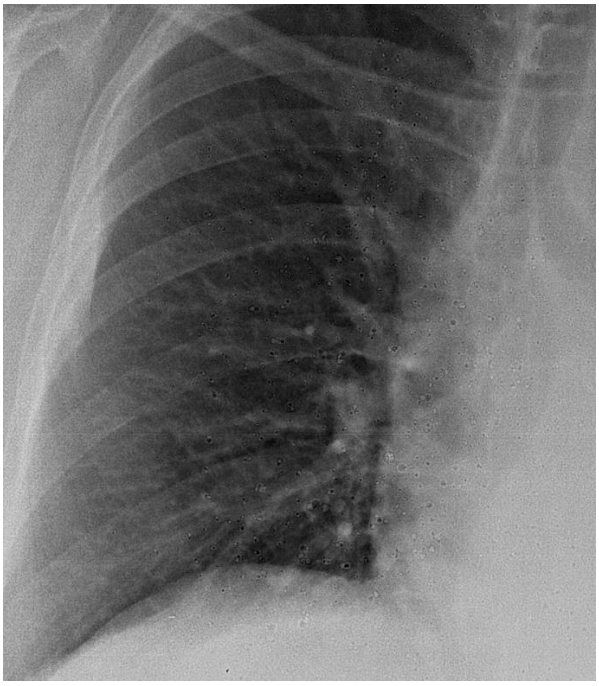


Рисунок 9 – Пример ошибки при заполнении карточки библиотеки НД: неверное указание анатомической локализации (грудь) для НД с признаками ишемического инсульта

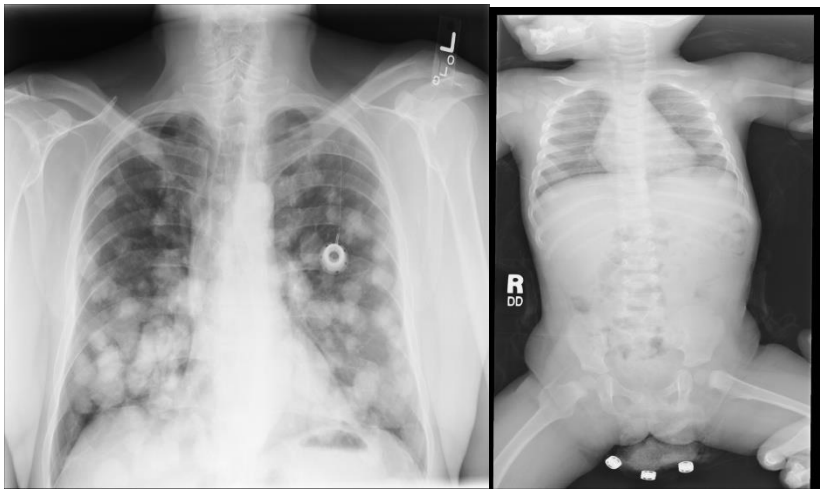
С ростом количества направлений Эксперимента, а также научных исследований в ГБУЗ «НПКЦ ДиТ ДЗМ» в целом становилось все труднее осуществлять процессы управления при создании и использовании НД. На первых этапах для этого использовались стандартные инструменты управления проектами, в частности «Битрикс24», однако они не адаптированы под специфику работы с наборами данных. Информация находилась в разрозненных хранилищах, таблицах, «задачах» (пространство «Битрикс24», в которых ведется работа), визуализация была затруднена, информация дублировалась. В результате выявленных уже в процессе использования ошибок до 30 % исследований требовало пересмотра и до 50 % – замены. Самое большое количество версий одного НД, полученное в результате замены и пересмотра исследований, – 10.

Следует отметить, что изучение НД, находящихся в открытом доступе, также выявило ряд серьезных ошибок, однако объективно оценить устойчивость и управляемость процессов формирования таких НД не представляется возможным, так как оценка проводится ретроспективно по доступным данным. Тем не менее одним из ярких примеров ошибок в НД явился NIH Chest X-rays dataset [105]. В описании НД указано, что он содержит 112 120 рентгенограмм грудной клетки, частично размеченных врачами, частично с помощью алгоритма обработки

естественного языка. Врачебная экспертиза выборочных исследований показала, что в НД допущено большое количество пропусков патологий, а также присутствовали исследования детей, исследования с артефактами и дефектами, часть исследований дублировалась (рисунок 10).



а



б в

Рисунок 10 – Примеры некачественных исследований в наборе данных NIH Chest X-rays dataset: а – исследование с артефактами (зернистость), б – тотальная диссеминация помечена как «без патологии», в – рентгенография ребенка

Таким образом, в ходе анализа процессов формирования и использования НД на первом этапе Эксперимента в 2020 г., а также других, находящихся в открытом доступе, были выявлены следующие проблемы:

– отсутствие стандартизированного алгоритма создания НД;

– отсутствие методов и инструментов автоматизации сбора данных и анонимизации данных;

– отсутствие специальных инструментов контроля качества НД;

– отсутствие специальных инструментов управления процессами создания, хранения и использования НД;

– разрозненное хранение данных;

– отсутствие единой стандартизированной терминологии, принципов наименования и классификации НД и их сопроводительной документации.

В ходе выполнения диссертационного исследования был сформулирован жизненный цикл НД и разработаны:

– методика формирования НД лучевой диагностики;

Методы формирования унифицированных названий НД

– принципы систематизации и классификации НД и метаинформации,

– реестр наборов данных –инструмент управления и контроля качества процессов создания и использования НД;

– обоснование минимального объема выборки НД для тестирования СИИ.

3.2 Жизненный цикл и алгоритм формирования наборов данных

Жизненный цикл – развитие системы, продукции, услуги, проекта или другой создаваемой изготовителем сущности от замысла до вывода из эксплуатации [106]. Жизненный цикл НД состоит из следующих этапов (рисунок 11):

– инициирования;

– планирования;

– формирования;
– регистрации и публикации;
– использования;
– смены версии;
– удаления и архивации.

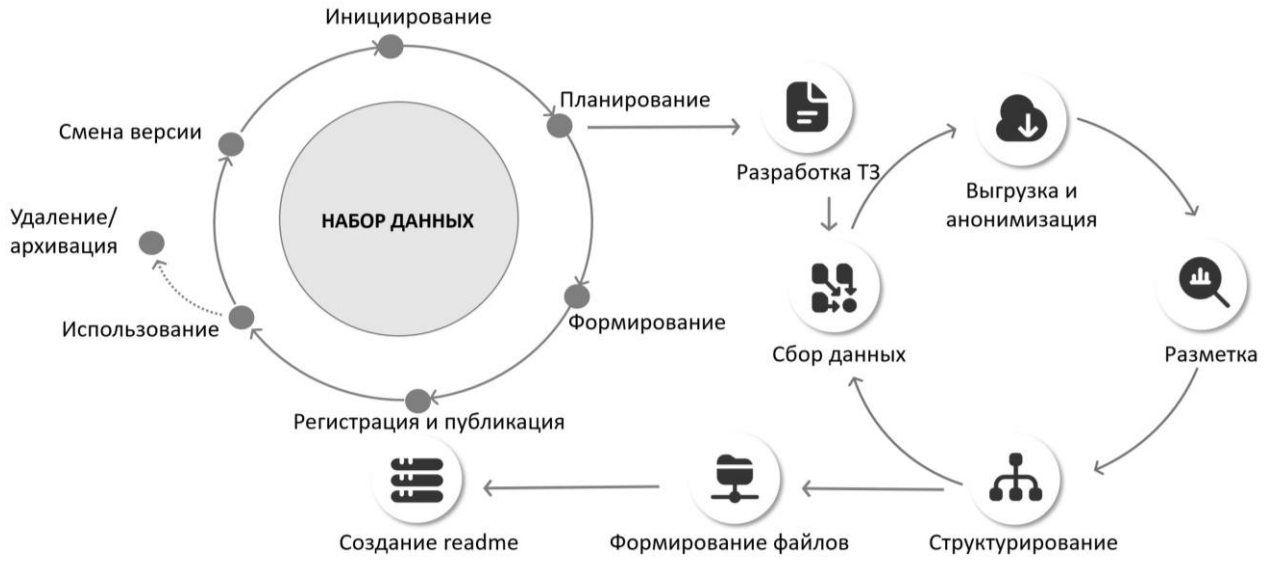


Рисунок 11 – Жизненный цикл НД (слева) и алгоритм формирования НД (справа)

3.2.2 Инициирование

Работа над проектом по созданию НД начинается задолго до начала сбора данных, а именно в тот момент, когда возникает потребность в создании НД, когда появляется какая-либо задача, требующая для своего решения НД, т. е. в первую очередь необходимо обозначить цель. Именно цель определяет дальнейшие ключевые параметры и процессы формирования НД. На основе вариантов цели создания и дальнейшего использования НД предложена классификация на типы:

I. Проведение тестирований для оценки функционала (функциональное тестирование) и критериев диагностической точности, настройки СИИ (калибровочное тестирование) [6].

II. Самотестирование техническое – проведение самостоятельной проверки разработчиками способности СИИ обрабатывать исследования с диагностических устройств разных производителей [107].

III. Самотестирование диагностическое – проведение самостоятельной проверки корректности клинической интерпретации исследований СИИ.

IV. Выполнение клинических испытаний – оценка безопасности и эффективности медицинского изделия согласно [108].

V. Выполнение технических испытаний – оценка соответствия характеристик СИИ требованиям нормативно-правовой, технической и эксплуатационной документации согласно [49].

VI. Проведение разметки текстовых протоколов с помощью программ автоматизированного анализа текстов (например, MedLabel [51]).

VII. Проведение научных исследований.

VIII. Разработка СИИ: обучение и дообучение алгоритмов ИИ.

IX. Обучение специалистов: используются при обучении и проведении тестирований специалистов (например, врачей-рентгенологов).

X. Для категории национальных стандартов с набором данных.

Следует отметить, что во исполнение принципа разумной бережливости Национальной стратегии один и тот же НД может создаваться для решения различных задач (например, проведение научных исследований и разработка ИИ).

Кроме того, на этапе инициирования необходимо определить ответственных за проект по созданию НД, источник финансирования, а также сформировать базовые диагностические требования (БДТ) и базовые функциональные требования (БФТ) [90]. БДТ – это требования к содержащейся в НД информации, необходимой для решения поставленных задач и достижения цели (модальность исследования, целевая патология, критерии отнесения исследований к классам и т. д.). БФТ – это требования к СИИ, включающие в себя описание технических особенностей отображения результатов клинических исследований (серия изображений, толщина срезов, окно визуализации и т. д.) [90]. На основании этих

требований на следующем этапе жизненного цикла формируется техническое задание (ТЗ).

3.2.3 Планирование

Этот этап является самым важным и в дальнейшем определяет эффективность процесса создания НД и качество результата. Именно от того, насколько тщательно будет распланирован процесс создания, распределены ресурсы, оценены риски и назначены инструменты контроля, будет зависеть не только результат, т.е. НД, но и все дальнейшие задачи, для которых он создавался. На данном этапе регламентируются:

– сроки проекта (в том числе поэтапно);
– распределение финансовых ресурсов;
– распределение кадровых ресурсов: необходимо распланировать роли и объем работы руководителя, менеджера проекта, аналитиков, технических специалистов, разметчиков и экспертов [55];
– определение рисков проекта;
– критерии и точки контроля качества проекта.

Кроме того, результатом этапа планирования является ТЗ – основной документ, в котором прописаны все аспекты создания НД. ТЗ регламентирует все процессы так, чтобы на его основе можно было воспроизвести разработку НД [55]. В рамках данной работы был разработан шаблон ТЗ, который содержит следующие ключевые блоки:

– общая информация (название, информация о заказчике и исполнителях, о ресурсах и финансировании, о сроках подготовки и т.д),
– информация о назначении (цель, решаемая задача, направление Эксперимента и т.д),

- информация о процессе сбора данных (критерии отбора данных – наименование и параметры процедуры, даты проведения исследований, половозрастные ограничения, тип медицинской организации и т.д.)
- информация о выгрузке из МИС (используемые инструменты формирования выгрузки, ключевые и фильтрующие слова для поиска по текстовым протоколам, количество необходимых исследований и т.д.)
- требования к обработке данных (требования, способы и инструменты анонимизации, проверка корректности тегов, способы защиты интеллектуальной собственности и т.д)
- информация о разметке (инструмент разметки, требования к разметчикам, критерии согласованности, параметры разметки, критерии брака и т.д.)

Также на этом этапе осуществляется подготовка инструкции и таблицы (или специальной формы) для разметчиков. Инструкция содержит следующие блоки:

- Общая информация: наименование НД и роли разметчиков (врач-разметчик, врач-эксперт и т.д.);
- Порядок разметки: количество размечаемых исследований, стратегия разметки и аудита;
- Критерий бракованных исследований: перечисление признаков, по которым исследования относятся к «браку», не размечаются и не включаются в итоговый набор данных;
- Критерии отнесения к классу «технический дефект»: в случае необходимости наличия класса «технический дефект» (для функциональных тестирований), необходимо указать требования к количеству и критерии отнесения к этому классу;
- Параметры исследования: характеристика исследований, подходящих для разметки(наличие/отсутствие контраста, толщина среза; проекция/плоскость; окно визуализации/режим);
- ПО для разметки;

– Измеряемые величины: подробное описание величин и характеристик, определяемых на исследовании. В соответствии с БДТ и БФТ, необходимо перечислить все величины, которые должен определить или измерить врач. Для каждой из них требуется определить:

а) в задаче классификации: список возможных классов, критерии отнесения исследования к каждому из них с поясняющими иллюстрациями, примерами и т.д.;

б) в задаче выполнения измерения/вычисления: подробная методика проведения измерений с поясняющими иллюстрациями, примерами и т.д., все необходимые формулы с расшифровкой обозначений;

в) необходимо наличие четкой инструкции, что делать с исследованием, если оно нестандартное, но не подходит под критерии брака (наприер, аномалия развития);

Эти данные должны быть подтверждены руководствами или другими литературными источниками.

– Алгоритм работы: пошаговая последовательность действий разметчика. Алгоритм полно и однозначно отражает порядок разметки на всех этапах, от открытия исследования и до сохранения произведенной разметки. Должен включать полный список полей формы разметки, для каждого из которых должна быть описана суть развернуто, чтобы исключить возникновение двусмысленности. В случае наличия более 1 роли, алгоритм должен быть указан для каждой. Для полей также должны быть определены:

а) единицы измерения заполняемой величины (единица измерения не должна содержаться в ячейке значения);

б) точность измерения — количество знаков после запятой или округление до конкретного разряда;

в) единый порядок действий при невозможности выполнения разметки, включая способ заполнения ячейки (например, «n/a» или иное обозначение);

г) уровень разметки, определяющий необходимость внесения данных о серии, номере среза и т.д., где была измерена/определена конкретная величина;

д) при необходимости сохранения врачом снимков экрана — объект, масштаб, порядок и место его сохранения;

е) при наличии цветовой кодировки ячеек — какие данные и в каком формате будут выделены;

– Список литературы: использовавшиеся при составлении инструкции источники в формате пронумерованного списка.

3.2.4 Алгоритм формирования

На данном этапе происходит непосредственно процесс сбора данных, их разметки, структуризации, анонимизации, формирования файлов данных и подготовка сопроводительной документации.

В рамках данной диссертационной работы был разработан алгоритм формирования НД (рисунок 12), позволяющий не только систематизировать и учесть все основные аспекты создания НД, но и способствующий созданию ПО для автоматизации этих процессов. Согласно Национальной стратегии разработка методологий описания, сбора и разметки данных и механизмов контроля за их соблюдением является основным направлением повышения доступности и качества данных для технологий ИИ [2]. Такая методология, реализованная в программных модулях, позволяет создавать путем их объединения единый программно-аппаратный комплекс, что, в свою очередь, позволит осуществить централизованную полуавтоматическую подготовку НД и оптимизирует процессы сбора, аннотации и обработки данных. Это сэкономит время и ресурсы

медицинских исследователей и врачей, обеспечит развитие ТИИ и в итоге будет способствовать повышению качества лучевой диагностики.

Разработанный алгоритм основывается на получении данных из медицинских информационных систем (МИС) МО, так как благодаря цифровизации здравоохранения большинство медицинских данных, полученных от пациентов, хранится именно в них, однако он может быть адаптирован и для работы с базами данных и другими источниками. Как отмечалось ранее, алгоритм разрабатывался на ЕРИС ЕМИАС как источнике медицинских данных. Важным аспектом при работе с медицинскими данными является обеспечение информационной безопасности и защита персональных данных. Поэтому в данной работе использовался модуль анонимизации, позволяющий удалять персональные данные как из данных, полученных из ЕРИС ЕМИАС, так и из DICOM-тегов самих диагностических исследований. Ниже представлен алгоритм формирования НД в подробном виде (рисунок 12).

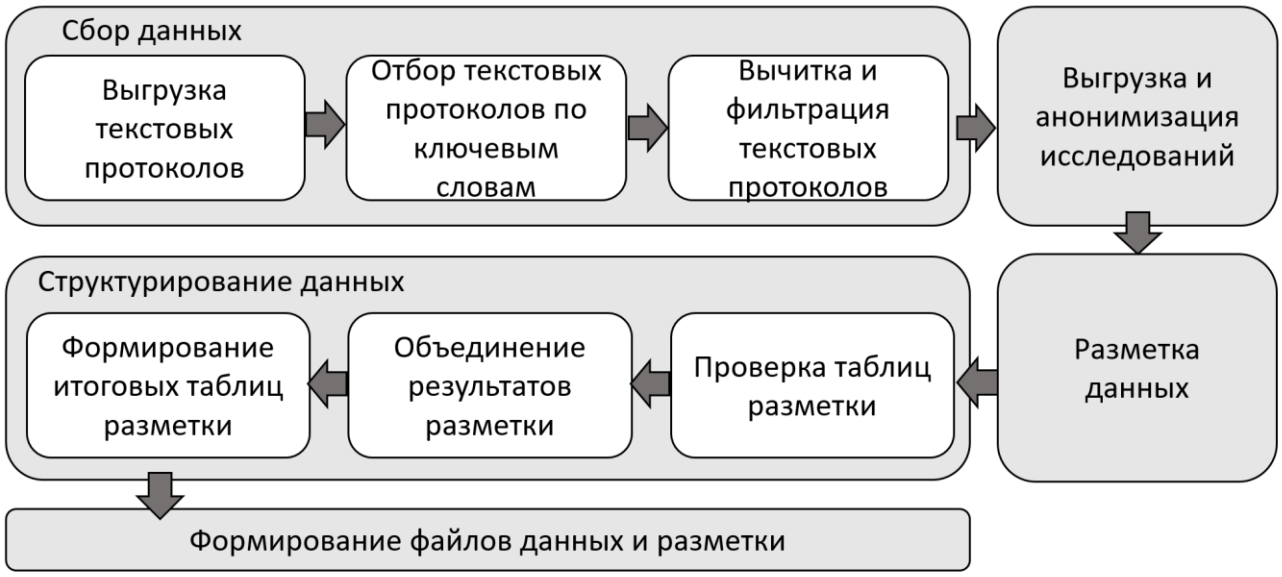


Рисунок 12 – Алгоритм формирования набора данных

Алгоритм состоит из следующих этапов:

1. Сбор данных:

– Выгрузка текстовых протоколов исследований из МИС. С помощью специальных программных средств, согласно требованиям, описанным в ТЗ (критерии включения, невключения), происходит фильтрация исследований по заданным параметрам (модальность, вид услуги, возраст, пол, МО, период проведения исследования и т. д.) и выгрузка текстовых протоколов исследований, полученных от врачей-рентгенологов и хранящихся в МИС.

– Отбор текстовых протоколов по ключевым словам. Анализ текстовых протоколов на предмет возможного соответствия требованиям ТЗ можно производить вручную с помощью врача-рентгенолога, однако гораздо быстрее и проще использовать методы автоматизированного анализа текстов, разделив процесс на 2 шага.

а) Первый: фильтрация по так называемым ключевым словам и стоп-словам с помощью специальных алгоритмов обработки естественного языка. Ключевые слова, в контексте анализа текстовых протоколов, – это слова и/или словосочетания, указывающие на наличие той или иной патологии или признака. Стоп-слова указывают на отсутствие целевой патологии или признака.

б) Второй: вычитка и фильтрация экспертами текстовых протоколов на соответствие исследуемой патологии с целью формирования выборки: проводится врачом-специалистом по данному направлению. Благодаря предварительному этапу автоматической фильтрации текстовых протоколов объем работы для такого специалиста существенно снижается по сравнению с отбором исследований врачом-рентгенологом вручную в ходе рутинной практической деятельности (пропорционально частоте встречаемости целевой патологии в выборке).

2. Выгрузка и анонимизация отобранных исследований: проводится специалистом в защищенном контуре. С помощью специального ПО происходит удаление персональных данных из DICOM-тегов, с изображений, а также замена уникальных идентификаторов.

Результатом этапов сбора и выгрузки и анонимизации данных является список отобранных исследований и соответствующие ему DICOM-файлы.

3. Разметка данных.

Это самый сложный, наименее автоматизированный, операторозависимый этап создания НД, который во многом определяется выбранным методом верификации данных (методы верификации представлены в главе 5). Наименее ценной и самой простой будет являться разметка методами анализа текстовых протоколам, т. е. то, что происходило на предыдущих этапах. Согласно классификации по типам разметки данных [55] (рисунок 13), ее ценность возрастает в следующей последовательности: определение факта наличия или отсутствия патологии, ее локализация, ее сегментация. Локализация – обозначение области интереса простой геометрической фигурой. Сегментация – обозначение области интереса путем попиксельного обведения ее границ (маска). Кроме того, классификация находки (например, по общепринятым шкалам – RADS [109], ASPECTS [110]), а также дополнительные данные медицинской карты (результаты лабораторных и инструментальных исследований, анамнез, данные осмотра, поставленный клинический диагноз) повышают ценность разметки.

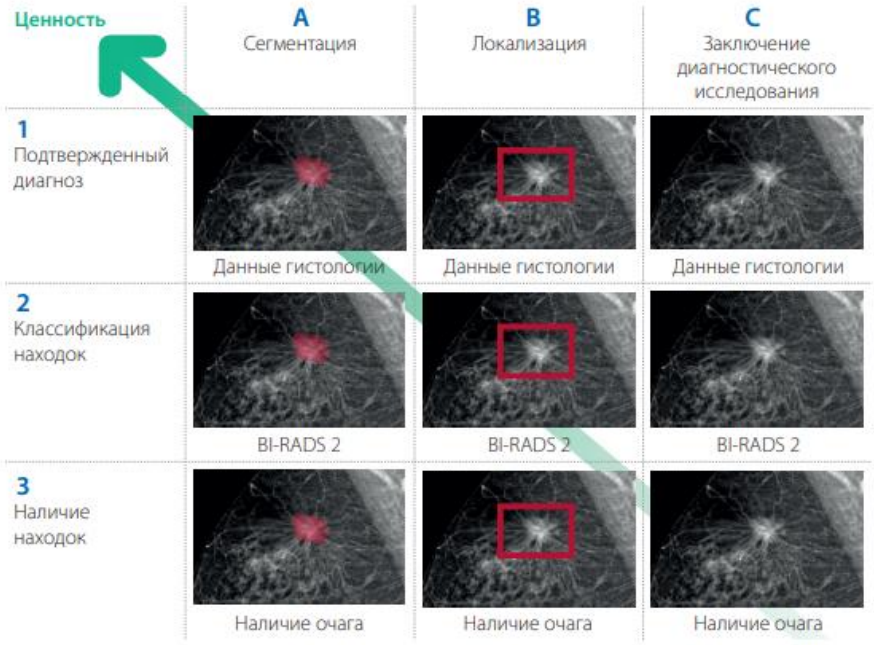


Рисунок 13 – Классификация наборов данных по типам разметки

Результаты разметки (т. е. измерения, классификация и определение наличия целевых признаков или патологий) заносятся в специальные таблицы или формы для последующего анализа. Сложность заполнения таких таблиц напрямую зависит от количества параметров разметки (целевых признаков, измерений, лейблов) и их типа (бинарные, мультиклассовые, регрессионные, текстовые). При увеличении количества параметров возрастает количество ошибок заполнения таблиц.

Результатом этапа разметки являются таблицы «с разметкой», т. е. заполненные разметчиками по каждому заявленному признаку параметры, и DICOM-файлы с дополнительной серией, на которой представлены результаты сегментации или локализации целевых признаков (маски), в случае если они проводились.

4. Структурирование данных:

– Проверка таблицы разметки специалистом по работе с данными. Все таблицы, полученные на этапе разметки, должны быть проверены на заполняемость, отсутствие дубликатов, достаточность и соответствие ТЗ. В случае нахождения ошибок и пропусков они возвращаются на предыдущий этап.

– Объединение результатов разметки. В случае выбора стратегии разметки данных, где участвуют 2 и более разметчиков, данные необходимо объединить в соответствии с выбранной стратегией.

– Формирование итоговых таблиц разметки. Итоговые таблицы формируются в соответствии с требованиями ТЗ (количество исследований, итоговые параметры разметки). Также желательно наличие минимальной сопроводительной информации в самом файле (например, в виде дополнительной вкладки, поля, строки) и/или зашифрованной в его названии (принципы формирования названий освещены в главе 5).

Результатом данного этапа являются структурированные таблицы с разметкой и/или дополнительная серия с масками в формате DICOM (NIfTI и др.) в случае, если проводились сегментация и локализация.

5. Формирование файлов данных и разметки.

Результатом данного этапа является репозиторий со структурированными файлами таблиц с разметкой, исследования в формате DICOM и дополнительная серия с масками в формате DICOM в случае, если проводились сегментация и локализация.

3.2.5 Регистрация и публикация

После завершения процессов сборки НД необходимо подготовить readme-файл и зарегистрировать НД – внести в реестр (раздел 3.3) сведения о его готовности и всю доступную информацию о нем, включая ссылку на место хранения (метаданные).

Готовность набора данных определяется ответственным за набор данных и заказчиком в соответствии с сформированным на этапе планирования техническим заданием. Сверяется соответствие модальности, количества исследований, лейблов, соотношения классов, критериев включения, невключения, исключения и других параметров. В данном процессе важную роль играет реестр наборов данных, который может выполнять роль инструмента контроля качества при приеме (публикации) НД: при внесении или проверке реальных данных в реестре фиксируется соответствие или несоответствие техническому заданию. В случае необходимости внесения изменений на более поздних этапах, эта информация также указывается в реестре, что обеспечивает прозрачность процесса создания набора данных.

Создание сопроводительного текстового файла (readme-файла) для дальнейшего использования и передачи на публикацию – важный аспект работы с НД. Readme-файл содержит краткую информацию о созданном НД: описание в свободной форме, модальность, целевую патологию/признак(и), методы верификации, период сбора, популяционные параметры, авторский состав, аффилиацию, информацию о цитировании, информацию о регистрации и

распространении и т.д. На рисунке 14 представлены фрагменты readme-файла, разработанного в ГБУЗ «НПКЦ ДиТ ДЗМ».

Readme-файл рекомендуется хранить в форматах pdf и md. Формат pdf является стандартным для хранения документации. Тем не менее конечные пользователи НД – это зачастую разработчики, а формат хранения md легко читаем на всех платформах независимо от ПО или операционных систем. Формат md упрощает обмен информацией о проектах в среде, где производится разработка программных продуктов, обеспечивает простоту редактирования, переноса и парсинга информации. Кроме того, при создании большого количества проектов (в том числе НД) он позволяет создавать программные модули для быстрого формирования readme-файла, что способствует автоматизации процессов создания НД. Документация в формате md легко интегрируется с системами управления версиями, что упрощает отслеживание изменений и совместную работу над документацией. Также рекомендуется формировать readme-файл на двух языках (русском и английском), что существенно расширяет географию использования НД. Таким образом, хранение в двух форматах на двух языках покрывает всю целевую аудиторию: зарубежных и отечественных исследователей и разработчиков.

Readme-файл необходимо хранить вместе с набором данных, как в защищенных локальных хранилищах, так и в открытых библиотеках наборов данных, обеспечивая обмен информацией с другими исследователями и разработчиками в соответствии с принципами преемственности и повторного использования.

MosMedData: РГ ОГК с наличием и отсутствием легочных узлов тип VII

> Набор данных содержит результаты рентгенографии органов грудной клетки с признаками легочных узлов, а также без признаков (норма). Данные исследования были собраны в отделениях лучевой диагностики лечебных учреждений города Москвы в период с 16.11.2017 по 06.04.2022. Двумя врачами-рентгенологами был проведен просмотр исследований на предмет наличия легочных узлов. (согласно глоссарию общества Фляйшнер (Fleischner) «легочный узел» определяется как образование более 6 мм, но менее 30 мм, расположенными в паренхиме легких). Для балансировки набора данных были добавлены исследования РГ ОГК без патологии. Дополнительно производился полный пересмотр финального набора данных врачом-экспертом (рентгенологом со стажем работы более 10 лет) и верификация исследований по результатам комьютерной томографии, выполненной не позднее 14 дней после рентгенографии. В итоговом наборе данных содержится 100 исследований РГ ОГК, из них 50 с патологическими изменениями (на КТ солидный или субсолидный узел размером более 6 мм) и 50 исследований РГ ОГК без патологии (отсутствие узлов на КТ).

а

Классы разметки

C-1

Методы верификации

Пересмотр специалистом, исследование другой модальности (компьютерная томография)

Ключевые слова

MosMedData, рентгенография, радиология,РГ, грудная клетка, легочные узлы, злокачественные новообразования легких, рак

Язык

Английский, русский

Версия набора данных

1.0.0

Постоянная ссылка

https://mosmed.ai/datasets/

б

Обзор данных

Параметр	Значение
Количество исследований, ед.	100
Количество пациентов, чел.	100
Распределение по полу, чел. (М/ Ж)	47/ 51
Распределение по возрасту, лет (мин./ медиана/ макс.)	18/ 58/ 87
Распределение по классам, ед. (С патологией/ Без патологии)	50/ 50

в

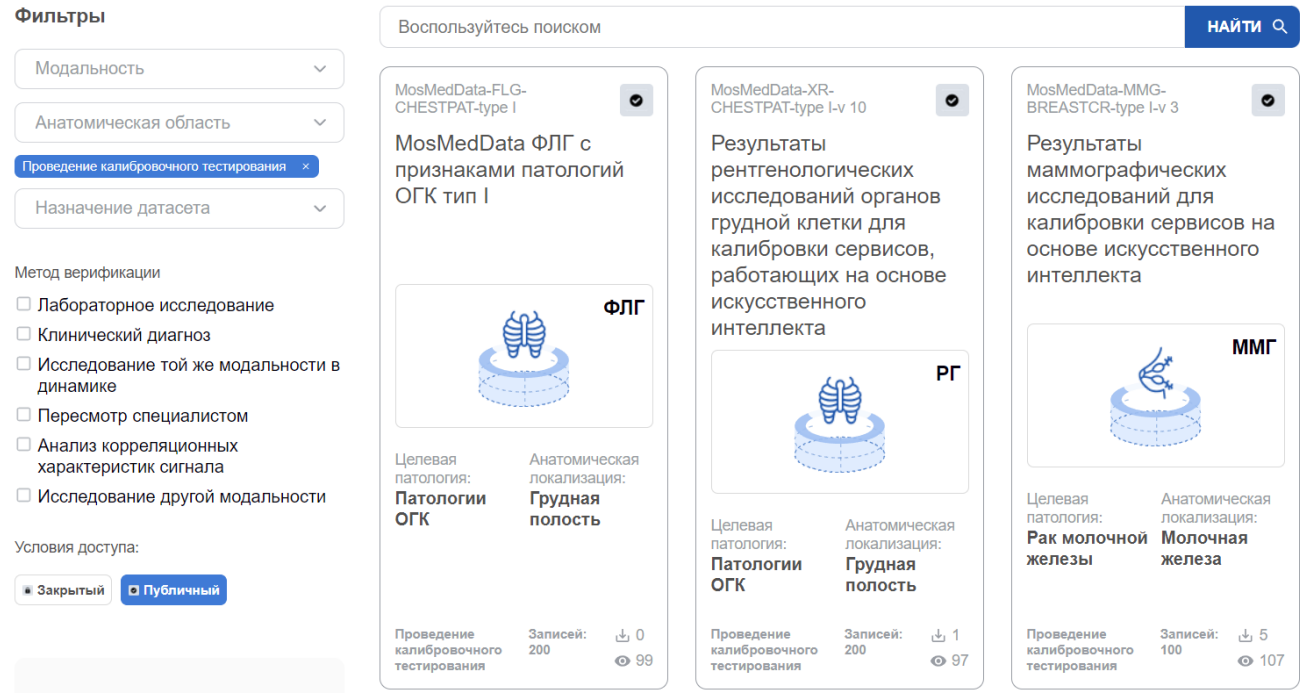
Рисунок 14 – Фрагменты readme-файла: а – краткое описание, б – информация о наборе данных, в – популяционные параметры

Публикация НД осуществляется в зависимости от типа доступа либо в локальном защищенном хранилище, либо в открытых библиотеках НД. По типам доступа НД можно разделить на закрытые, открытые и закрытые с общедоступными примерами.

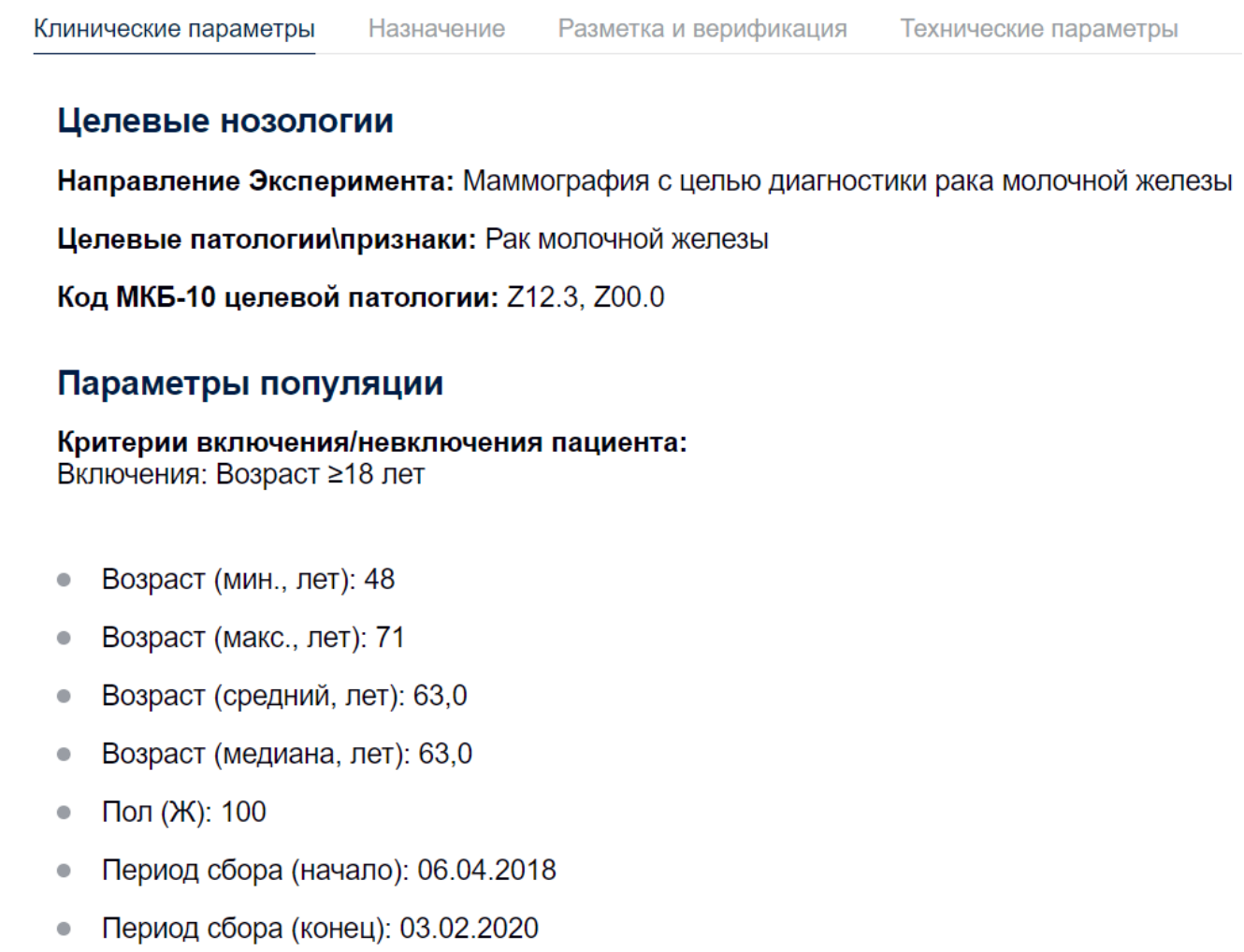
Даже закрытые НД соответствуют принципам разумной бережливости Национальной стратегии. Они могут использоваться повторно (целиком или частично) для решения других задач в рамках одного учреждения. Тем не менее огромную роль в развитии ТИИ играют опубликованные в открытом доступе НД. Библиотека НД – систематизированное собрание НД, доступных для использования. Они способствуют реализации принципа поддержки конкуренции в области ТИИ [2] за счет обеспечения компаний по разработке СИИ наборами данных с целью обучения и тестирования их программных продуктов. Библиотеки представляют собой каталог карточек, в которых представлена систематизированная информация об опубликованных НД (рисунок 15) [111].

Такое наглядное отображение информации о НД, удобство ее представления, возможность фильтрации по заданным параметрам стали возможными благодаря ее четкой структуризации и классификации. Это позволяет разработчикам и исследователям изучать предоставленные данные и оперативно принимать решение об их применимости для выполнения конкретных задач.

Важным аспектом создания библиотек является соответствие единым принципам организации данных. Этому может способствовать создание реестров НД и синхронизация с ними: полуручной или полностью автоматизированный перенос данных из реестра в библиотеки также способствует ускорению процессов публикации и снижению количества ошибок представления метаинформации.



а



б

Рисунок 15 – Библиотека наборов данных mosmed.ai/datasets: а – каталог библиотеки, б – фрагмент структурированной карточки библиотеки НД

Использование

В зависимости от поставленной цели НД может использоваться для обучения, дообучения и тестирования СИИ, для научных исследований, клинических и технических испытаний. Также НД может быть опубликован в открытых источниках и зарегистрирован в качестве результата интеллектуальной деятельности (РИД). Кроме того, на этапе использования необходимо четко регламентировать место хранения, формат и доступ. Например, для НД, предназначенных для тестирования, актуально наличие файлов в двух вариантах: с разметкой (для проверки результатов тестирования) и без разметки (для загрузки в СИИ). Кроме того, вместе с файлами данных необходимо хранить readme-файл и ТЗ.

Смена версии

Иногда в НД приходится вносить изменения уже после его регистрации и даже использования. Например, могут закрасться ошибки при создании, которые не увидели ранее, исследования могут быть утеряны, могут быть изменены БДТ и т.д. Для этого ранее была разработана система смены версионности [95], которая успешно применялась в Эксперименте и оказалась очень удобной. Она позволила сделать процесс создания НД для тестирования СИИ максимально прозрачным и удобным. Если происходит смена версии, то жизненный цикл набора данных начинается заново.

Удаление и архивация

Безвозвратно удалять данные нежелательно, так как из вышесказанного становится понятно, что создание НД – это сложный и дорогостоящий процесс. Данные можно использовать повторно и для других задач. Тем не менее возможны ситуации, когда их приходится удалять (например, это оговорено юридически). Во всех иных случаях НД по возможности следует сохранять в архиве.

3.3 Методы стандартизации и систематизации. Реестр как инструмент управления и контроля качества

Как отмечалось выше, медицинские данные имеют сложную, разнообразную и неоднозначную структуру. Методы сбора и способы организации данных также варьируют в зависимости от самих данных и задач, для которых они требуются. Вследствие этого появляется множество разнообразных НД, что усложняет их поиск и применение. Трудности возникают, например, при поиске большого количества данных для задач обучения. Еще один пример, который возник при проведении Эксперимента – необходимость в дополнительном НД с той же спецификацией, но с другим набором исследований (вариант) для целей повторного тестирования СИИ. Кроме того, очевидно, что разработка и использование СИИ напрямую зависит от качества НД, на которых она обучается и тестируется.

Для решения задач организации процессов управления и контроля качества при работе с НД требуется систематизация и структуризация информации о нем. Такая задача была реализована в виде реестра наборов данных. Это систематизированный перечень сведений обо всех НД ГБУЗ «НПКЦ ДиТ ДЗМ», ведущийся уполномоченным сотрудником, с целью упорядочивания деятельности по формированию и использованию НД [112]. На начало 2025 г. реестр состоял из 103 полей и 690 записей, включая смену версии. Все поля сгруппированы по разделам, соответствующим жизненному циклу НД. В приложении А представлено их краткое описание [112].

Принципы формирования названий и идентификаторов НД

Очень важным аспектом в процессе использования НД является его наименование и идентификация. В предыдущей главе указано, что в реестре имеется 4 названия. Такое количество обусловлено, во-первых, особенностями Эксперимента, во вторых – публикацией НД в открытом доступе. НД, предназначенные для тестирования СИИ в рамках Эксперимента, имеют вариативность и строгую схему использования: сначала проводится первичное

тестирование на первом варианте НД, в случае необходимости повторного использования – вторичное на втором варианте НД, в некоторых случаях – третичное. Чтобы не перегружать идентификаторы публичных НД дополнительной информацией, а также отражать в нем наиболее актуальные параметры (название проекта актуальнее для внутреннего использования, а наименование учреждения – для публичного), было разработано 2 версии идентификаторов: публичная и внутренняя. Ниже представлена структура всех видов наименований и идентификаторов:

1. Предварительное название. Формируется в свободной форме на самом раннем этапе жизненного цикла – инициализации. Основная цель – внести инициированный НД в реестр и обозначить начало работы над проектом.

2. Внутренний идентификатор. Предназначен для внутреннего использования, необходим для однозначной идентификации НД (уникален). Сформирован из ключевых параметров НД, что позволяет при его небольшой длине получить основную информацию о его содержании: принадлежность проекту, год создания, цель создания, целевая модальность, нозология, анатомическая область, вариант и версия. Кроме того, так как в него включены параметры, отображенные в реестре, он может формироваться в автоматическом режиме. Структура представлена на рисунке 16. Для НД, предназначенных для тестирования, предусматривается два файла: с разметкой (для оценки работы СИИ) и без разметки (для направления компании – разработчику СИИ), что также отражается в названии этих файлов путем добавления «full_markup» в название файла с разметкой.

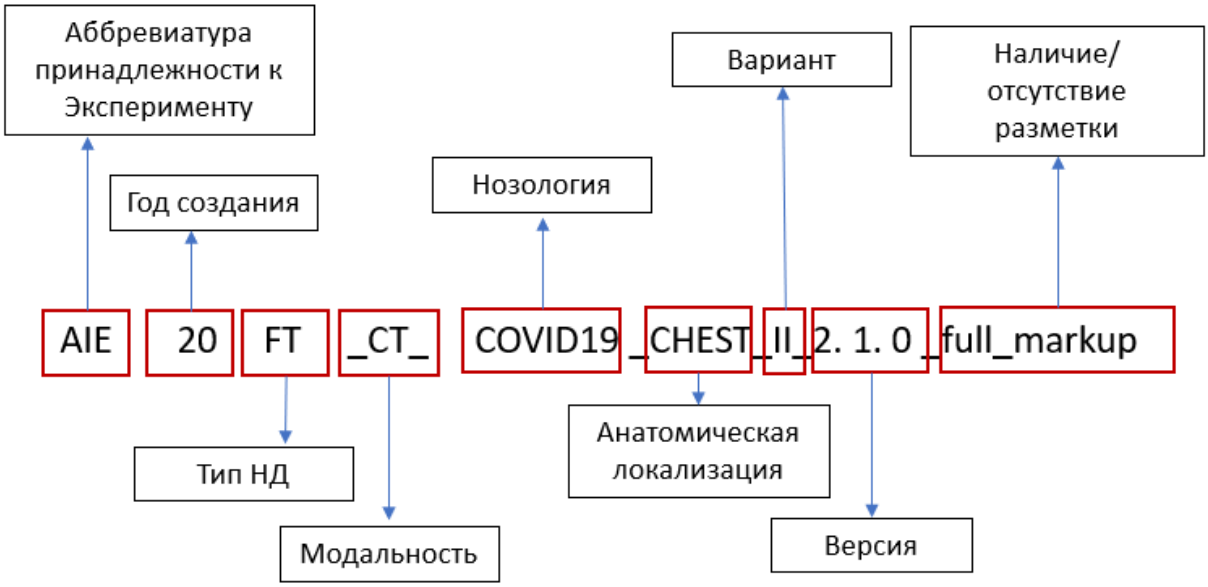


Рисунок 16 – Структура внутреннего идентификатора НД

Для удобства восприятия и исключения путаницы между версиями и типом НД разработана таблица соответствия обозначения цели создания НД (таблица 2).

Таблица 2 – Типы НД по их назначению, сокращенные обозначения

Тип	Идентификатор	Назначение НД	
I	FT/CT	Functional testing / calibration testing	Проведение функционального / калибровочного тестирования
II	TST	Technical Self-test	Проведение технического самотестирования
III	DST	Diagnostic Self-test	Проведение диагностического самотестирования
IV	ClT	Clinical Test	Выполнение клинических испытаний
V	TT	Technical Test	Выполнение технических испытаний
VI	NLP	Natural Language Processing	Проведение разметки текстовых протоколов с помощью программ автоматизированного анализа текстов
VII	SS	Scientific Study	Проведение научных исследований
VIII	AID	Artificial Intelligence Design	Разработка ИИ
IX	ST	Specialists Training	Для обучения специалистов
X	GOST	ГОСТ (государственный стандарт)	Для национальных стандартов

3. Публичный идентификатор. Предназначен для идентификации НД при публикации в открытом доступе (уникален). Содержит в своей структуре основные параметры, представляющие интерес внешнему пользователю: организация сбора данных, целевая модальность, патология, анатомическая область, целевое применение. Под порядковым номером версии НД понимается порядковый номер НД, для которого вся остальная часть названия идентична. Необходим исключительно для однозначной идентификации. Структура представлена на рисунке 17.

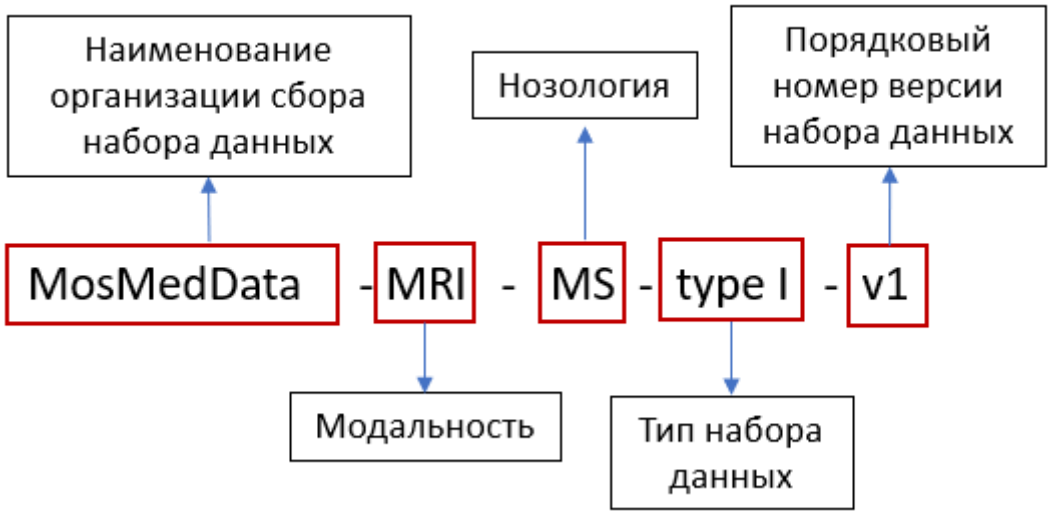


Рисунок 17 – Структура публичного идентификатора НД

4. Публичное название. Не является уникальным. Формируется на русском языке, отражая в структуре те же данные, что и публичный идентификатор, только в полном, связанном виде. Используется для публикации НД в открытых источниках (библиотеках, публикациях, РИД и т. д.), а также при формировании readme-файла (в т. ч. при его переводе на английский язык). Неуникальность обусловлена тем, что одно полное публичное название не содержит никаких данных о версиях и вариантах и может включать в себя несколько наборов данных. Например, 2 варианта НД для калибровочного тестирования и 2 варианта для функционального будут объединены под одним полным названием с целью регистрации РИД. Структура представлена на рисунке 18.

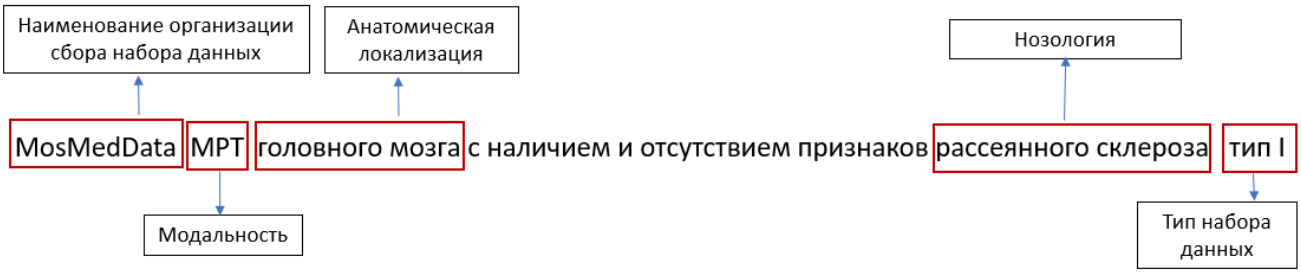


Рисунок 18 – Структура публичного названия НД

На первых этапах создания НД была неочевидна необходимость большого количества полей (в реестре) и сложной структуры наименований НД. Однако в дальнейшем, с ростом количества НД потребовалась и была разработана такая структура. Для сравнения: в ходе ревизии библиотеки НД и хранилищ в 2024 г. для НД, не имевших структурированного названия, понадобилась выгрузка, разархивация и идентификация путем изучения readme-файла или таблиц с разметкой, что потребовало больших затрат ресурсов: времени, сотрудников, а также места на рабочих персональных компьютерах, особенно при выгрузке объемных файлов. Тем не менее благодаря наличию реестра удалось идентифицировать НД, восстановить корректные данные, при необходимости поправить карточки библиотеки.

Кроме того, внутренний идентификатор является частью автоматизации процесса тестирования СИИ: все НД типа I хранятся в локальном хранилище в виде таблиц с разметкой (full_markup) и без нее, и при необходимости проведения тестирования производится запрос НД (обращение к ответственному за выдачу) с указанием ИИ-сервиса, направления (модальность, нозология и анатомическая локализация) и вида тестирования (функциональное/калибровочное, первичное или повторное). Все запрашиваемые параметры (за исключением названия ИИ-сервиса) закодированы в названии НД, что позволяет сотруднику, ответственному за выдачу, быстро найти требуемый НД. Кроме того, благодаря внесенной в название кодировке версии и хранению всех предыдущих версий возможно

отследить, на каком НД тестировался тот или иной сервис, что часто необходимо не только в рамках решения задач Эксперимента, но и научных.

Такая, на первый взгляд, сложная система названий оправдана в организации всех основных процессов работы с НД с учетом количества и разнообразия задач, решаемых с их использованием в ГБУЗ «НПКЦ ДиТ ДЗМ».

Она обеспечивает:

1. удобную идентификацию и обращение к НД для всех сотрудников (в том числе разных структурных подразделений и разных специальностей), задействованных в рамках Эксперимента (на начало 2025 г. отмечено регулярное использование более чем 280 НД), а также для других научных и практических задач в области инструментальной диагностики в ГБУЗ «НПКЦ ДиТ ДЗМ»;
2. единые принципы формирования названий для регистрации НД в качестве РИД, формирования readme-файла, упоминания в научных публикациях;
3. удобную идентификацию и представление НД в каталоге библиотеки mosmed.ai;
4. удобную систему хранения НД в архиве.

Функции реестра наборов данных

Большое количество полей реестра, охватывающих самые разные аспекты, начиная с клинических и технических параметров НД и заканчивая постановкой цели, распределением ресурсов и планированием проекта, обусловлены широким спектром задач, которые решаются с использованием реестра. Они могут стоять перед специалистами разных областей и структурных подразделений. Например, руководителей проекта, структурных подразделений, участвующих в процессах создания НД, будут в большей степени интересовать вопросы качества выполнения работ, соблюдения сроков, обозначенных при планировании, распределения ресурсов, результативности и эффективности использования. Врачей и исследователей – клинические параметры, инженеров и разработчиков – технические и т. д. Исходя из этого, были определены основные функции реестра НД:

1. Контроль качества данных.

Согласно европейским рекомендациям по ведению медицинских реестров [68] среди основных факторов, влияющих на качество, выделяют: управление, качество данных, безопасность. Контроль качества тесно связан с процессами управления, тем не менее в рамках нашей работы они являются разными функциями реестра. Под контролем качества данных понимается:

2. Проверка параметров НД на соответствие ТЗ, БДТ (раздел «Карточка НД») при внесении информации. Кроме того, реестр способствует автоматизации процессов создания и использования НД, а также позволяет осуществить автоматизированный контроль ввода данных, что уменьшает вероятность ошибки оператора. Так, был реализован модуль внесения популяционных параметров с проверкой возраста согласно ТЗ, что позволило не только оперативно выгружать параметры НД, но и контролировать такой параметр невключения, как возраст пациента, а также проводить проверку на наличие дубликатов.

3. Прозрачность, надежность и воспроизводимость разработок (разделы «Карточка НД», «Смена версии»). Ведение реестра позволяет подробно и структурированно зафиксировать ключевую информацию о НД, что в случае возникновения вопросов или аналогичных задач обеспечит удобный доступ к данным без необходимости поиска ТЗ и участников создания НД. Если такой информации будет недостаточно, в реестре также предусмотрены ссылки на более детальные документы (БДТ, ТЗ, файлы разметки, файлы изображений), а также Ф. И. О. специалистов.

Следует отметить, что функцию реестра по контролю качества данных необходимо разделять с качеством самого реестра (несмотря на то, что первое напрямую следует из второго). Качество реестра определяется результативностью его использования, а также точностью и полнотой его заполнения [68, 112].

4. Управление.

Управление – организационная основа всех процессов жизненного цикла НД, обеспечение ресурсов (финансовых, людских и технических) для их

функционирования [68]. Управленческие задачи чрезвычайно важны как в процессе создания одного конкретного НД, так и при ведении проектов, включающих в себя большое количество НД, причем с их увеличением значимость верно выстроенных процессов управления резко возрастает. Реестр позволяет не только организовывать деятельность по созданию и использованию НД, но и способствует пониманию стратегии более глобальных задач и проектов, для которых эти НД создавались. Например, анализ результативности использования НД по целевой патологии позволяет определить тенденции к изучению конкретной патологии или, напротив, выявить направления, ранее не изучавшиеся, но представляющие интерес. При внедрении реестра НД в деятельность ГБУЗ «НПКЦ ДиТ ДЗМ» удалось оптимизировать процессы создания НД за счет реализации следующих управленческих задач:

1. контроль сроков и порядка выполнения работ по созданию НД (раздел «Планирование»);
2. распределение ресурсов при создании НД (раздел «Планирование»);
3. оценка результативности использования НД (выписки и справки), в том числе с целью предоставления отчетности по проектам (все разделы, преимущественно «Использование»);
4. оптимизация ресурсов, в т. ч. возможность повторного использования данных (все разделы);
5. доступ к данным.

Реестр НД способствует решению задач по централизации и упорядочиванию хранения НД и информации о них за счет следующих параметров:

1. ссылки на хранение (поля «Хранение в архиве», «Ссылки на хранение файлов с разметкой и без», на «Ссылка на БДТ»);
2. ответственные лица (поля «Заказчик», «Ответственный за БДТ», «Ответственный за НД» , «Разметчики» , «Авторы»);
3. информация о НД (все разделы);
4. оперативное формирование библиотеки НД (раздел «Карточка НД»).

Централизованный доступ к информации о НД позволяет решать самые разные задачи, и одной из них является формирование наполнения библиотек НД. Карточки библиотеки синхронизированы с полями реестра, что позволяет быстро выгружать структурированную информацию о НД, а это, в свою очередь, оптимизирует процесс публикации, минимизирует количество ошибок, а также делает возможным поиск по настраиваемым параметрам (модальность, метод верификации, тип НД). Публикация в таком наглядном виде позволяет потенциальным пользователям изучать информацию о НД и оперативно принимать решение о возможности их применения для своих задач. В целом это способствует развитию информационно-коммуникационной инфраструктуры для обеспечения доступа к данным [2].

Одним из ключевых параметров при создании и использовании НД является метод верификации входящих в него данных, поэтому необходимо было разработать классификацию, максимально обобщающую методы верификации, что, в свою очередь, позволит настроить удобный поиск НД по заданным параметрам. В рамках данной диссертационной работы была разработана такая классификация (таблица 3).

Таблица 3 – Методы верификации данных

Метод верификации	Пример
Исследование другой модальности	Для верификации патологии на рентгенографическом исследовании: компьютерная томография той же области
Лабораторное исследование	Гистологическая верификация злокачественных новообразований предстательной железы
Исследование той же модальности в динамике	Для верификации признаков кровоизлияния в головной мозг: признаки кровоизлияния в заключении компьютерной томографии в динамике
Клинический диагноз	Установленный диагноз U07.1 по данным медицинской карты
Пересмотр специалистом	Пересмотр разметчиком и экспертом
Согласно тексту описания исследования	Поиск ключевых слов в тексте описания исследования

5. Автоматизация.

Автоматизация процессов – один из ключевых способов реализации Национальной стратегии по разработке и развитию СИИ. Повышению доступности и качества данных способствует в том числе и автоматизация на всех этапах создания и использования НД и разработка единой платформы подготовки НД, что, в свою очередь, возможно реализовать при наличии четких методологий формирования НД, принципов их стандартизации и структуризации. Реестр является одним из инструментов автоматизации и позволяет решить следующие задачи:

1. генерация readme-файлов (на двух языках в двух форматах): на языке Python был реализован код, позволяющий в автоматическом режиме генерировать readme-файл, используя поля реестра;
2. формирование карточек библиотек НД;
3. проверка параметров на соответствие ТЗ;
4. автоматизация создания НД.

Необходимо отметить, что функции реестра между собой тесно взаимосвязаны. Например, качество данных напрямую зависит от процессов управления и автоматизации, централизованный доступ позволяет оптимизировать управленческие задачи, а формированию качественных библиотек НД способствует автоматизация заполнения карточек и создания readme-файла.

Можно выделить еще одну функцию реестра, которая лежит в основе всех остальных: получение информации. Она может быть востребована для совершенно разных задач. К примеру, необходимо описать НД для публикации исследования в научном журнале: наличие реестра позволяет не тратить ресурсы на поиски этой информации (тем более может возникнуть такая ситуация, что после анонимизации эта информация будет утеряна, например популяционные параметры). Еще один пример ситуации – это подготовка документации для оформления РИД: в реестре содержится практически вся информация (за исключением примеров наполнения

НД, которые находятся по ссылке в разделе «Использование»), необходимая для оформления документации.

Регламент ведения реестра

Для обеспечения заявленного функционала реестра, а также максимальной результативности его использования необходимо регламентировать процессы управления и контроля качества самого реестра. Применение надлежащих принципов управления должно обеспечивать четкий и легкий способ сбора данных на всех этапах [112]. Для этого были разработаны сопутствующие документы:

1. Правила создания, изменения и использования НД и их учет. Они опираются на описанные выше алгоритм создания НД и его жизненный цикл и регламентируют порядок взаимодействия персонала при создании, изменении и использовании НД.
2. Руководство по заполнению реестра НД. В нем содержится подробное описание всех полей реестра, ссылки, справочники, примеры и регламент предоставления данных.
3. Шаблон для заполнения реестра. Представляет собой таблицу, в которой содержатся все поля реестра, указана краткая инструкция по заполнению каждого поля, а также примеры заполнения. Шаблон в полном или сокращенном виде предоставляется ответственному за НД для внесения необходимой информации.

Представленная документация предназначена для обеспечения качества заполнения реестра, то есть своевременности, точности и полноты. Процессы управления реестром могут также базироваться на обратной связи, то есть частоте и целесообразности его использования. Целесообразность напрямую зависит от масштабов разработок: если ведется создание единичных НД, то нерационально тратить ресурсы на заполнение более чем 100 полей, хотя, в случае их регулярного использования, возможно ведение специальных журналов учета. Напротив, при большом количестве разработок, их разнообразии и потенциале новых задач оправдано не только ведение полноценного реестра, разработка модулей

автоматизации его заполнения, но и создание регламента так называемых процессов ETL (Extract, Transform, Load – извлечение, преобразование, загрузка) [115].

Еще один важный вопрос, который необходимо осветить, – это обеспечение безопасности информации не только в контексте самих НД, но и для реестра как базы данных, в том числе и возможной непубличной информации. Аспекты обеспечения безопасности персональных данных НД отражены в поле реестра «Анонимизация», а обеспечение безопасности содержимого реестра должно строго регламентироваться внутренними приказами на основании действующего законодательства [37] и быть прописано в сопроводительной документации.

3.4 Ошибки, возникающие при создании наборов данных, и методы их устранения

Алгоритм формирования НД, сформулированный жизненный цикл, стандартизация и классификация данных, а также опыт создания большого количества НД позволили выявить ряд систематических ошибок и предложить пути их устранения. Ниже представлены виды таких ошибок на каждом этапе жизненного цикла и алгоритма создания НД.

Инициирование и планирование

1. Самой критичной ошибкой являются нечетко сформулированные цель и задачи, что влияет на качество ТЗ и в дальнейшем на качество НД. В наихудшем случае это может привести к созданию НД, который невозможно применять для решения предполагаемых задач. С целью предотвращения этой ошибки в ТЗ необходимо указывать тип (согласно таблице 2), цель создания, назначение, область применения и/или целевую аудиторию использования НД.

2. Некачественное, формальное ТЗ не только является следствием предыдущего пункта, но может иметь место и при четко поставленных целях и задачах. Отсутствие ключевых параметров НД (например, лейблы, представленность классов, итоговый состав) или их неявный вид также могут

привести к критичным ошибкам на любом из следующих этапов и, в итоге, к невалидному НД. Для предотвращения таких ошибок необходимо утверждение формы ТЗ, в которой будут максимально подробно и четко обозначены все аспекты создания НД [55]. ТЗ должно быть понятно заказчику и ключевым участникам процесса создания НД.

3. Недостаточная проработка процесса планирования работ по созданию НД. Для понимания распределения ресурсов, обозначения корректных сроков выполнения работ и критериев качества необходимо верное представление всех процессов ключевыми участниками проекта. Для этого нужно соблюдать регламент подготовки НД, а также использовать инструменты управления, как общепринятые (например, диаграмма Ганта), так и специализированные (реестр НД). Кроме того, зачастую трудно распланировать новые процессы, если они ранее не выполнялись. Для этого необходимо изучение тематической научной литературы, опыта аналогичных работ, а также проводить пилотные проекты.

Для предотвращения ошибок на этапах инициирования и планирования следует привлекать мультидисциплинарную команду для правильной и корректной формулировки необходимых требований (БДТ, БФТ, ТЗ, инструкции, регламенты и иная сопроводительная информация) и рационального планирования всех процессов.

Формирование НД

1. Неизбирательность фильтрации текстовых протоколов, следствием чего может стать нехватка исследований с заданными параметрами в дальнейшем. Для предотвращения такого рода ошибок необходимо привлечение профильного специалиста для подбора ключевых слов (при разработке алгоритмов обработки естественного языка) и/или увеличение объема выгрузки данных (объем выгрузки необходимо рассчитывать исходя из данных встречаемости целевого признака в популяции).

2. Длительная выгрузка данных. Необходимо планирование с учетом загрузки каналов связи.

3. Выгруженные данные повреждены и/или не соответствуют критериям отбора. При планировании необходимо внести запас на возможный брак в данных.

4. Наличие внедренных в медицинское изображение персональных данных, которые невозможно удалить, дефектов, а также отсутствие изображения в МИС. Решением этой проблемы могут быть средства автоматизированного контроля персональных данных и дефектов [47, 48, 116, 117].

5. Наличие пропущенных, некорректных значений, дубликатов в таблицах разметки. Разметка данных – самый сложный и трудозатратный этап создания НД, и наибольшее количество ошибок возникает именно здесь. Например:

4. пустые значения (причина может быть неясна: пустое, потому что пропущено или потому что невозможно разметить);

5. текстовые символы вместо числовых (например, интервал или несколько размеров в одной ячейке);

6. неверная размерность (например, в процентах вместо абсолютных значений);

7. неверная точность (например, необходима точность в мм, а указана в см);

8. опечатки (отсутствие разделителя разряда, лишние/недостающие/ неверные цифры). Наиболее сложный для выявления тип ошибок;

9. присутствует разметка исследования, помеченного как брак;

10. дубликаты (одно и то же исследование размечено несколько раз).

Существует несколько причин появления ошибок при разметке, от которых зависят способы их устранения. Во-первых, как отмечалось выше, это ошибки в сопроводительной документации. Во-вторых, отсутствие понимания процессов создания НД и опыта разметки у врачей-рентгенологов. Зачастую врачи-разметчики, являясь выпускниками разных школ, имеют разное представление об описании рентгенологических исследований (включая различную терминологию). Свой опыт и знания по описанию исследований они

экстраполируют на процесс разметки несмотря на то, что она преследует совсем иные цели. Поэтому необходим тщательный инструктаж (в том числе сопроводительная документация), обучение, тестирование специалистов и обратная связь мультидисциплинарной команды непосредственно при создании НД. В-третьих, таблицы разметки в зависимости от задач могут быть очень большими, сложными и неудобными, что повышает вероятность возникновения случайных ошибок. И наконец, человеческий фактор, на который влияют опыт, загруженность, личные качества специалиста, условия труда, удобство оборудования и ПО.

Одним из универсальных способов решения вышеперечисленных проблем является автоматизация процессов разметки, например путем создания специальных платформ с удобным интерфейсом и всем необходимым для разметки инструментарием. Заполнение структурированной формы намного удобнее, чем таблиц разметки, исключает ряд случайных ошибок. Кроме того, на таких платформах можно создавать ограничения на формат/размер данных, вносимых в ячейку, что исключает ошибки некорректного внесения данных. Добавление информационных вкладок и сносок к полям облегчает понимание требований к данным. Ограничение на пустые значения, а также ввод чек-боксов позволяют избавиться от пропущенных значений и внести ясность в возникновение незаполненных полей. Такая платформа позволит получать данные в структурированном, единообразном, машиночитаемом виде, что не только уменьшит количество ошибок, но и будет способствовать дальнейшей автоматизации процесса формирования НД, что, в свою очередь, ускорит его и повысит качество конечного продукта.

Регистрация и заполнение реестра НД

1. Отсутствие в сопроводительной документации информации, необходимой для заполнения реестра. Для корректного функционирования реестра как инструмента управления необходимо обеспечить точность и полноту его заполнения. Для предотвращения такого рода ошибки необходимо создание

стандартизированных форм сопроводительной документации, а также модулей автоматизации извлечения информации из сопроводительной документации и файлов данных.

2. Несоответствие данных требованиям ТЗ. В этом случае реестр НД сам выступает в роли инструмента предотвращения ошибок, если это не произошло на более ранних этапах, однако процесс можно также оптимизировать путем создания автоматизированных модулей проверки данных.

3. Ошибки, отсутствие необходимой информации в readme-файле, внесение правок в НД после создания readme-файла. Для предотвращения таких ошибок необходимо тщательное заполнение реестра НД, соблюдение регламента формирования НД, а также создание модуля автоматической генерации readme-файла на основе данных из реестра. Такой модуль был разработан на базе подготовленного реестра и позволил существенно сократить временные затраты на этапе создания readme-файла. Так, данный процесс в ручном режиме (путем занесения информации из реестра в шаблон) занимал не менее часа. Кроме того, для конвертации формата md в pdf требовалась помощь сторонних ресурсов, что не всегда возможно из соображений информационной безопасности, поэтому первоначально был разработан модуль, куда вносилась необходимая информация в ручном режиме и формировались файлы в двух форматах, а в дальнейшем было реализовано автоматическое считывание из реестра. Данный процесс позволил сократить время подготовки readme-файла до 10 секунд.

Таким образом, соблюдение этапов алгоритма формирования НД, а также использование реестра в ходе создания НД в течение 2 лет позволили выявить наиболее типичные ошибки, а также предложить пути их предотвращения и устранения. Кроме того, внедрение методологии формирования НД позволило регламентировать все процессы, создавать больше НД, как следствие, проводить больше тестирований по большему количеству направлений, вовлекать меньше немедицинских специалистов в работу (количество врачей, вовлекаемых в процесс создания НД, зависит преимущественно от сложности разметки и количества

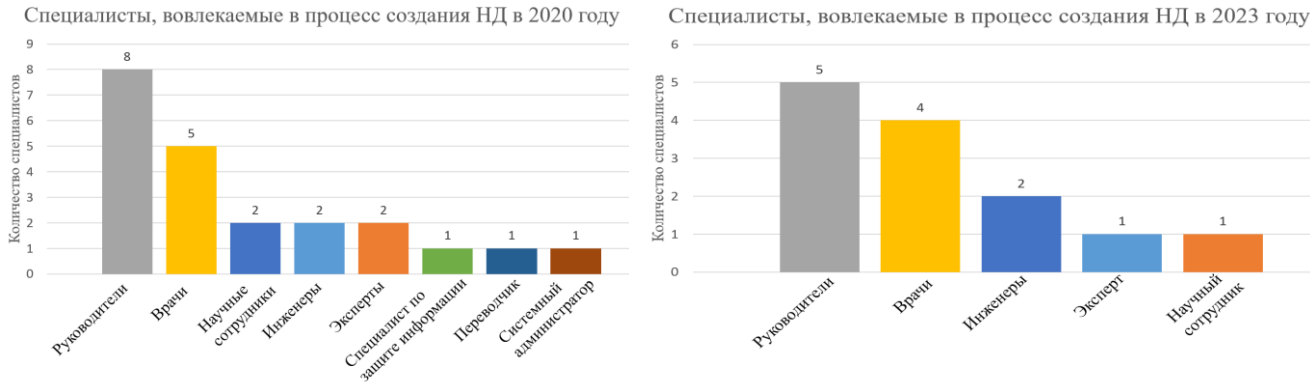
размечаемых исследований). В таблице 4 представлен ряд параметров, позволяющих оценить изменения в процессе создания и использования НД, которые были получены в результате внедрения разработанной методологии в практическую деятельность. Так, количество созданных за год НД увеличилось втрое, количество сервисов, участвующих в эксперименте – также втрое, при этом сложность структуры НД возросла (увеличение количества лейблов), а количество сотрудников, задействованных в процессе формирования НД, уменьшилось.

Таблица 4 – Изменения в процессах создания и использования НД в 2023 г. по сравнению с 2020 г.

Параметр	2020 год	2023 год
Количество созданных НД	58	186
Количество направлений Эксперимента (по каждому направлению созданы НД)	4	58
Количество сервисов в Эксперименте (прошли тестирование на НД)	18	57
Количество тестирований	65	204
Среднее количество лейблов НД	1,5	9
Среднее количество сотрудников, привлекаемых в процесс создания НД (за исключением врачей)	13	8
Время, затрачиваемое на создание readme-файла	от 1 часа	от 10 секунд
Количество смен версии в среднем на 1 НД	0,75	0,05
Максимальное количество смен версий	10	3

Безусловно, одним из показательных параметров является время, затраченное на создание НД, однако в условиях ГБУЗ «НПКЦ ДиТ ДЗМ» оценить его не представляется возможным ввиду задействования большого количества сотрудников и их многозадачности. На рисунке 19 показаны специалисты, вовлекаемые в процесс создания НД в среднем за 2020 и 2023 гг. На первых этапах требовалось большое количество специалистов разной направленности

(рисунок 19): руководители различных подразделений и дирекций, врачи (большинство врачей также являлось научными сотрудниками), научные сотрудники, инженеры, эксперты (как врачи, так и эксперты в области анализа данных, машинного обучения и т. д.), специалист по защите информации, переводчик, системный администратор. В 2023 г., после внедрения разработанной методологии количество и состав группы, участвующей в создании НД, изменился: сотрудников стало меньше, вовлечение сотрудников по трем специальностям (защита информации, переводчик, системный администратор) не требовалось. При этом сложность НД возросла: увеличилось количество лейблов, разнообразие представленной информации. Так, в 2020 г. НД преимущественно содержали 1–2 лейбла и бинарную разметку («норма»/«патология»), тогда как в 2023 г. количество лейблов достигало 74, и они содержали как бинарную разметку, так и категориальную и регрессионную.



а б

Рисунок 19 – Специалисты, вовлекаемые в процесс создания НД в ГБУЗ «НПКЦ ДиТ ДЗМ»: а – в 2020 году (до внедрения методологии формирования НД), б – в 2023 году (после внедрения методологии)

В целом структуризация информации и этапность процессов позволяют не только избежать ошибок, тщательно распланировать и регламентировать ход работы, но и осуществить автоматизацию этих процессов путем создания отдельных модулей и дальнейшего их объединения в платформу подготовки НД.

3.5 Обоснование минимального объема выборки и баланса классов набора данных для тестирования систем искусственного интеллекта в лучевой диагностике

Для того чтобы ответить на вопрос, каков должен быть минимальный размер набора НД и баланс классов при проведении тестирования СИИ с бинарным исходом, в рамках данной диссертационной работы было проведено экспериментальное исследование поведения AUC ROC в зависимости от объема выборки и баланса классов на реальных данных. AUC ROC был выбран в качестве целевого параметра оценки диагностической точности по причине его интегрального характера (учитывает и чувствительность, и специфичность) и независимости от настройки порога принятия решения. Кроме того, AUC ROC является наиболее популярным параметром оценки работы СИИ, оценивается во многих работах, в том числе в Эксперименте, что позволяет сравнивать и сопоставлять результаты исследований [65, 90].

В результате описанного в материалах и методах эксперимента с многократным повторением выборок различного баланса классов и размера и расчёта площади под характеристической кривой, мы получили зависимости средних значений AUC ROC от объемов выборки (рисунок 20) для каждого баланса класса. Отмечается снижение амплитуды разброса средних значений с увеличением объема выборки, при этом для сбалансированных выборок (50%) амплитуда разброса меньше, чем у несбалансированных (10%) и ее снижение имеет более резкий характер.

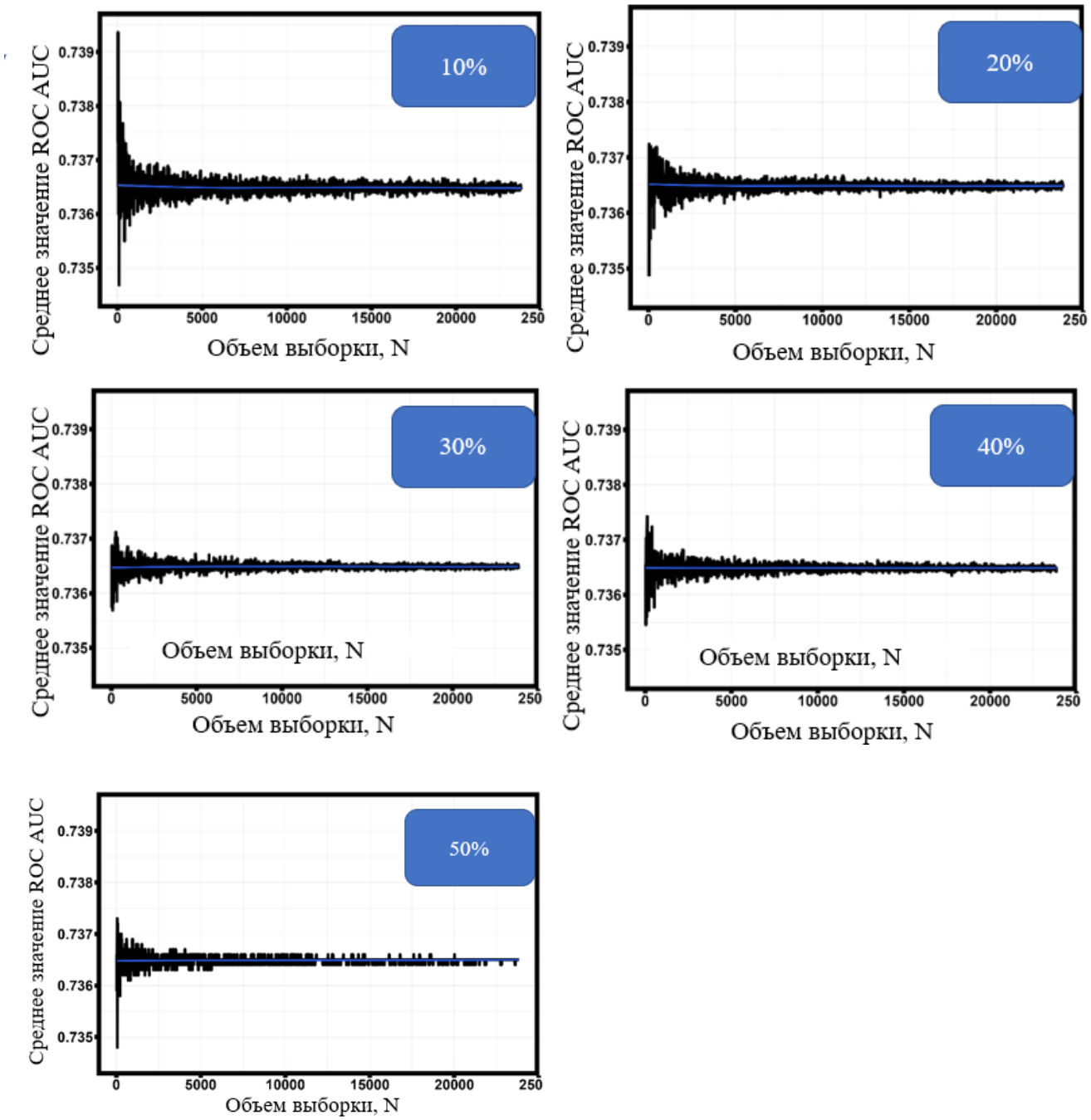


Рисунок 20 – Поведение средних значений AUC ROC для различных балансов классов «норма»/«патология» для НД1: синяя линия показывает аппроксимирующую кривую; процент исследований с «патологией» указан в правом верхнем углу каждого графика

При анализе полученных значений отмечается тенденция к снижению отклонения средних значений AUC ROC при балансировке классов (от доли патологии 10 % до 50 %) (рисунок 21).

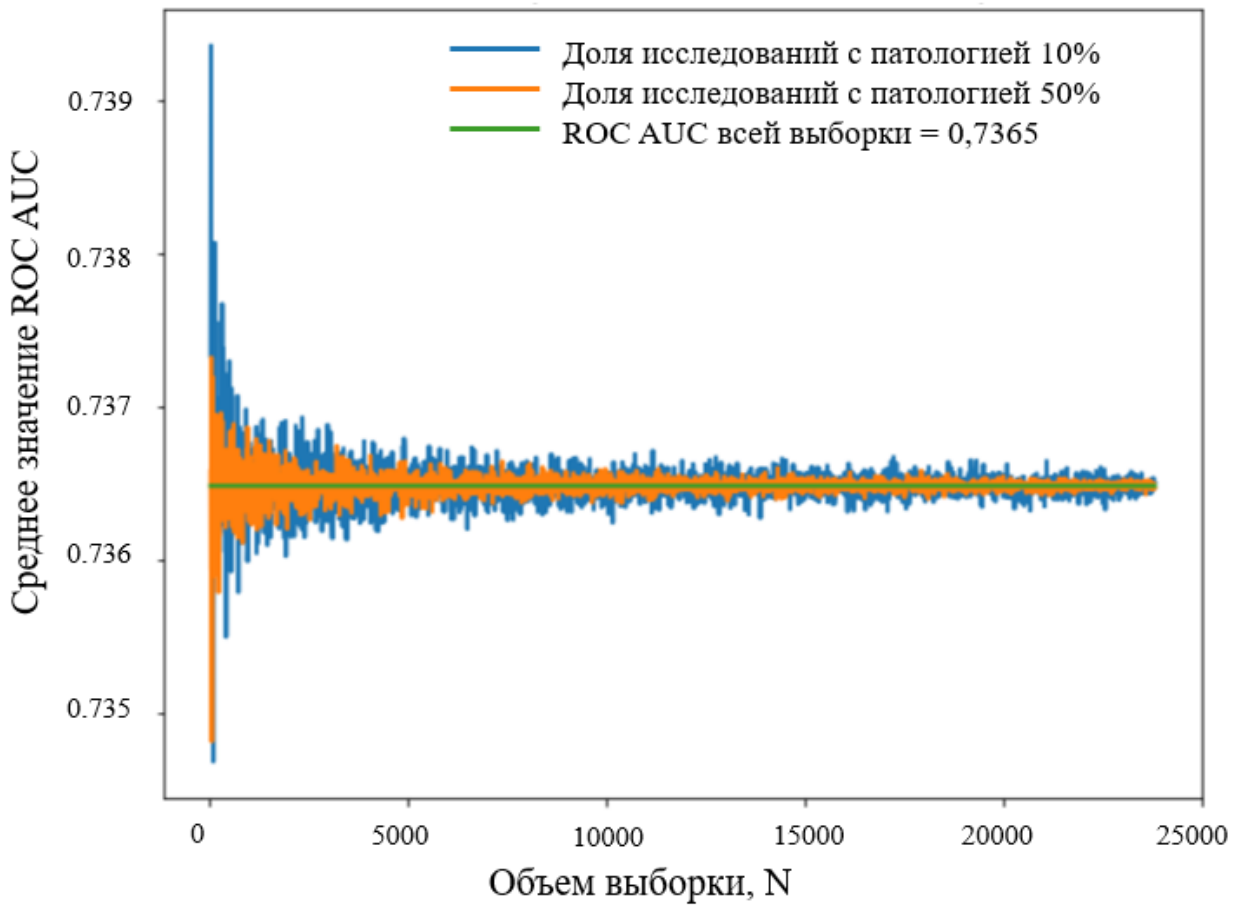


Рисунок 21 – Зависимость средних значений AUC ROC от объема выборки: для балансов классов 50 % патологии (оранжевый график) и 10 % (синий график), зеленым цветом обозначен AUC ROC всей выборки

Зависимость средних AUC ROC непрерывна в диапазоне изменений числа исследований и имеет колебательный характер, что позволяет представить ее в виде периодической функции *F(n)* и определить спектральную плотность:

$$\widehat{F(n)} = \sum_{j=1}^{N} F_j(n) * \exp(-2\pi i(\gamma, n_j)), \quad (17)$$

где n – количество образцов,

N – общее количество исследований,

γ – аргумент спектральной функции:

$$\gamma = Re(\widehat{F(n)})/Im(\widehat{F(n)}), \quad (18)$$

где *Re((F(n))* – вещественная часть спектральной функции,

Im((F(n)) – мнимая часть спектральной функции.

Далее был проведен Фурье-анализ средних значений AUC ROC. Результаты вычисления аргумента (18) спектральной функции (17) в зависимости от количества испытаний, полученные с помощью Фурье-анализа, представлены на рисунке 22.

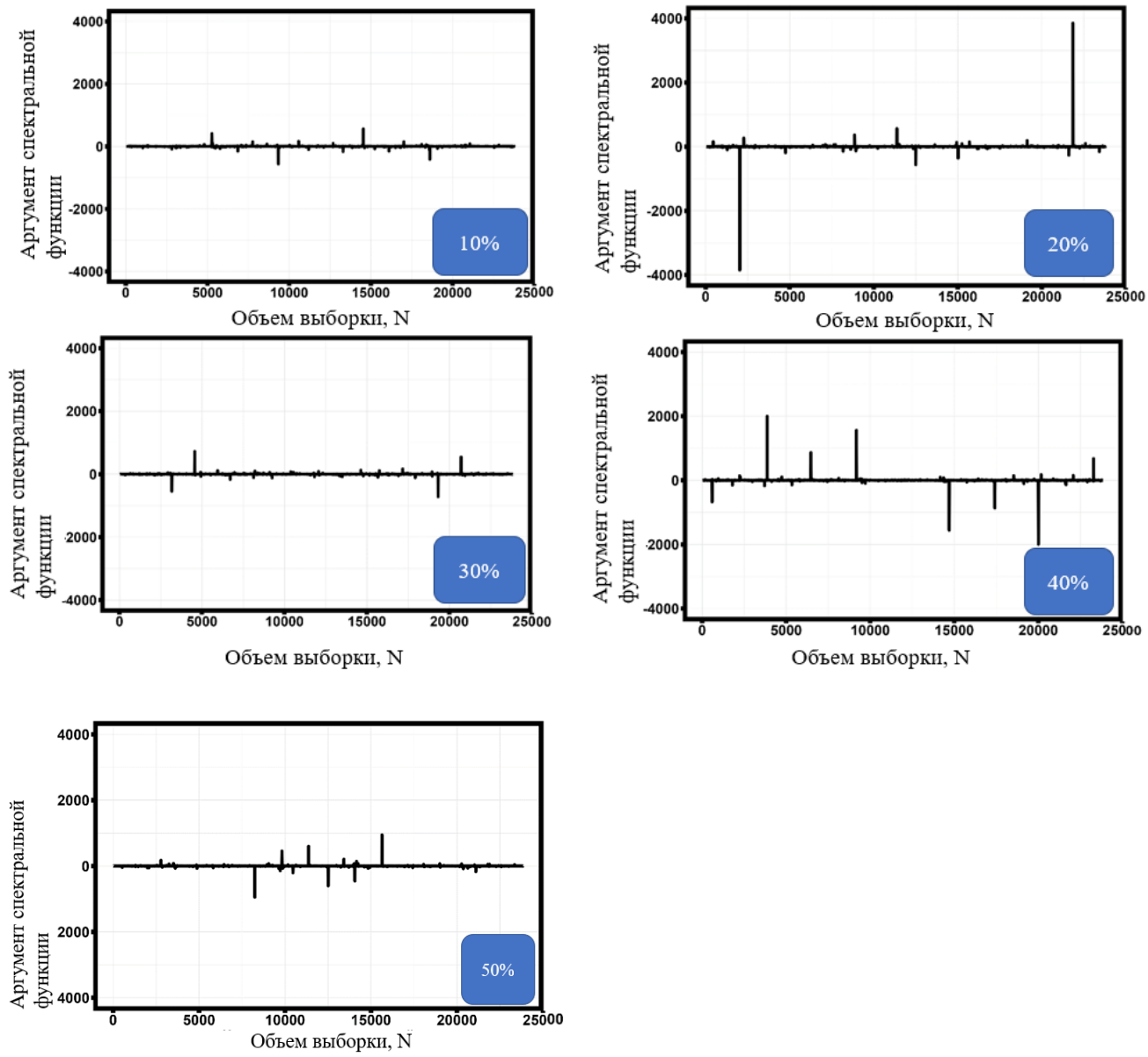


Рисунок 22 – Зависимость аргумента спектральной функции AUC ROC от количества исследований для разных балансов классов для НД1. Доля «патологии» указана внизу справа на графиках

Для балансов 20, 30, 40, 50 % можно выделить два основных паттерна поведения главных максимумов и минимумов аргумента спектральной функции AUC ROC, для баланса 10 % они менее выражены.

Далее был проведен анализ основных максимумов и минимумов аргумента спектральной функции на наличие симметрии (рисунок 23):

$$\gamma(n) + \gamma(n_T - n) = 0, \tag{19}$$

где n_T – точка симметрии аргумента спектральной функции.

Рисунок 23 – Зависимость количества исследований, соответствующих главным максимумам и минимумам аргумента спектральной функции AUC ROC, от доли «патологии» в балансе классов «норма»/«патология»

На рисунке 23 синим цветом обозначена середина интервала между первыми максимумами и минимумами аргумента спектральной функции, которая для всех рассматриваемых балансов классов получилась одинаковой и составила 11 940 исследований. Полученное значение является точкой перехода $n_{T.}$ Необходимо отметить, что Фурье-анализ использовался исключительно с целью определения точки перехода и дальнейшего анализа типа распределения. Анализ поведения аргумента спектральной функции в данной работе не проводился.

Далее с помощью метода максимального правдоподобия был определен ближайший тип распределения функций до и после точки перехода по минимуму критериев Акаике и Байеса (таблица 5).

Таблица 5 – Типы распределений до и после точки перехода n_T (11 940 исследований)

№	Доля «патологии» в балансе «норма»/«патология»	Тип распределения до n_T	Тип распределения после n_T
1	0,1	Коши	Нормальное
2	0,2	Коши	Нормальное
3	0,3	Коши	Логистическое
4	0,4	Коши	Логарифмическое нормальное
5	0,5	Коши	Логистическое

Согласно результатам анализа, для всех балансов классов до точки перехода сохраняется распределение Коши. После точки перехода тип распределения изменяется: нормальное – для балансов 0,1, 0,2, логистическое – для балансов 0,3, 0,5, и логарифмическое нормальное – для 0,5.

Далее был проведен анализ коэффициента вариации с целью определения однородности значений AUC ROC до точки перехода (11 940 исследований) для каждого баланса классов (рисунок 24). Для определенного ранее распределения Коши коэффициент вариации равен:

$$K = \frac{\gamma}{x_0} \tag{20}$$

где γ – масштабный параметр в распределении Коши,

x_0 – параметр сдвига в распределении Коши.

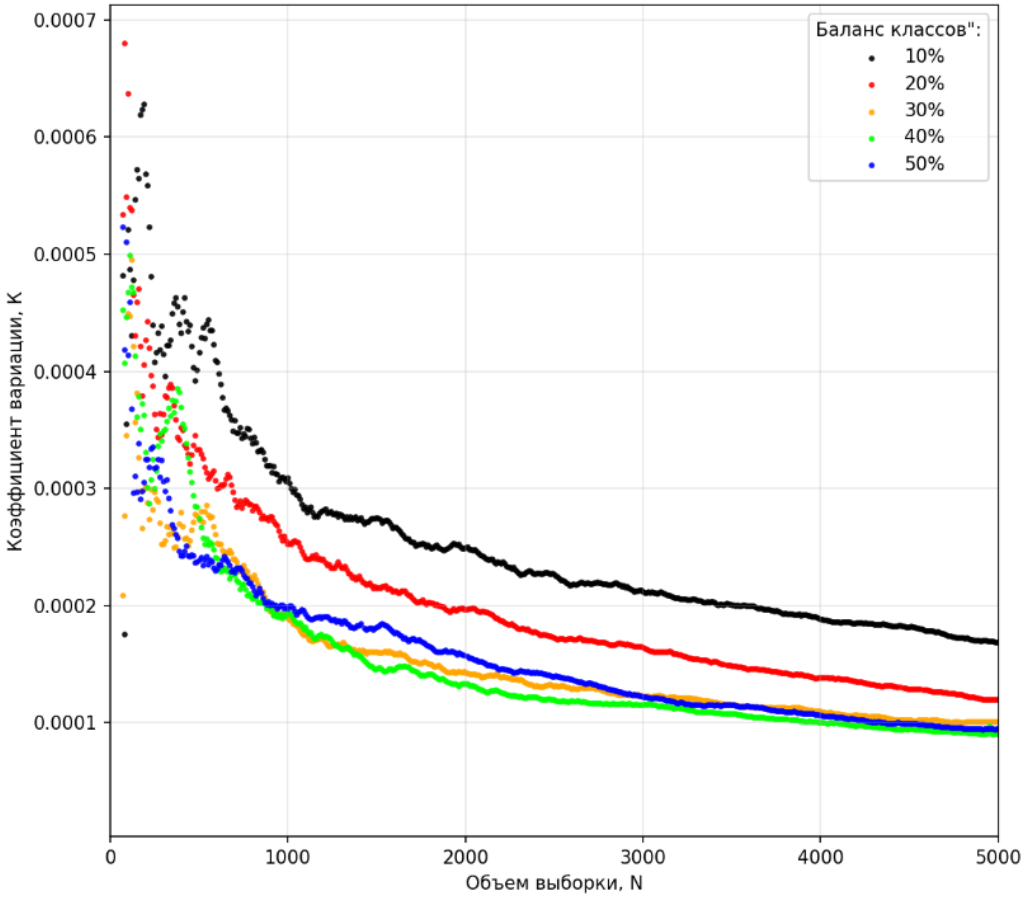


Рисунок 24 – Коэффициент вариации значений AUC ROC в зависимости от количества исследований для разных балансов классов для НД1.

При анализе зависимости коэффициента вариации средних AUC ROC от объема выборки был определен его максимум, который соответствует объему выборки, при котором наблюдается наибольшее отклонение AUC ROC от среднего значения. В соответствии с максимумом коэффициента вариации размер выборки с долей патологии 0,1 (10%) составил 190 исследований. Для выборок 0, 2 (20%) – 80 исследований, 0,3 (30%) – 120 исследований, 0,4 (40%) – 110 исследований, 0,5 (50%) – 70 исследований соответственно.

Такой же алгоритм был применен к двум другим НД, и был получен следующий результат (таблица 6).

Таблица 6 – Объем выборки для каждого исследуемого НД, при котором достигается максимум коэффициента вариабельности для разных балансов классов

Баланс классов (доля патологии, %	Количество исследований для НД1 (ММГ), шт.	Количество исследований для НД2 (ММГ), шт.	Количество исследований для НД3 (РГ), шт.	Максимальное количество исследований по трем НД, шт.
10	190	80	100	190
20	80	80	80	80
30	130	150	70	150
40	110	120	70	120
50	70	80	70	80

Полученные результаты сопоставимы друг с другом, однако наблюдаются различия в выборках при балансах 0,1, 0,3 и 0,4, что обусловлено, вероятнее всего, объемом исходного НД, который подвергался анализу. Тем не менее на балансах 0,2 и 0,5 результаты практически совпадают. Из полученных по трем НД результатов были выделены наибольшие значения объемов выборки для каждого баланса классов для дальнейшего использования на практике. Таким образом, по результатам проведенного анализа минимальный размер выборки для оценки ROC AUC на основании максимума коэффициента вариации составляет: для доли «патологии» 50 % при объеме выборки 80 исследований, 40 % – 120, 30 % – 150, 20 % – 80, 10 % – 190 исследований.

Предложенный в рамках данной работы подход отличается от предыдущих, основанных преимущественно на достижении заданной мощности [86–88]. Полученные результаты для сбалансированной выборки сопоставимы также с результатами предыдущей работы по расчету объема выборки для оценки количества технологических дефектов в НД [118], где рекомендуемый объем выборки составил 80 исследований.

Кроме того, был не только рассмотрен вопрос определения объема выборки, но и изучено поведение AUC ROC в зависимости от баланса классов: отклонение средних значений AUC ROC уменьшается при балансировке выборки по целевому признаку.

На основании полученных результатов предлагается следующий алгоритм оценки СИИ с помощью критерия AUC ROC при проведении внешних валидационных тестирований:

1. Создание НД с заданным (исходя из возможностей исследователя согласно таблице 5) балансом классов и количеством исследований.
2. Проведение тестирования СИИ на полученном НД.
3. Оценка результатов тестирования с определением AUC ROC.
2. Определение доверительного интервала для AUC ROC с помощью метода бутстрэппинга.
3. Использование нижней границы доверительного интервала в качестве оценки AUC ROC при сравнении с референсным значением.

Кроме того, важным выводом из данной работы является то, что на небольших объемах выборок наблюдаются более сильные отклонения средних значений ROC AUC, а в дальнейшем, с ростом размера выборки они снижаются. Это свидетельствует о том, что на практике значения AUC ROC будут отличаться от полученных при валидационном тестировании. Поэтому впоследствии при использовании СИИ рекомендуется проводить регулярный мониторинг [16].

ЗАКЛЮЧЕНИЕ

В конце XX века произошел значительный скачок в развитии информационных технологий, приведший к увеличению объема данных и появлению новых способов их обработки. Это нашло отражение во многих сферах, включая медицину, где началось активное использование информационных систем и ТИИ для диагностики различных патологий и прогнозирования течения заболеваний. ТИИ, включая машинное обучение и алгоритмы нейронных сетей, находят широкое применение в медицинской, в частности лучевой, диагностике. Однако все это было бы невозможно без качественных, репрезентативных НД.

Создание таких НД является одной из главных задач при разработке и тестировании СИИ. При этом необходимо учитывать сложность представления и организации медицинской информации, включая вопросы защиты персональных данных, в частности составляющих врачебную тайну. Создание стандартизированных методологий, принципов организации информации, единых представлений о структуре данных, их концептуализация и формализация позволяют не просто выработать четкий алгоритм действий при создании НД, но и учесть максимально возможное количество важных аспектов и предотвратить ряд ошибок и нерациональное использование ресурсов.

Данная диссертационная работа посвящена созданию такой универсальной методологии создания НД для тестирования СИИ в лучевой диагностике. Жизненный цикл и алгоритмы, приведенные в работе, позволят не только определить порядок действий исследователя при создании НД, но и внедрить систему менеджмента качества, то есть определить структуру, функции, процедуры, процессы и ресурсы, необходимые для организации управленческих процессов при работе с НД [49].

Результаты, представленные в данной диссертационной работе, соответствуют задачам:

1. Оценить управляемость, надежность и устойчивость процессов формирования НД, применяемых при разработке и тестировании программных средств анализа медицинских изображений в лучевой диагностике. Данные литературных источников, НД, находящиеся в открытом доступе, а также собственный опыт показали отсутствие единых алгоритмов, учитывающих важнейшие аспекты создания НД.

Следует отметить, что на самых ранних этапах Эксперимента уже были сформулированы некоторые принципы классификации НД, их наименования и организации бизнес-процессов [95, 119]. Одним из самых полезных организационных решений было введение системы версионности при внесении изменений, а также ведение перечня НД, состоящего из 8 полей [95]. Однако в дальнейшем с ростом количества направлений и НД, появлением новых, более сложных задач такое количество оказалось недостаточным и потребовало расширения и унификации параметров НД. В настоящий момент реестр содержит 101 поле, сгруппированное в соответствии с этапами жизненного цикла НД: инициирование, планирование, карточка НД (соответствует этапу формирования), использование, смена версии. Самый большой раздел – «Карточка» – содержит 69 полей, дополнительно сгруппированных по разделам: идентификация, назначение, клинические, популяционные, технические параметры и параметры разметки. Все поля реестра используются для различных задач: управленческие, контроль качества, описание НД в публикациях, РИД, формирование библиотеки НД, выдача НД для тестирования ИИ-сервисов и другое.

Разрозненность процессов, нестандартизированность информации, отсутствие классификаций, их смешение и отсутствие преемственности приводят к возникновению большого числа ошибок, дублированию информации, неоднозначности представления данных и неэффективному использованию ресурсов. Вследствие этого возникают экономические, временные и кадровые потери не только непосредственно при создании НД, но и в дальнейшем в процессе его использования, например при разработке и тестировании СИИ, что может

привести к фатальным последствиям, особенно когда речь идет о жизни и здоровье населения.

2. Обосновать принципы систематизации НД в лучевой диагностике в виде реестра и разработать концепцию выбора и применения глоссария и тезауруса для описания процессов, связанных с созданием и использованием НД. В результате проделанной работы определены важнейшие систематизирующие параметры НД, доработаны и разработаны классификации ключевых аспектов представления информации о НД. Систематизирующие принципы синхронизированы в соответствии с жизненным циклом НД с целью избегания неоднозначности понимания процессов, определены семантические связи между терминами, использующимися в контексте создания и использования НД, сформирован глоссарий и тезаурус терминов. Определенные систематизирующие параметры и классификации реализованы и представлены в виде реестра НД, который является не только перечнем основных характеристик НД, но и благодаря синхронизации с жизненным циклом НД и введению полей на этапах инициирования, планирования и использования (сроки, финансирование, ответственные лица, ссылки на сопутствующую документацию, хранение, количество тестирований, публикации и т. д.) является инструментом управления и контроля качества.

Кроме того, функция контроля качества реализуется и благодаря внедрению процессов автоматизации проверки корректности данных при заполнении реестра. Также реестр – основа для автоматизированного процесса формирования readme-файла, наличие которого вместе с файлами данных чрезвычайно важно. Именно этот файл будет изучать разработчик или исследователь для принятия решения об использовании НД для его задач (в случае, если информация о НД в достаточном объеме не отражена в библиотеке, где он хранится). Кроме того, при хранении на локальных компьютерах или в хранилище данных это основной информационный документ, который позволяет оперативно, без изучения таблиц разметки или

DICOM-файлов найти требуемый НД, а также способствует преемственности передачи данных между сотрудниками.

Разрозненное, закрытое, нестандартизованное хранение данных препятствует развитию ТИИ [96, 120]. Реестр – такая структура хранения данных, которая позволит максимально результативно их использовать, а при возможности открытой публикации позволит оперативно формировать наглядные библиотеки НД с целью развития конкуренции в сфере разработок на основе ТИИ [2].

3. Разработать подход к определению минимального размера НД для тестирования СИИ в лучевой диагностике. В рамках данной диссертационной работы предложен способ оценки объема выборки валидационных НД для определения диагностической точности СИИ. По результатам проведенного эксперимента на НД объемом более 300 000 исследований, определены зависимости поведения критериев диагностической точности от объема выборки и баланса классов. Предложено использование фиксированных размеров НД с заданным балансом классов и разработан задел для формирования методологий исследования различных критериев точности в зависимости от объемов выборки и балансов классов. Данный подход в дальнейшем может быть расширен на другие критерии диагностической точности и другие направления.

4. Создать, внедрить и оценить эффективность методов стандартизации и оптимизации процессов формирования НД. Существующие работы, описывающие процесс создания НД, как правило освещают частные случаи подготовки НД [18-20, 53, 54, 56-59]. В условиях Эксперимента (большое количество направлений, разные ИИ-сервисы, архитектура которых является коммерческой тайной, регулярные расширения и обновления направлений) требовалась единая методология создания НД для лучевой диагностики. Разработанная методология формирования НД является универсальной для НД, предназначенных для тестирования СИИ в лучевой диагностике и представляет собой следующий комплекс методик и инструментов для их реализации:

– Жизненный цикл НД;

– Методика формирования НД;

– Методы формирования унифицированных названий НД;

– Методы стандартизации и классификации НД и метаинформации;

– Реестр наборов данных – инструмент управления и контроля качества процессов создания и использования НД;

– Инструменты и методы автоматизации при подготовке НД: поиск по ключевым словам, выгрузка и анонимизация, выгрузка популяционных данных и проверка на возраст и дубликаты;

– Сопутствующая документация: регламент создания НД, шаблон ТЗ, шаблон инструкции разметчика;

– Метод обоснования размера и баланса классов НД.

Разработанные методы и инструменты легли в основу нового проекта национального стандарта 1.11.164-1.261.24 «Наборы данных для тестирования алгоритмов. Методы контроля набора данных на универсальность и структурированность» [121]. Кроме того, она может быть применена и/или адаптирована не только для других НД в лучевой диагностике, но и в других направлениях медицинской диагностики. Данная методология успешно внедрена и используется для создания НД как в Эксперименте, так и для решения научных задач. Этапность создания НД позволила не только оптимизировать организационные процессы, упростить планирование и оценку результатов, избежать ряда ошибок, существенно расширить количество направлений для тестирования и последующего внедрения СИИ, но и начать процесс автоматизации подготовки НД и дальнейшего создания платформ подготовки НД, что является приоритетным направлением повышения доступности и качества данных, необходимых для развития ТИИ согласно Национальной стратегии.

Так, на базе данной методологии в ГБУЗ «НПКЦ ДиТ ДЗМ» для удобства и ускорения процессов создания НД, а также предотвращения ошибок и решения проблем, описанных в главе 6, была создана специальная платформа подготовки НД [122]. На рисунке 25 представлен ее внешний вид. Платформа имеет

модульную структуру, соответствующую этапам жизненного цикла: сбор данных осуществляется в модулях поиска исследований (фильтрация, отбор и предварительная разметка по текстовым протоколам, включая подмодуль Label Studio для их пересмотра [123]) и выгрузки и анонимизации, а разметка и структурирование – в модуле разметки, при этом дополнительный модуль контроля качества исследований позволяет проводить анализ на наличие дефектов в DICOM-файлах в автоматическом режиме [116]. Управление процессами создания и использования НД реализовано через реестр, в нем же происходит формирование технического задания, readme-файла, контроль смены версионности, формирование таблиц и форм разметки, назначение ответственных.

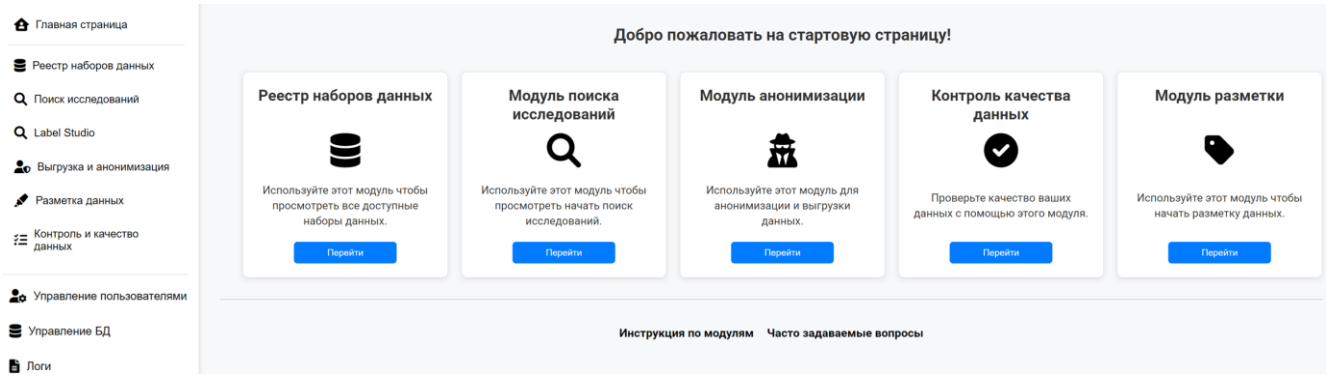


Рисунок 25 – Внешний вид главной страницы платформы подготовки наборов НД

Преимуществом платформы является удобный интерфейс, который не требует знаний в области программирования и позволяет работать различным пользователям. Автоматизация, реализованная на платформе, позволяет ускорить все процессы создания НД, оптимизирует ресурсы, способствует снижению количества ошибок и обеспечивает безопасность работы с данными. При этом платформа реализована с учетом принципов стандартизации, описанных в главе 5, что позволяет всем участникам процесса не только быть в одном информационном поле, но и обучаться работе с НД благодаря регламентированным шагам и сопровождающим их инструкциям.

Результаты диссертационной работы были успешно внедрены в практическую деятельность ГБУЗ «НПКЦ ДиТ ДЗМ» и показали свою эффективность.

Основываясь на принципах описанной методологии, непрерывно совершенствуя и оптимизируя ее, было создано более 460 НД (включая смену версии). Часть НД регулярно выкладывается в открытый доступ, преимущественно для самотестирования, с целью предоставить разработчикам возможность самостоятельно проверить работоспособность их моделей. На начало 2025 г. библиотека mosmed.ai насчитывала 71 открытый НД, из них 57 – для самотестирования диагностического, 8 – для научных исследований, 3 – для самотестирования технического, 3 – для проведения калибровочного тестирования, а также 5 НД для национальных стандартов с доступом «по запросу». На начало 2025 г. в библиотеке mosmed.ai зафиксировано 6494 скачиваний НД (рисунок 26), находящихся в открытом доступе, и более 212 000 просмотров карточек НД, что указывает на востребованность открытых данных среди исследователей и разработчиков.

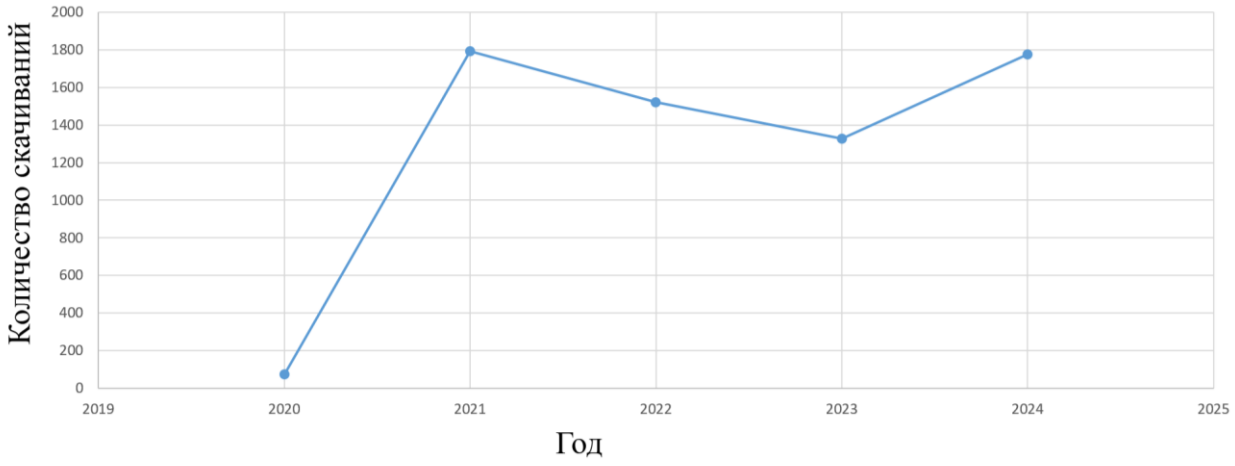


Рисунок 26 – Статистика скачиваний НД в библиотеке mosmed.ai.

НД библиотеки mosmed.ai активно используются как зарубежными, так и отечественными исследователями. По запросу «MosMedData» в реферативных научных базах имеется в общей сложности 48 публикаций (без учета публикаций

сотрудников ГБУЗ «НПКЦ ДиТ ДЗМ»), в которых используются НД, созданные в ГБУЗ «НПКЦ ДиТ ДЗМ», из них: 43 – зарубежных (Pubmed – 8, Arxiv – 6, MedRxiv – 23, Elibrary – 6), 5 – российских (Elibrary). Таким образом, это является подтверждением качества создаваемых НД, их внешней независимой валидацией.

Большинство из созданных НД используются в Эксперименте в рамках функциональных и калибровочных тестирований. В период с 2021 по 2024 г. проведено 515 тестирований ИИ-сервисов (рисунок 27), благодаря которым на конец 2024 г. в Эксперимент было интегрировано 59 ИИ-сервисов, проанализировавших за все время более 14 000 000 исследований. Следует отметить, что процесс тестирования сопровождается высококвалифицированной экспертной поддержкой. Результаты работы ИИ-сервисов дополнительно пересматривались врачами-рентгенологами, с опытом работы по данному виду исследования не менее 5 лет. В случае непрохождения тестирования эксперты подтверждали, что ошибки вызваны некорректной работой ИИ-сервиса, а не ошибками в разметке НД. Это исключает попадание ИИ-сервисов, не соответствующих требованиям, в практическую деятельность. Таким образов за все время проведения Эксперимента из двухсот протестированных ИИ-сервисов только 60 были отобраны для дальнейшей работы [124]. При этом, ИИ-сервисы, прошедшие такое тестирование, регулярно подвергаются мониторингу и подтверждают свое качество [12].

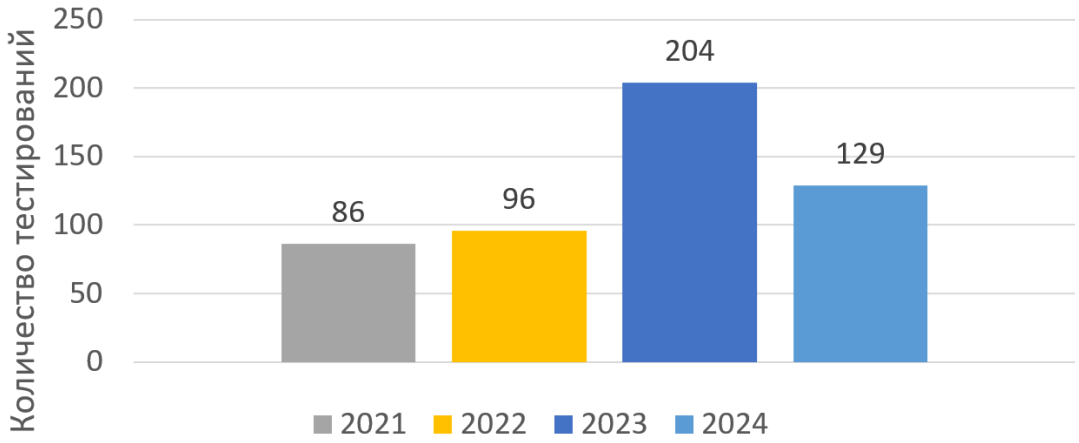


Рисунок 27 – Количество тестирований, проведенных на созданных в ГБУЗ «НПКЦ ДиТ ДЗМ» НД в 2021–2024 гг.

Алгоритм формирования НД является основой для создания специализированной платформы подготовки НД, разработка которой ведется в рамках научно-исследовательской и опытно-конструкторской работы «Разработка платформы подготовки наборов данных лучевых диагностических исследований» (№ ЕГИСУ: 123031500003-8).

Реестр ведется на регулярной основе с начала 2022 г., совершенствуясь в процессе использования. На начало 2025 г. он насчитывает 690 записей. Являясь сводной централизованной таблицей по всей информации о НД, реестр позволяет сэкономить время при поиске необходимых данных. В ГБУЗ «НПКЦ ДиТ ДЗМ» все работы ведутся в корпоративной системе «Битрикс 24». Благодаря этому весь ход работы зафиксирован в задачах. Тем не менее при необходимости извлечения какой-либо информации о НД время, затрачиваемое на поиск необходимых данных, варьирует от нескольких минут до получаса в зависимости от сложности НД. Время, необходимое для поиска с помощью реестра, составляет, как правило, несколько секунд.

Кроме того, на основе полей реестра разработан генератор readme, который позволил ускорить процесс создания readme-файла с нескольких часов в ручном режиме до нескольких секунд в автоматическом (ускорение процесса до 97 %).

Реестр регулярно используется в следующих задачах:

1. создание НД (контроль соответствия технической документации, обращение к ответственным лицам);
2. генерация readme-файла (в автоматическом режиме на основе информации из полей реестра);
3. сводная информация по запросу для отчетности или поиска НД для повторного использования (например, количество НД по заданным критериям);
4. поиск информации о НД (например, для научных исследований);
5. использование ссылки на место хранения с целью автоматизации процессов тестирования в Эксперименте.

ПЕРСПЕКТИВЫ ДАЛЬНЕЙШЕЙ РАЗРАБОТКИ ТЕМЫ

На основе проведенной работы планируются следующие шаги по расширению функционала реестра и оптимизации его использования:

1. создание дашборда на основе реестра для оперативного и наглядного доступа к ключевой информации;
2. создание инструментов автоматического заполнения реестра согласно техническим требованиям, а также автоматической проверки на их корректность;
3. автоматическое заполнение карточек библиотек НД информацией из реестра;

Разработанные методики и инструменты позволят врачам и исследователям экономить время и ресурсы в процессе сбора и аннотации данных, что в конечном счете существенно ускорит подготовку НД и поможет сделать их доступными для медицинских исследователей, врачей и специалистов в области машинного обучения. Систематизация и унификация способствуют повышению качества НД и создают управляемую, удобную информационно-коммуникационную инфраструктуру для обеспечения доступа к данным, что улучшит качество ТИИ и ускорит их внедрение в практическую деятельность с широкой областью применения и большим масштабом реализации. Планируется масштабирование результатов, адаптация под другие направления медицины, автоматизация путем создания и объединения программных модулей, что позволит построить программно-аппаратный комплекс централизованной полуавтоматической подготовки НД, что оптимизирует процессы сбора, аннотации и обработки данных. Это, в свою очередь, поможет сэкономить время и ресурсы медицинских исследователей и врачей и в конечном итоге повысит качество медицинской помощи.

ВЫВОДЫ

1. К основным недостаткам процессов создания и использования НД, применяемых при разработке и тестировании программных средств анализа медицинских изображений в лучевой диагностике, относятся: отсутствие единой методологии подготовки НД, регламентированной этапности процессов создания НД, стандартизированной сопроводительной документации, единых принципов классификации и систематизации, а также отсутствие специальных инструментов управления, используемых на всех этапах жизненного цикла НД.

2. Разработан жизненный цикл НД, соответствующие ему систематизирующие группы параметров и классификации НД по цели создания и методам верификации. Создан, апробирован и внедрен в практическую деятельность реестр НД лучевой диагностики, состоящий из 101 поля и более чем 500 записей, позволяющий реализовать процессы управления, контроля качества, обеспечивающий централизацию хранения данных, а также способствующий автоматизации формирования НД (сокращение на 97 % длительности подготовки отдельных этапов), проверки качества и их публикации, в том числе в открытых источниках. Создан глоссарий и тезаурус (список сокращений, список определений), используемые при создании НД в лучевой диагностике.

3. Разработана методика оценки диагностической точности для СИИ с бинарным исходом в зависимости от объема выборки и баланса классов, основанная на многократном моделировании процессов формирования выборок различного объема и состава с последующим анализом коэффициента вариации. Для 10 % доли «патологии» максимум коэффициента вариации достигается при количестве исследований, равном 190; для 20 % доли – 80 исследований; для 30 % доли – 150 исследований, для 40 % доли – 120 исследований, а для 50 % доли – 80 исследований. На основании полученных данных предложена следующая стратегия оценки диагностической точности СИИ при проведении валидационного тестирования: сравнение нижней границы доверительного интервала ROC AUC,

рассчитанной путем бутстрэппинга, с референсным значением.

4. Сформирована, апробирована и внедрена в клиническую практику методология создания НД лучевой диагностики, на базе которой было создано более 460 НД и зарегистрировано в качестве РИД более 40 наборов данных по различным модальностям и патологиям, а также был проведен ряд научных медицинских исследований в области ИИ на основе созданных НД.

ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ

Органам исполнительной власти субъектов Российской Федерации:

1. Реестр НД рекомендуется в качестве инструмента управления и контроля качества в процессах создания и использования медицинских НД.

Руководителям МО и разработчикам СИИ:

2. При создании НД для тестирования СИИ рекомендуется организовывать процесс подготовки и использования НД с учетом этапов жизненного цикла, а также применять алгоритм создания НД, состоящий из следующих шагов: сбор, разметка, структурирование данных, формирование файлов данных, создание readme-файла.

3. При создании НД для оценки диагностической точности СИИ с бинарным исходом использовать следующие рекомендации по минимальному размеру и составу НД: доля патологии 10 % – 190 исследований; доля 20 % – 80 исследований; доля 30 % – 150 исследований, доля 40 % – 120 исследований, доля 50 % – 80 исследований, а также при оценке ROC AUC использовать нижнюю границу доверительного интервала, рассчитанную путем бутстрэппинга, при сравнении с референсным значением.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

АРМ	–	автоматизированное рабочее место
ГБУЗ «НПКЦ ДиТ ДЗМ»	–	Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»
БДТ	–	базовые диагностические требования
БФТ	–	базовые функциональные требования
ЕРИС ЕМИАС	–	Единый радиологический информационный сервис Единой медицинской информационно-аналитической системы
ИИ	–	искусственный интеллект
КТ	–	компьютерная томография
МИС	–	медицинская информационная система
МКБ	–	Международная классификация болезней
ММГ	–	маммография
МО	–	медицинская организация
МРТ	–	магнитно-резонансная томография
НД	–	набор данных
ПО	–	программное обеспечение
РГ	–	рентгенография
РИД	–	результат интеллектуальной деятельности
СИИ	–	система искусственного интеллекта
ТЗ	–	техническое задание
ТИИ	–	технологии искусственного интеллекта
ФС	–	федеральный справочник
ЭВМ	–	электронно-вычислительная машина

AUC ROC – (receiver operating characteristic area under the curve) площадь под кривой рабочей характеристики приемника

BI-RADS – (Breast Imaging-Reporting and Data System) стандартизированная шкала оценки результатов маммографии

FDA – (Food and Drug Administration) федеральное агентство США, отвечающее за контроль безопасности товаров при продаже пищевых продуктов и медикаментов

LOINC – (Logical Observation Identifiers Names and Codes) наименования и коды идентификаторов логического наблюдения

PACS – (Picture Archiving and Communication System) система архивирования и передачи изображений

SNOMED CT – (Systematized Nomenclature of Medicine Clinical Terms) систематизированная машинно-обрабатываемая медицинская номенклатура

СПИСОК ТЕРМИНОВ

Анонимизация: действия, в результате которых происходит безвозвратная утрата способности данных быть связанными с конкретным субъектом, даже если будет использована какая-то дополнительная информация.

Библиотека НД: систематизированное собрание НД, доступных для использования.

Данные: информация, представленная в формализованном виде, пригодном для ее передачи, интерпретации и обработки с участием человека или автоматическими средствами.

Вариант НД: НД с той же спецификацией, но с другим набором исследований.

Внешняя (независимая) валидация: тестирование СИИ на данных, не использовавшихся ранее при обучении и тестировании, с целью определения ее критериев диагностической точности.

Единица НД: минимальный структурный элемент набора данных.

Жизненный цикл: развитие системы, продукции, услуги, проекта или другой создаваемой сущности от замысла до вывода из эксплуатации.

Искусственный интеллект: комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые как минимум с результатами интеллектуальной деятельности человека.

ИИ-сервис: СИИ в Эксперименте.

Калибровочное тестирование: тестирование ИИ-сервиса с целью оценки его диагностической точности и проверки соответствия заявленным значениям.

Клинический мониторинг: проверка корректности описания исследования и заключения исследования.

Ключевые слова: в контексте анализа текстовой информации – слова и/или словосочетания, указывающие на наличие той или иной патологии или признака.

Лейбл (параметр): признак, который подвергается разметке при создании НД.

Локализация: обозначение области интереса простой геометрической фигурой.

Машинное обучение: процесс автоматического обучения и совершенствования поведения СИИ на основе обработки массива обучающих данных без явного программирования.

Модальность: метод получения изображения с помощью различных медицинских устройств, которые используются для визуализации в медицинской диагностике.

Мониторинг: регулярная проверка работоспособности СИИ с целью выявления технологических дефектов (технологический мониторинг) и контроля клинической корректности (клинический мониторинг).

Набор данных (база данных): состав данных, которые структурированы или сгруппированы по определенным признакам, соответствуют требованиям законодательства Российской Федерации и необходимы для разработки программ для электронных вычислительных машин на основе ИИ.

Направление Эксперимента: сочетание вида (модальности) лучевого исследования, анатомической области и целевой клинической задачи (нахождение патологических признаков, автоматизация рутинных измерений) для работы ИИ-сервисов.

Обезличивание (псевдонимизация, деперсонализация) персональных данных: действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных.

Порог срабатывания (порог cut-off): такая вероятность наличия целевого признака, при достижении которой принимается решение о наличии этого признака.

Предразметка: способ предварительного отбора информации при создании набора данных, например с использованием какого-либо алгоритма.

Разметка (аннотация, маркировка): этап обработки структурированных и неструктурированных данных, в процессе которого данным (в том числе текстовым документам, фото- и видеоизображениям) присваиваются идентификаторы,

отражающие тип данных (классификация данных), и (или) осуществляется интерпретация данных для решения конкретной задачи, в том числе с использованием СИИ.

Реестр НД: систематизированный перечень сведений обо всех НД в учреждении (или группе учреждений), ведущийся уполномоченным сотрудником, с целью упорядочивания деятельности по формированию и использованию НД.

Самотестирование: тестирование СИИ с целью оценки возможности обработки целевых данных, наличия и работоспособности заявленного функционала.

Сегментация: обозначение области интереса путем попиксельного обведения ее границ (маска).

Сопроводительный текстовый файл (readme-файл): документ, содержащий краткую структурированную информацию о НД.

Стоп-слова: это слова и/или словосочетания, указывающие на отсутствие целевой патологии или признака при анализе текстовой информации.

Технологии искусственного интеллекта: технологии, основанные на использовании ИИ, включая компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы ИИ.

Технологический мониторинг: проверка на наличие различных типов дефектов.

Уровень разметки: уровень, на котором происходит разметка в зависимости от используемых данных: пациент, исследование, серия, изображение.

Функциональное тестирование: тестирование ИИ-сервиса с целью технологической проверки его интегрируемости в ЕРИС ЕМИАС и оценки наличия и работоспособности заявленного функционала.

Эксперимент: эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения этих технологий в системе здравоохранения.

СПИСОК ЛИТЕРАТУРЫ

1. Гусев, А. В. Создание региональных фрагментов ЕГИСЗ: текущие результаты и анализ программ дальнейшего развития информационных систем в области здравоохранения / А. В. Гусев // Врач и информационные технологии. – 2013. – Т. 6. – С. 15–25.

2. Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» // ГАРАНТ [Электронный ресурс]. – https://base.garant.ru/72838946/ (дата обращения: 24.03.2024).

3. Гусев, А. В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения / А. В. Гусев // Врач и информационные технологии. – 2017. – №. 3. – С. 92–105.

4. Benjamens, S. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database / S. Benjamens, P. Dhunnoo, B. Meskó // NPJ Digital Medicine. – 2020. – Vol. 3, No. 1. – P. 1–8.

5. Алексеева М. Г. Искусственный интеллект в медицине / М. Г. Алексеева, А. И. Зубов, М. Ю. Новиков // Международный научно-исследовательский журнал. – 2022. – Т. 7, № 121 (Часть 2). – С. 10–13.

6. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента: монография / под ред. Ю. А. Васильева, А. В. Владзимирского. – М. : Издательские решения, 2022. – 388 с.

7. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis / S. E. Hickman, R. Woitek, E. P. V. Le, et al. // Radiology. – 2022. – Vol. 302, No. 1. – P. 88–104.

8. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis / C. Leibig, M. Brehmer, S. Bunk, et al. // Lancet Digit Health. – 2022. – Vol. 4, No. 7. – P. e507–e519.

9. Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis / J. Liu, J. Lei, Y. Ou, et al. // Clinical and Experimental Medicine. – 2023. – Vol. 23, No. 6. – P. 2341-2356.

10. Применение компьютерного зрения для профилактических исследований на примере маммографии / К. М. Арзамасов, Ю. А. Васильев, А. В. Владзимирский [и др.] // Профилактическая медицина. – 2023. – Т. 26, № 6. – С. 117.

11. Первые 10000 маммографических исследований, выполненных в рамках услуги «описание и интерпретация данных маммографического исследования с использованием искусственного интеллекта / Ю. А. Васильев, А. В. Владзимирский, К. М. Арзамасов [и др.] // Менеджер здравоохранения. – 2023. – № 8. – С. 54–67.

12. Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы [Электронный ресурс]. – URL: https://mosmed.ai/ai/ (дата обращения: 24.03.2024).

13. Тарифное соглашение 2023 – ДЗМ [Электронный ресурс]. – URL: https://mosgorzdrav.ru/ru-RU/document/default/view/2490.html (дата обращения: 24.03.2024).

14. Ngiam, K. Y. Big data and machine learning algorithms for health-care delivery / K.Y. Ngiam, I. W. Khor // Lancet Oncol. – 2019. – Vol. 20, No. 5. – P. e262–e273.

15. Оценка зрелости технологий искусственного интеллекта для здравоохранения : методические рекомендации / сост. Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.]. – ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. – 28 с. – (Серия «Лучшие практики лучевой и инструментальной диагностики» ; вып. 123).

16. Оценка зрелости технологий искусственного интеллекта для здравоохранения: методология и ее применение на материалах московского эксперимента по компьютерному зрению в лучевой диагностике / И. А. Тыров, Ю. А. Васильев, К. М. Арзамасов [и др.] // Врач и информационные технологии. – 2022. – Т. 4. – С. 76–92.

17. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis / R. Aggarwal, V. Sounderajah, G. Martin, et al. // NPJ Digital Medicine. – 2021. – Vol. 4, No. 1. – P. 1–23.

18. Особенности создания базы данных нейроонкологических 3D МРТ-изображений для обучения искусственного интеллекта / Е. В. Амелина, А. Ю. Летягин, Б. Н. Тучинов [и др.] // Сибирский научный медицинский журнал. – 2022. – Т. 42, № 6. – С. 51–59.

19. Косарева, А. А. Алгоритм подготовки набора данных для обучения нейронных сетей на примере задачи анализа радиологических изображений лёгких / А. А. Косарева // Доклады БГУИР. – 2023. – Т. 21, № 1. – С. 66–73.

20. Окунев, С. В. Рассмотрение способов формирования наборов данных для обучения нейронных сетей / С. В. Окунев // Вестник науки и образования. – 2020. – № 2 (80), Часть 3. – С. 16-19.

21. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group / V. Sounderajah, H. Ashrafian, R. Aggarwal, et al. // Nature Medicine. – 2020. – Vol. 26, No. 6. – P. 807–808.

22. Cabitza, F. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies / Cabitza F., Campagner A. // Int J Med Inform. – 2021. – Vol. 153. – P. 104510.

23. ГОСТ Р 59895-2021. Национальный стандарт РФ: «Технологии искусственного интеллекта в образовании. Общие положения и

терминология» (утв. и введен в действие приказом Федерального агентства по техническому регулированию и метрологии от 26.11.2021 № 1617-ст) // ГАРАНТ [Электронный ресурс]. – URL: https://base.garant.ru/403602944/ (дата обращения: 06.04.2024).

24. Artificial Intelligence in Pathology / H. Y. Chang, C. K. Jung, J. I. Woo, et al. // J Pathol Transl Med. – 2019. – Vol. 53, No. 1. – P. 1–12.

25. ЕРИС – LAVAL – Информационные системы и технологии для медицины [Электронный ресурс]. – URL: http://lvlmed.ru/eris/ (дата обращения: 24.03.2024).

26. Шулькин, И. М. Управление на основе данных в лучевой диагностике: оценка результативности модели единого радиологического информационного сервиса / И. М. Шулькин, А. В. Владзимирский // Менеджер здравоохранения. – 2022. – № 7. – С. 68–80.

27. Указ Президента Российской Федерации от 07.05.2018 № 204 «О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года» [Электронный ресурс]. – URL: http://www.kremlin.ru/acts/bank/43027 (дата обращения: 24.03.2024).

28. External validation of prognostic models: what, why, how, when and where? / C. L. Ramspek, K. J. Jager, F. W. Dekker, et al. // Clin Kidney J. – 2020. – Vol. 14, No. 1. – P. 49-58.

29. Machine learning for technical skill assessment in surgery: a systematic review / K. Lam, J. Chen, Z. Wang, et al. // NPJ Digit Med. – 2022. – Vol. 5, No. 1. – P. 24.

30. Свидетельство о государственной регистрации программы для ЭВМ № 2023665713 Российская Федерация. Веб-платформа технологического и клинического мониторинга результатов работы алгоритмов анализа цифровых медицинских изображений : № 2023664691 : заявл. 11.07.2023 : опубл. 19.07.2023 / Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.] ; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

31. Свидетельство о государственной регистрации программы для ЭВМ № 2022617324 Российская Федерация. Веб-инструмент для выполнения ROC анализа результатов диагностических тестов: № 2022616046: заявл. 05.04.2022: опубл. 19.04.2022 / С. П. Морозов, А. Е. Андрейченко, С. Ф. Четвериков [и др.]; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

32. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness / S. Vollmer, B. A. Mateen, G. Bohner, et al. // BMJ. – 2020. – Vol. 368. – P. l6927.

33. Demler, O. V. Misuse of DeLong test to compare AUCs for nested models. / O. V. Demler, M. J. Pencina, R. B. D'Agostino Sr // Stat Med. – 2012. – Vol. 31, No. 23. – P. 2577-2587.

34. Pauly, M. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data / M. Pauly, T. Asendorf, F. Konietschke // Biom J. – 2016. – Vol. 58, No. 6. – P. 1319–1337.

35. ROC Analysis [Электронный ресурс]. – URL: https://roc-analysis.mosmed.ai/ (дата обращения: 08.02.2025).

36. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases / X. Wang, Y. Peng, L. Lu, et al. // Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. – Institute of Electrical and Electronics Engineers Inc, 2017. – P. 3462–3471.

37. Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 № 149-ФЗ (последняя редакция) \ КонсультантПлюс [Электронный ресурс]. – URL: https://www.consultant.ru/document/cons_doc_LAW_61798/ (дата обращения: 17.03.2023).

38. Межгосударственный стандарт ГОСТ 34.320-96 «Информационные технологии. Система стандартов по базам данных. Концепции и терминология для концептуальной схемы и информационной

базы» (введен в действие постановлением Государственного комитета Российской Федерации по стандартизации и метрологии от 22.02.2001 № 87-ст) // ГАРАНТ [Электронный ресурс]. – URL: https://base.garant.ru/71444660/ (дата обращения: 06.04.2024).

39. Мирошниченко, Е. А. К формальному определению понятия база данных / Е. А. Мирошниченко // Проблемы информатики. – 2011. – № 2. – С. 83-87.

40. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review / W. Sun, Z. Cai, Y. Li, et al. // J Healthc Eng. – 2018. – Vol. 2018. – P. 4302425.

41. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets / D. Newman-Griffis, G. Divita, B. Desmet, et al. // J Am Med Inform Assoc. – 2021. – Vol. 28, No. 3. – P. 516.

42. UMLS Metathesaurus – CPT (CPT – Current Procedural Terminology) – Metadata [Электронный ресурс]. – URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT/metadata.html (дата обращения: 19.04.2022).

43. SNOMED - Home | SNOMED International [Электронный ресурс]. – URL: https://www.snomed.org/ (дата обращения: 19.04.2022).

44. Logical Observation Identifiers Names and Codes [Электронный ресурс]. – URL: https://loinc.org/ (дата обращения: 19.04.2022).

45. RadLex Term Browser [Электронный ресурс]. – URL: http://radlex.org/ (дата обращения: 15.04.2024).

46. Filice, R. W. Biomedical Ontologies to Guide AI Development in Radiology / R. W. Filice, C. E. Kahn // J Digit Imaging. – 2021. – Vol. 34, No. 6. – P. 1331–1341.

47. Использование нейронных сетей для поиска нарушений укладки пациента на рентгенограммах органов грудной клетки / А. А. Борисов, Ю. А.

Васильев, А. В. Владзимирский [и др.] // Программные системы: теория и приложения. – 2023. – Т. 14, № 3. – С. 95–113.

48. Применение технологий искусственного интеллекта как способ обеспечения качества выполнения рентгенографии органов грудной клетки / А. А. Борисов, Ю. А. Васильев, А. В. Владзимирский [и др.] // Менеджер здравоохранения. – 2023. – № 7. – С. 91–101.

49. ГОСТ Р 59921.5-2022. Национальный стандарт РФ «Системы искусственного интеллекта в клинической медицине. Часть 5. Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов» (утв. и введен в действие приказом Федерального агентства по техническому регулированию и метрологии от 31.03.2022 № 180-ст) / ГАРАНТ [Электронный ресурс]. – URL: https://base.garant.ru/404786491/ (дата обращения: 06.04.2024).

50. The National Cancer Informatics Program (NCIP) Annotation and Image Markup (AIM) Foundation Model / P. Mongkolwat, V. Kleper, S. Talbot, D. Rubin // J Digit Imaging. – 2014. – Vol. 27, No. 6. – P. 692.

51. Свидетельство о государственной регистрации программы для ЭВМ № 2020664321 Российская Федерация. MedLabel – автоматизированный анализ медицинских протоколов: № 2020663035: заявл. 27.10.2020: опубл. 11.11.2020 / С. П. Морозов, А. Е. Андрейченко, Ю. С. Кирпичев [и др.]; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

52. Обзор современных средств разметки цифровых диагностических изображений / Ю. А. Васильев, Е. Ф. Савкина, А. В. Владзимирский [и др.] // Казанский медицинский журнал. – 2023. – Т. 104, № 5. – С. 750–760.

53. RIL-Contour: a Medical Imaging Dataset Annotation Tool for and with Deep Learning / K. A. Philbrick, A. D. Weston, Z. Akkus, et al. // J Digit Imaging. Springer. – 2019. – Vol. 32, No. 4. – P. 571.

54. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem / B. G. S. Cruz, M. N. Bossa, J. Sölter, A. D. Husch // Med Image Anal. – 2021. – Vol. 74. – P. 102225.

55. Васильев, Ю. А. Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта : учебное пособие / Ю. А. Васильев, К. М. Арзамасов, А. В. Владзимирский [и др.]. – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», Издательские решения по лицензии Ridero, 2024. – 140 с.

56. Борбат, А. М. Первый российский набор данных гистологических изображений патологических процессов молочной железы / А. М. Борбат, С. В. Лищук // Врач и информационные технологии. – 2020. – № 3. – С. 25–30.

57. Методология и инструментарий создания обучающих выборок для систем искусственного интеллекта по распознаванию рака легкого на КТ-изображениях / Н. С. Кульберг, М. А. Гусев, Р. В. Решетников [и др.]. // Здравоохранение Российской Федерации. – 2021. – Т. 64, № 6. – С. 343–350.

58. Multi-expert annotation of Crohn’s disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network / A. de Maissin, R. Vallée, M. Flamant, et al. // Endosc Int Open. – 2021. – Vol. 9, № 7. – P. E1136.

59. Image analysis for retinopathy of prematurity / M. F. Chiang, R. Gelman, M. E. Martinez-Perez, et al. // Journal of AAPOS – 2009. – Vol. 13, No. 5. – P. 438-445.

60. Expert Image Analysis – Medical Metrics [Электронный ресурс]. – URL: https://medicalmetrics.com/services/expert-image-analysis/ (дата обращения: 16.03.2024).

61. Борисов, Р. С. Протокол обработки наборов данных для их публикации в открытых источниках / Р. С. Борисов, А. А. Ефименко // Правовая информатика. – 2021. – № 2. – С. 59–70.

62. Методические рекомендации по применению приказа Роскомнадзора от 05.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных» от 13.12.2013 [Электронный ресурс]. – URL: https://docs.cntd.ru/document/420281168 (дата обращения: 24.03.2024).

63. Working Party Article 29 – Opinion 05-2014 on Anonymisation Techniques : Nederlandse overheid : Free Download, Borrow, and Streaming : Internet Archive [Электронный ресурс]. – URL: https://archive.org/details/blg-749795 (дата обращения: 24.03.2024).

64. Datasheets for Datasets. Documentation to facilitate communication between dataset creators and consumers / T. Gebru, J. Morgenstern, B. Vecchione, et al. // Commun ACM. – 2021. – Vol. 64, No. 12. – P. 86-92.

65. Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology / A. Homeyer, C. Geißler, L. O. Schwen, et al. // Modern Pathology. – 2022. – Vol. 35, No. 12. – P. 1759–1769.

66. Arts, D. G. T. Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework / D. G. T. Arts, N. F. de Keizer, G.-J. Scheffer, et al. // J Am Med Inform Assoc. – 2002. – Vol. 9. – P. 600–611.

67. Reda, O. A systematic literature review on data quality assessment / O. Reda, N. C. Benabdellah, A. Zellou // Bulletin of Electrical Engineering and Informatics. – 2023. – Vol. 12, – № 6. – P. 3736–3757.

68. Zaletel, M. Methodological guidelines and recommendations for efficient and rational governance of patient registries / M. Zaletel, M. Kralj. – National Institute of Public Health, 2015.

69. Data quality assessment and improvement / R. Silvola, J. Härkönen, O. Vilppola, et al. // Int J Bus Inf Syst. – 2016. – Vol. 22, No. 1. – P. 62–81.

70. Oakden-Rayner, L. Exploring Large-scale Public Medical Image Datasets / L. Oakden-Rayner // Acad Radiol. – 2020. – Vol. 27, No. 1. – P. 106–112.

71. Bland, J. M. The tyranny of power: is there a better way to calculate sample size? / J. M. Bland // BMJ. – 2009. – Vol. 339, No. 7730. – P. 1133–1135.

72. Hidden Variables in Deep Learning Digital Pathology and Their Potential to Cause Batch Effects: Prediction Model Study / M. Schmitt, R. C. Maron, A. Hekler, et al. // J Med Internet Res. – 2021. – Vol. 23, No. 2. – P. e23436

73. Preparing medical imaging data for machine learning / M. J. Willemink, W. A. Koszek, C. Hardell, et al. // Radiology. – 2020. – Vol. 295, No. 1. – P. 4–15.

74. Impact of dataset size on classification performance: An empirical evaluation in the medical domain / A. Althnian, D. AlSaeed, H. Al-Baity, et al. // Applied Sciences. – 2021. – Vol. 11, No. 2. – P. 1–18.

75. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis / A. J. Larrazabal, N. Nieto, V. Peterson, et al. // Proc Natl Acad Sci U S A. – 2020. – Vol. 117, No. 23. – P. 12592–12594.

76. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation / A. D. Schütte, J. Hetzel, S. Gatidis, et al. // NPJ Digital Medicine. – 2021. – Vol. 4, No. 1. – P. 141.

77. Kohli, M. D. Medical Image Data and Datasets in the Era of Machine Learning – Whitepaper from the 2016 C-MIMI Meeting Dataset Session / M. D. Kohli, R. M. Summers, J. R. Geis // J Digit Imaging. – 2017. –Vol. 30, No. 4. – P. 392–399.

78. Understanding Biases and Disparities in Radiology AI Datasets: A Review / S. Tripathi, K. Gabriel, S. Dheer, et al. // Journal of the American College of Radiology. – 2023. – Vol. 20, No. 9. – P. 836–841.

79. The Clinician and Dataset Shift in Artificial Intelligence / S. G. Finlayson, A. Subbaswamy, K. Singh, et al. // N Engl J Med. – 2021. – Vol. 385, No. 3. – P. 283.

80. Hajian-Tilaki, K. Sample size estimation in diagnostic test studies of biomedical informatics / K. Hajian-Tilaki // J Biomed Inform. – 2014. – Vol. 48. – P. 193–204.

81. Fox, N. Sampling and Sample Size Calculation / Fox N., Hunn A. // The NIHR RDS for East Midlands. – 2009. – Vol. 1, No. 1.

82. Lehr, R. Sixteen S-squared over D-squared: a relation for crude sample size estimates / R. Lehr // Stat Med. – 1992. – Vol. 11, No. 8. – P. 1099–1102.

83. Altman, D. G. How large a sample? / D. G. Altman // Statistics in Practice. – London, UK: British Medical Association. – 1982.

84. Estimation of required sample size for external validation of risk models for binary outcomes / M. Pavlou, C. Qu, R. Z. Omar, et al. // Stat Methods Med Res. – 2021. – Vol. 30, No. 10. – P. 2187–2206.

85. Бусыгина, Ю. С. Оценка точности диагностики COVID-19 на КТ-исследованиях алгоритмами искусственного интеллекта / Ю. С. Бусыгина, Т. М. Бобровская, С. С. Семенов // X международный молодёжный научный медицинский форум «Белые цветы», посвященный 150-летию С.С. Зимницкого : сборник тезисов ; Казань, 12–14 апреля 2023 года. – Казань : Казанский ГМУ, 2023. – С. 939.

86. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models / Y. Vergouwe, E. W. Steyerberg, M. J. Eijkemans, J. D. Habbema // J Clin Epidemiol. – 2005. – Vol. 58, No. 5. – P. 475–483.

87. Collins, G. S. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study / G. S. Collins, E. O. Ogundimu, D. G. Altman // Stat Med. – 2016. – Vol. 35, No. 2. – P. 214–226.

88. A calibration hierarchy for risk models was defined: from utopia to empirical data / B. van Calster, D. Nieboer, Y. Vergouwe, et al. // J Clin Epidemiol. – 2016. – Vol. 74. – P. 167–176.

89. Calibration: the Achilles heel of predictive analytics / B. van Calster, D. J. McLernon, M. van Smeden, et al. // BMC Med. – 2019. – Vol. 17, No. 1. – P. 230.

90. Клинические испытания программного обеспечения на основе интеллектуальных технологий (лучевая диагностика) : методические рекомендации / С. П. Морозов, А. В. Владзимирский, В. Г. Кляшторный [и др.]. – Москва : ГБУЗ «НПКЦ ДиТ ДЗМ», 2019. – 33 с. – (Серия «Лучшие практики лучевой и инструментальной диагностики»; вып. 23).

91. Minimum sample size for external validation of a clinical prediction model with a binary outcome / R. D. Riley, T. P. A. Debray, G. S. Collins, et al. // Stat Med. – 2021. – Vol. 40, No. 19. – P. 4230–4251.

92. MIMIC-IV, a freely accessible electronic health record dataset / A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. // Sci Data. – 2023. – Vol. 10, No. 1. – P. 1.

93. Knowledge Graph-Enabled Cancer Data Analytics / S. M. S. Hasan, D. Rivera, X. C. Wu, et al. // IEEE J Biomed Health Inform. – 2020. – Vol. 24, No. 7. – P. 1952.

94. Черняков, А. Н. Обзор информационных платформ – источников наборов данных для построения моделей машинного обучения в ритейле / А. Н. Черняков // Инновации и инвестиции. – 2023. – № 3.

95. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике / Н. А. Павлов, А. Е. Андрейченко, А. В. Владзимирский [и др.] // Digital Diagnostics. – 2021. – Vol. 2, No. 1. – P. 49–66.

96. Wang, J. W. Registries, Databases and Repositories for Developing Artificial Intelligence in Cancer Care / J. W. Wang, M. Williams // Clin Oncol (R Coll Radiol). – 2022. – Vol. 34, No. 2. – P. e97–e103.

97. Datasets [Электронный ресурс]. – URL: kaggle.com/datasets/ (дата обращения: 24.03.2024).

98. Collections | IDC [Электронный ресурс]. – URL: https://portal.imaging.datacommons.cancer.gov/collections/ (дата обращения: 24.03.2024).

99. Larobina, M. Thirty Years of the DICOM Standard / M. Larobina // Tomography. – 2023. – Vol. 9, No. 5. – P. 1829-1838.

100. Breast Imaging Reporting & Data System | American College of Radiology [Электронный ресурс]. – URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads (дата обращения: 23.01.2024).

101. C.2.2 Patient Identification Module [Электронный ресурс]. – URL: https://dicom.nema.org/Medical/dicom/2016d/output/chtml/part03/sect_C.2.2.html (дата обращения: 24.03.2024).

102. How many bootstrap replicates are necessary? / N. D. Pattengale, M. Alipour, O. R. P. Bininda-Emonds, et al. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin, Heidelberg, – 2009. – Vol. 5541 LNBI. – P. 184–200.

103. Myung, I. J. Tutorial on maximum likelihood estimation / I. J. Myung // J Math Psychol. – 2003. – Vol. 47, No. 1. – P. 90–100.

104. Возможности и ограничения использования инструментов машинной обработки текстов в лучевой диагностике / Д. Ю. Кокина, В. А. Гомболевский, К. М. Арзамасов [и др.] // Digital Diagnostics. – 2022. – Т. 3, – № 4. – С. 374–383.

105. NIH Chest X-rays [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/nih-chest-xrays/data (дата обращения: 25.03.2024).

106. ГОСТ Р 59921.6-2021. Системы искусственного интеллекта в клинической медицине. Часть 6. Общие требования к эксплуатации – ФГБУ «Институт стандартизации» [Электронный ресурс]. – URL: https://www.gostinfo.ru/catalog/Details/?id=6877735 (дата обращения: 06.04.2024).

107. Методология тестирования и мониторинга программного обеспечения на основе технологий искусственного интеллекта для медицинской диагностики / Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.] // Digital Diagnostics. – 2023. – Т. 4, № 3. – С. 252–267.

108. ГОСТ Р 59921.1-2022. Системы искусственного интеллекта в клинической медицине. Часть 1. Клиническая оценка [Электронный ресурс]. – URL: https://internet-law.ru/gosts/gost/78211/ (дата обращения: 06.04.2024).

109. Breast Imaging Reporting & Data System | American College of Radiology [Электронный ресурс]. – URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads (дата обращения: 28.03.2024).

110. Use of the Alberta Stroke Program Early CT Score (ASPECTS) for assessing CT scans in patients with acute stroke – PubMed [Электронный ресурс]. – URL: https://pubmed.ncbi.nlm.nih.gov/11559501/ (дата обращения: 28.03.2024).

111. Наборы данных [Электронный ресурс]. – URL: https://mosmed.ai/datasets/ (дата обращения: 24.03.2024).

112. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике / Ю. А. Васильев, Т. М. Бобровская, К. М. Арзамасов [и др.] // Менеджер здравоохранения. – 2023. – № 4. – С. 28–41.

113. Федеральный справочник инструментальных диагностических исследований [Электронный ресурс]. – URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1471/version/2.15 (дата обращения: 15.04.2024).

114. Справочник НСИ. Анатомические локализации [Электронный ресурс]. – URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477/version/4.3 (дата обращения: 15.04.2024).

115. Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review / Y. Peng, F. Bathelt, R. Gebler, et al. // JMIR Med Inform. – 2024. – Vol. 12, № 1. – P. e52967.

116. Свидетельство о государственной регистрации программы для ЭВМ № 2023619686 Российская Федерация. Веб-инструмент для контроля качества датасетов : № 2023617136 : заявл. 13.04.2023: опубл. 15.05.2023 / Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.]; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

117. Свидетельство о государственной регистрации программы для ЭВМ № 2023617333 Российская Федерация. Модуль контроля качества результатов диагностических исследований по РГ ОГК: № 2023615822: заявл. 28.03.2023: опубл. 07.04.2023 / Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.]; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

118. Подходы к формированию выборки для контроля качества работы систем искусственного интеллекта в медико-биологических исследованиях / С. Ф. Четвериков, К. М. Арзамасов, А. Е. Андрейченко [и др.] // Современные технологии в медицине. – 2023. – Т. 15, № 2. – С. 19–25.

119. Регламент подготовки наборов данных с описанием подходов к формированию репрезентативной выборки данных. Часть 1 : методические рекомендации / сост. С. П. Морозов, А. В. Владзимирский, А. Е. Андрейченко [и др.] – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», 2021. – 40 с. – (Серия «Лучшие практики лучевой и инструментальной диагностики», вып. 103).

120. The FAIR Guiding Principles for scientific data management and stewardship / M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, et al. // Scientific Data. – 2016. – Vol. 3, No. 1. – P. 1–9.

121. ПНС 1.11.164-1.261.24 «Наборы данных для тестирования алгоритмов. Методы контроля набора данных на универсальность и структурированность» [Электронный ресурс]. – URL:https://base.garant.ru/411572439/ (дата обращения: 29.09.2025).

122. Свидетельство о государственной регистрации программы для ЭВМ № 2025610804 Российская Федерация. Платформа подготовки наборов данных: № 2024691653: заявл. 20.12.2024: опубл. 14.01.2025 / Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.]; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

123. Label Studio. [Электронный ресурс]. – URL: https://labelstud.io/ (дата обращения: 30.01.2025).

124. Искусственный интеллект в лучевой диагностике: Per Aspera Ad Astra / Ю. А. Васильев, Ю. А. Алымова, К. М. Арзамасов [и др.]. – Москва : ГБУЗ "НПКЦ ДиТ ДЗМ", 2025. – 493 с. – ISBN 978-5-0067-5622-9

ПРИЛОЖЕНИЕ А

Перечень и краткое описание полей реестра наборов данных

I. Инициирование:

1. Предварительное (рабочее) название – название в произвольной форме, предназначенное для идентификации НД на этапе инициации проекта.
2. Заказчик / контактное лицо – Ф. И. О. сотрудника, являющегося заказчиком проекта и/или контактным лицом.
3. Финансирование – источник финансирования.
4. Финансирование (англ.) – источник финансирования на английском языке.
5. Тип – номер, согласно классификации НД по цели создания (таблица 2).
6. Ответственный за формирования БДТ – Ф. И. О. сотрудника, ответственного за формирование БДТ.
7. Дата начала работы над БДТ.
8. Дата утверждения БДТ.
9. Ссылка на БДТ – ссылка на место хранения или публикации БДТ.

II. Планирование:

10. Ссылка на место хранения ТЗ.
11. Планируемая дата начала формирования НД.
12. Планируемая дата завершения формирования НД.
13. Актуальный статус – на каком этапе работ находится НД на момент актуализации реестра. По завершению всех работ, ставится статус «готов».
14. Дата смены статуса – дата на момент актуализации статуса.
15. Комментарий к статусу.
16. Разметчики – Ф. И. О. специалистов, ответственных за разметку данных.

17. Ответственный – Ф. И. О. сотрудника, ответственного за создание НД.

III. Карточка НД (соответствует этапу формирования):

Идентификация

18. Номер НД – порядковый номер в реестре.

19. Внутренний идентификатор – уникальный идентификатор НД для внутреннего использования (в т. ч. для наименования файлов данных).

20. Публичный идентификатор – уникальный идентификатор набора данных для публикации в открытом доступе (в т. ч. для наименования файлов данных).

21. Полное наименование – полное название НД на русском языке для публикации в открытом доступе, оформления РИД и т.д.

22. Год создания НД.

23. Версия – версия НД в формате А.Б.В. [95].

24. Условия доступа – открытый, закрытый или закрытый с публичными примерами.

25. Авторы – Ф. И. О. авторского состава НД.

Клинические параметры

26. Модальность – аббревиатура вида диагностического исследования согласно стандарту DICOM.

27. Уникальный идентификатор анатомической локализации – номер согласно ФС инструментальных диагностических исследований [113].

28. Код Radlex – код анатомической локализации целевой области согласно справочнику RadLex [45].

29. Код LOINC – код анатомической локализации целевой области согласно справочнику LOINC [44].

30. Наименование анатомической локализации (русское) согласно ФС инструментальных диагностических исследований – полное наименование анатомической локализации целевой модальности согласно указанному справочнику [113] на русском языке.

31. Наименование анатомической локализации (русское) согласно ФС анатомических локализаций – полное наименование анатомической локализации целевой области согласно указанному справочнику [114] на русском языке.

32. Наименование анатомической локализации (английское) согласно RadLex – полное наименование анатомической локализации целевой модальности согласно указанному справочнику [45] на английском языке.

33. Наименование анатомической локализации (английское) согласно ФС анатомических локализаций – полное наименование анатомической локализации целевой области согласно указанному справочнику [114] на английском языке.

34. Внутренний код – сокращенное название целевого признака на английском языке, формируется исходя из МКБ-10, предназначен для формирования идентификатора.

35. Название нозологии – название целевой патологии/признака на русском языке, согласно коду МКБ-10.

36. Код МКБ-10 – код МКБ-10 целевой патологии.

37. Код МКБ-10 направляющего диагноза.

38. Код услуги ЕРИС – код согласно справочнику услуг ЕРИС ЕМИАС.

39. Критерии включения/ невключения пациента – критерии, на основании которых принимается решение о включении/ невключении обследуемого в НД.

Популяционные параметры

40. Претестовая вероятность патологии – частота встречаемости целевой патологии/ признака в популяции.

41. Возраст (мин., лет) – возраст самого младшего обследуемого в НД.

42. Возраст (макс., лет) – возраст самого старшего обследуемого в НД.

43. Возраст (средний, лет) – среднее значение возраста в НД.

44. Возраст (мин., лет) – медианное значение возраста в НД.

45. Пол (М) – количество обследуемых мужского пола в НД.

46. Пол (Ж) – количество обследуемых женского пола в НД.

47. Пол (не определено) – количество обследуемых, данные о поле которых отсутствуют.

48. География сбора – названия МО, в которых происходил сбор данных или географический субъект, в котором происходил сбор данных.

49. Период сбора (начало) – дата проведения самого раннего исследования в НД.

50. Период сбора (конец) – дата проведения самого позднего исследования в НД.

51. Поток – тип МО, в которых происходил сбор данных: амбулаторный, стационарный, специализированный.

52. Эпидемиологическая обстановка – состояние распространенности инфекционной болезни людей на территории сбора данных на момент сбора.

53. Источник данных – фантомные, синтетические, пациенты.

Назначение (область применения):

54. Клиническая/ практическая/ научная задача создания НД – задача в соответствии с поставленной целью, которую планируется решить с помощью НД.

55. Направление Эксперимента – направление согласно приказу порядка и условий проведения Эксперимента.

56. Ключевые слова для поиска НД (на русском) – слова или словосочетания на русском языке, наиболее точно описывающие тематику области, для которой создается НД.

57. Ключевые слова для поиска НД (на английском) – слова или словосочетания на английском языке, наиболее точно описывающие тематику области, для которой создается НД.

58. Вид тестирований – калибровочное, функциональное или самотестирование (для НД типа I, II, III);

59. Вариант – номер НД с той же спецификацией, но с другим набором исследований (первичный, вторичный и т. д.).

Параметры разметки:

60. Способы предразметки – способ предварительного отбора информации в НД, например, с использованием какого-либо алгоритма.

61. Уровень разметки – уровень, на котором происходит разметка в зависимости от используемых данных: пациент, исследование, серия, изображение.

62. Тип разметки мультилейбл – определяется в случае количества лейблов более одного.

63. Количество лейблов – количество параметров (полей, признаков), по которым производится разметка.

64. Названия лейблов – наименования параметров, по которым производилась разметка.

65. Характер разметки – бинарная, категориальная, регрессионная, текстовая, указывается для каждого лейбла.

66. Уровень детализации лейблов – уровень, на котором происходит разметка каждого параметра: исследование и/или серия и/или изображение, а также сегментация или локализация области интереса в случае разметки на изображении.

67. Количество классов – количество вариантов всех возможных для данного лейбла объектов с заданным значением метки (указывается для каждого лейбла в случае классификации данных).

68. Названия классов – названия всех классов в соответствии с п. 69.

69. Количество по классам – количество единиц НД в каждом классе.

70. Класс разметки – классификация разметки по способу верификации данных (рисунок 13).

71. Метод верификации – метод, с помощью которого

верифицировались данные (таблица 3).

72. Количество специалистов – количество специалистов, участвующих в разметке одного исследования НД.

73. Опыт специалистов – стаж работы врачей и экспертов, участвующих в разметке.

74. Временной промежуток между входными данными и данными верификации – для данных, верифицированных с помощью других методов диагностики или исследований в динамике.

75. Данные медицинской карты – используемая при верификации информация из медицинской карты пациента.

76. Критерии отнесения к классам – по каким критериям каждая единица НД относилась к тому или иному классу.

Технические параметры

77. Критерии включения/ исключения исследования в НД – критерии, по которым исследование включалось или не включалось в НД.

78. Протоколы и условия сбора данных – протоколы проведения исследований, включенных в НД.

79. Единичная запись НД: объект разметки – объект, подаваемый на вход СИИ/разметчику.

80. Единичная запись НД: результат разметки – результат разметки, получаемый от СИИ/ разметчика.

81. Формат записи НД: объект разметки – формат объекта, подаваемого на вход СИИ/ разметчику.

82. Формат записи НД: результат разметки – формат результата, получаемого от СИИ/ разметчику.

83. Количество записей НД – количество единиц НД.

84. Общий объем НД (Гб).

85. Количество уникальных источников – количество диагностических устройств, с которых собирались данные.

86. Перечень моделей и производителей – модели и производители

устройств, с которых собирались данные.

87. Анонимизация – способ анонимизации данных.

88. Комментарий к НД.

Смена версии/утилизация

89. Смена версии/утилизация – указывается, был ли сформирован НД в результате смены версии другого НД или указывается информация об утилизации.

Использование НД

90. Актуальная версия для Эксперимента – возможность использования НД в Эксперименте.

91. Количество тестирований СИИ.

92. Научное сотрудничество – данные о научных сотрудничествах, в рамках которых использовался НД.

93. Научная публикация – данные о публикациях, в рамках которых использовался НД.

94. Другое – другие данные по использованию НД.

95. Доступ для разработчиков – ссылки, по которым доступен НД.

96. Необходимость регистрации – указывается, требуется ли регистрация НД в качестве РИД.

97. Статус регистрации – на каком этапе регистрации находится НД; на момент актуализации реестра, номер и дата выдачи свидетельства РИД по готовности.

98. Формат хранения НД – формат, в котором НД находится в хранилище.

99. Место хранения файла с разметкой – ссылка на файл с разметкой.

100. Место хранения файла без разметки – ссылка на файл без разметки.

101. Хранение в архиве – ссылка на НД в архиве.