

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)

Автореферат
диссертации на соискание ученой степени
кандидата медицинских наук

Самара – 2025

Работа выполнена в Государственном бюджетном учреждении здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ «НПКЦ ДиТ ДЗМ»)

Научный руководитель:

Арзамасов Кирилл Михайлович – доктор медицинских наук, ГБУЗ «НПКЦ ДиТ ДЗМ», руководитель научного отдела медицинской информатики, радиомики и радиогеномики.

Официальные оппоненты:

Мелдо Анна Александровна – доктор медицинских наук, профессор кафедры вычислительной техники федерального государственного автономного образовательного учреждения высшего образования Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина).

Камышанская Ирина Григорьевна – доктор медицинских наук, профессор кафедры лучевой диагностики Медицинского института ФГБОУ ВО «Санкт-Петербургский государственный университет».

Ведущая организация:

Институт программных систем им. А.К.Айламазяна Российской академии наук.

Защита диссертации состоится «__» _____ 2025 года в _____ часов на заседании диссертационного совета 21.2.061.08 при федеральном государственном бюджетном образовательном учреждении высшего образования «Самарский государственный медицинский университет» Министерства здравоохранения Российской Федерации по адресу: 443079, Российская Федерация, г. Самара, ул. К. Маркса, д.165 Б.

С диссертацией можно ознакомиться в библиотеке (443001, г. Самара, ул. Арцыбушевская, д. 171) и на сайте (<http://www.samsmu.ru/scientists/science/referats>) федерального государственного бюджетного образовательного учреждения высшего образования «Самарский государственный медицинский университет» Министерства здравоохранения Российской Федерации.

Автореферат разослан «__» _____ 2025 г.

Ученый секретарь диссертационного совета,
кандидат медицинских наук, доцент

Хивинцева Елена Викторовна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Технологии искусственного интеллекта (ТИИ) находят все более широкое применение во всех сферах нашей жизни, включая здравоохранение. Начавшаяся в 2011 г. масштабная цифровизация (Гусев А.В., 2013) 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. (Указ Президента Российской Федерации От 10.10.2019 г. № 490 «О Развитии Искусственного Интеллекта в Российской Федерации», 2019), 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 (Гусев А.В., 2017). 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) (Указ Президента Российской Федерации От 10.10.2019 г. № 490 “О Развитии Искусственного Интеллекта в Российской Федерации,” 2019). ТИИ применяются практически во всех направлениях медицины: радиологии, онкологии, офтальмологии, хирургии, фармацевтике, генетике, неврологии, психиатрии и т. д. (Benjamins S. et al., 2020; Алексеева М.Г. с соавт., 2022). 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 (Benjamens S. et al., 2020).

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) (Benjamens S. et al., 2020). 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) (Васильев Ю.А. с соавт., 2023). 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 (Васильев Ю.А. с соавт., 2023). В рутинной практике ТИИ также могут способствовать повышению качества оказания медицинской помощи и снижению трудозатрат. Одним из примеров является применение ТИИ в профилактической маммографии, для которой обязательно двойное чтение с целью минимизации пропуска патологии. Многочисленные исследования показали, что СИИ демонстрируют значения критериев диагностической точности (Hickman S.E. et al., 2022; Leibig C. et al., 2022; Liu J. et al., 2022), сопоставимые с врачами-рентгенологами, и могут использоваться в качестве «второго чтения» (Арзамасов К.М. с соавт., 2023; Васильев Ю.А. с соавт., 2023). Потребность в СИИ возрастает с каждым днем, о чем говорит динамика увеличения количества так называемых ИИ-сервисов (СИИ в Эксперименте) (Васильев Ю.А. с соавт., 2022), а также их внедрение в систему обязательного медицинского страхования в Москве (код процедуры – 001601 «Описание и интерпретация данных маммографического исследования с использованием искусственного интеллекта») (Васильев Ю.А. с соавт., 2023; Тарифное соглашение на оплату медицинской помощи, оказываемой по

территориальной программе обязательного медицинского страхования города Москвы на 2023г.).

Однако, несмотря на все преимущества применения ТИИ, тем не менее имеется ряд технических, этических и правовых ограничений (Benjamens S. et al., 2020; Ngiam K.Y. et.al., 2019). Оценивая зрелость СИИ для лучевой диагностики (Васильев Ю.А. с соавт., 2023; Тыров И.А. с соавт., 2022), 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 (Указ Президента Российской Федерации От 10.10.2019 г. № 490 «О Развитии Искусственного Интеллекта в Российской Федерации», 2019):

- 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 (Ngiam K.Y., 2019). Национальная стратегия 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 (Aggarwal R. et al., 2021), 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 (Амелина Е.В. с соавт., 2022; Окунев С.В., 2020; Косарева А.А., 2023), 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 (Sounderajah V. et al., 2020; Cabitza F., 2021), 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.

Степень разработанности темы исследования

Цифровизация и внедрение информационных технологий в практическую медицину позволили сделать серьезный скачок в ее развитии. Одним из ярких примеров является внедрение в лучевую диагностику ПО, позволяющего анализировать изображения и формировать заключения. Это поддерживается и Национальной стратегией (Указ Президента Российской Федерации От 10.10.2019 г. № 490 “О Развитии Искусственного Интеллекта в Российской Федерации,” 2019). В ней же большое внимание уделено НД – основе разработки и внедрения ИИ.

Разработчики и ученые нуждаются в качественных, репрезентативных НД. С этой целью создаются специализированные библиотеки (Павлов Н.А. с соавт, 2021), принципы организации и систематизации данных (Cabitz, F. ,2021; Gebru T. et al. 2021; Homeyer A. et al., 2022), подходы к созданию НД (Павлов Н.А. с соавт, 2021), специальное ПО для разметки данных (Васильев Ю.А. с соавт. 2023, Philbrick K.A. et al., 2019).

При этом существует ряд проблем, таких как сложность, неструктурированность и неоднозначность медицинских данных, сложность и высокая стоимость разметки, этические аспекты и безопасность при работе с медицинскими данными, отсутствие унифицированных методологий создания НД, ограниченное количество данных для публичного использования, их разрозненность и низкое качество. Представляется актуальным для решения этих проблем разработать правила систематизации информации, создать единую методологию формирования НД с обоснованием необходимого для тестирования ПО на основе технологий ИИ объема с целью оптимизации ресурсов и повышения качества создаваемого НД. Описанные процессы позволят создавать качественные НД, которые в дальнейшем необходимо размещать в открытом доступе в соответствии с принципами Национальной стратегии, что, в свою очередь, требует разработки определенных правил предоставления метаинформации с учетом обеспечения безопасности персональных данных.

Цели и задачи

Цель исследования – создание методологии формирования НД 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 в виде реестра 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 оценена применимость различных методов расчета минимального объема выборки для наборов медицинских данных и разработаны принципы расчета минимального размера НД на основании критериев диагностической точности для независимой оценки СИИ.
4. Разработана новая информационная технология – 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 методик 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, маммографии, компьютерной томографии и 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., 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.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В **первой главе** представлены результаты литературного обзора научных публикаций отечественных и зарубежных авторов. Актуальность диссертации заключается в необходимости разработки методологии формирования НД, принципов их систематизации и стандартизации, а также обоснования объема выборки и баланса классов НД для тестирования СИИ в лучевой диагностике. Большинство работ, посвящено частным вопросам подготовки НД, масштабные условия Эксперимента требовали унифицированной методологии и специализированных инструментов для работы с результатами лучевой диагностики. Создание и внедрение данных методов способствует созданию качественных НД и, как следствие, развитию и внедрению ПО на основе технологий ИИ в практическую медицинскую деятельность, что является одной из приоритетных задач Национальной стратегии.

Во **второй главе** представлены материалы и методы. Диссертационная работа состоит из двух частей: аналитической и экспериментальной, и поделена на этапы в соответствии с задачами (рисунок 1). Дизайн: исследование со смешанными методами. Используются преимущественно аналитические методы исследования (анализ, синтез, индукция, дедукция), в рамках экспериментальной части по изучению объема и состава НД – статистические методы, включая анализ характеристической кривой AUC ROC.

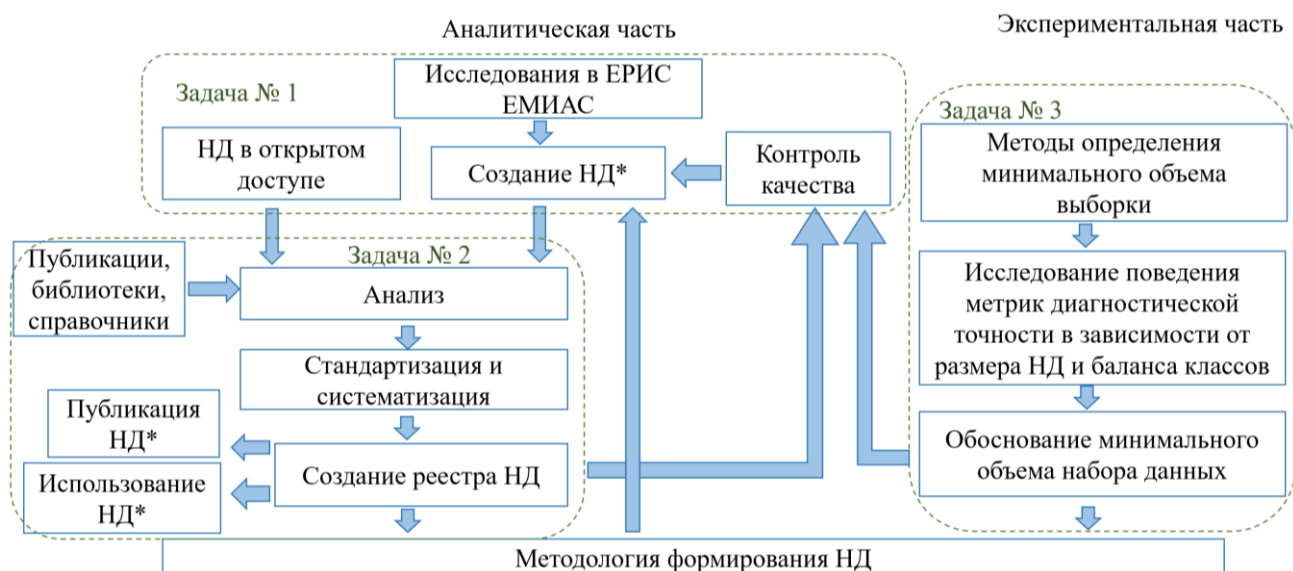


Рисунок 1 – Схема диссертационного исследования

В третьей главе представлены основные результаты. В первом параграфе проанализированы НД в открытом доступе, а также НД, процессы их создания и использования в ГБУЗ «НПКЦ ДиТ ДЗМ» на предмет оценки управляемости, надежности и устойчивости. Были сформулированы основные проблемы, выявлены их причины и поставлены задачи для реализации дальнейшей оптимизации процесса создания НД.

С целью развития исследовательской инфраструктуры и обеспечения доступа к данным была создана библиотека открытых НД *mosmed.ai*, которая требовала оптимизации процессов публикации, а также структуризации и стандартизации информации о НД для обеспечения наглядности и удобства, минимизации ошибок.

Расширение Эксперимента, внедрение новых ИИ-сервисов и появление новых направлений требовали создания большого количества разнообразных НД. Тем не менее отсутствие четких, регламентированных процессов создания и использования НД, а также специализированных инструментов управления этими процессами, на первых этапах приводило к тому, что в результате выявленных в ходе использования ошибок до 30 % исследований требовало пересмотра и до 50 % – замены.

Изучение открытых НД позволило преимущественно выявить ошибки на этапе использования, такие как большое количество пропусков патологий,

наличие исследований детей, исследований с артефактами и дефектами, дубликатов.

Таким образом, были выявлены следующие проблемы, связанные с отсутствием системы контроля качества:

- отсутствие стандартизированного алгоритма создания НД;
- отсутствие специальных 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.

Во **втором параграфе** представлены разработанный жизненный цикл, алгоритм формирования НД и принципы формирования сопроводительной документации (техническое задание, инструкция разметки). Жизненный цикл состоит из этапов инициирования, планирования, формирования, регистрации и публикации, использования, смены версии, удаления и архивации.

Также представлена классификация НД по цели создания (типы):

- I. Проведение тестирований СИИ;
- II. Самотестирование техническое (Васильев Ю. А. с соавт., 2023);
- III. Самотестирование диагностическое;
- IV. Выполнение клинических испытаний (ГОСТ Р 59921.1-2022);
- V. Выполнение технических испытаний (ГОСТ Р 59921.2-2022);
- VI. Проведение разметки текстовых протоколов с помощью программ автоматизированного анализа текстов;
- VII. Проведение научных исследований;
- VIII. Разработка СИИ;
- IX. Обучение специалистов;
- X. Для категории национальных стандартов с НД;

В ходе диссертационного исследования разработан алгоритм формирования НД лучевой диагностики, включающий в себя процессы сбора данных, их разметки, структуризации, анонимизации, формирования файлов данных и

подготовки сопроводительной документации (рисунок 2). Были разработаны методы оптимизации выполнения этапов.



Рисунок 2 – Алгоритм формирования набора данных

Описанные этапы жизненного цикла, алгоритм формирования НД и сопутствующая документация позволяют не только систематизировать и учесть все основные аспекты создания НД, но и способствуют разработке ПО для автоматизации этих процессов, что в перспективе позволит оптимизировать все процессы, ускорить формирование НД и повысить его качество.

В **третьем параграфе** представлены разработанные принципы стандартизации и систематизации информации о НД, сформированные в виде реестра и внедренные в практическую деятельность. Реестр НД – это систематизированный перечень сведений обо всех НД в организации, ведущийся уполномоченным сотрудником, с целью упорядочивания деятельности по формированию и использованию НД (Васильев Ю. А., 2023). На начало 2025 г. реестр состоит из 101 поля и 527 записей. Все поля сгруппированы по разделам, соответствующим жизненному циклу НД.

Реестр НД выполняет следующие функции:

1. Контроль качества данных: проверка параметров НД на соответствие техническому заданию и базовым диагностическим требованиям; прозрачность, надежность и воспроизводимость разработок за счет удобного доступа к данным и отслеживания изменений НД (версионность).

2. Управление: контроль сроков и порядка выполнения работ по созданию НД; распределение ресурсов при создании НД; оценка результативности использования НД; оптимизация ресурсов, повторное использование данных.

3. Доступ к данным: централизация и упорядочивание хранения НД и информации о них.

4. Автоматизация: генерация readme-файлов; заполнение библиотеки НД; проверка параметров на соответствие техническому заданию; основа платформы подготовки НД.

В рамках задач по стандартизации была также разработана классификация (таблица 1), обобщающая методы верификации, что позволило настроить удобный поиск НД по данному параметру.

Таблица 1 – Методы верификации данных

Метод верификации	Пример
Исследование другой модальности	Для верификации патологии на рентгенологическом 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

В процессе использования НД и реестра были разработаны принципы формирования названий и идентификаторов НД. В работе представлено 4 варианта наименований: предварительное название, публичное название, внутренний и внешний идентификатор. Структура разработана таким образом, чтобы из наименования можно было извлечь ключевую информацию о НД исходя из цели использования. Так, например, для внутреннего использования в названии указываются: принадлежность к проекту, год создания, тип (в соответствии с классификацией по цели создания), модальность исследования, целевая патология, анатомическая область, вариант, версия.

Использование разработанной структуры обеспечивает: удобную идентификацию и обращение к НД для всех сотрудников; единые принципы формирования названий для регистрации НД в качестве результата

интеллектуальной деятельности, формирования readme-файла, упоминания в научных публикациях; удобную идентификацию и представление НД в каталоге библиотеки; удобную систему хранения НД в архиве.

В четвертом параграфе представлены основные ошибки, выявленные в ходе процессов создания и использования НД, и пути их предотвращения и устранения, а также основные показатели, изменившиеся в ходе внедрения и совершенствования разрабатываемой методологии (таблица 2).

Таблица 2 – 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

В пятом параграфе описан эксперимент по изучению зависимости площади под характеристической кривой AUC ROC от объема выборки и баланса классов НД с бинарным исходом. В ходе работы из НД маммографических и флюорографических исследований моделировались выборки с различным объемом и балансом классов, рассчитывался показатель AUC ROC, анализировался коэффициент вариабельности. Был выявлен максимум коэффициента вариабельности для каждого изучаемого баланса классов (для доли патологии 10% – 190 исследований, для 20% и 50% – 80, для 30% – 150, для 40% – 120) и предложен алгоритм оценки AUC ROC при проведении внешних валидационных тестирований:

1. Создание НД с заданным балансом классов и объемом.
2. Проведение тестирования СИИ на полученном НД.

3. Оценка результатов тестирования с определением AUC ROC.
4. Определение доверительного интервала для AUC ROC с помощью метода бутстрэппинга.
5. Использование нижней границы доверительного интервала в качестве оценки AUC ROC при сравнении с референсным значением.

ЗАКЛЮЧЕНИЕ

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 НД – 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.

Разработанная методология положена в основу платформы автоматической подготовки НД (в рамках научно-исследовательской и опытно-конструкторской работы «Разработка платформы подготовки наборов данных лучевых диагностических исследований» № ЕГИСУ: 123031500003-8).

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».

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. Созданы 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 AUC ROC, 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 350 НД и зарегистрировано в качестве результата интеллектуальной деятельности более 40 НД 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: реестр НД рекомендуется в качестве инструмента управления и контроля качества в процессах создания и использования медицинских НД.

Руководителям медицинских организаций и разработчикам СИИ: 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; 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.

Перспективы дальнейшей разработки темы

В перспективе планируется адаптация методик и инструментов под НД лучевых исследований, обогащенных клинической информацией. Разработанная методология, включая инструменты также может быть расширена на другие направления медицинской диагностики.

Разработанная методика определения объема выборки и баланса классов в дальнейшем может быть апробирована на других направлениях медицинской диагностики, а также других критериях диагностической точности и лечь в основу методологии определения минимального размера НД для различных задач.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ

ДИССЕРТАЦИИ

1. Бобровская, Т. М. Подходы к формированию наборов данных в лучевой диагностике / Т. М. Бобровская, Ю. А. Васильев, Н. Ю. Никитин, К. М. Арзамасов // **Врач и информационные технологии**. – 2023. – № 4. – С. 14-23. – DOI: 10.25881/18110193_2023_4_14.

2. Васильев, Ю. А. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике / Ю. А. Васильев, Т. М. Бобровская, К. М. Арзамасов, С. Ф. Четвериков, А. В. Владзимирский, О. В. Омелянская, А. Е. Андрейченко, Н. А. Павлов, Л. Н. Анищенко // **Менеджер здравоохранения**. – 2023. – № 4. – С. 28-41. – DOI: 10.21045/1811-0185-2023-4-28-41.

3. Бобровская, Т. М. Объем выборки для оценки диагностической точности программного обеспечения на основе технологий искусственного интеллекта в лучевой диагностике / Т. М. Бобровская, Ю. А. Васильев, Н. Ю. Никитин, А. В. Владзимирский, О. В. Омелянская, С. Ф. Четвериков, К. М. Арзамасов // **Сибирский журнал клинической и экспериментальной медицины**. – 2024. – Т. 39, № 3. – С. 188-198. – DOI: 10.29001/2073-8552-2024-39-3-188-198.

4. Khoruzhaya, A. N. Expanded Brain CT Dataset for the Development of AI Systems for Intracranial Hemorrhage Detection and Classification / A. N. Khoruzhaya, T. M. Bobrovskaya, D. V. Kozlov, D. Kuligovskiy, V. P. Novik, K. M. Arzamasov, E. I. Kremneva // *Data*. – 2024. – Vol. 9, No. 2. – P. 30. – DOI: 10.3390/data9020030.

5. Arzamasov, K. An International Non-Inferiority Study for the Benchmarking of AI for Routine Radiology Cases: Chest X-ray, Fluorography and Mammography / K. Arzamasov, Y. Vasilev, A. Vladzymyrskyy, O. Omelyanskaya, I. Shulkin, D. Kozikhina, I. Goncharova, P. Gelezhe, Y. Kirpichev, T. Bobrovskaya, A. Andreychenko // *Healthcare*. – 2023. – Vol. 11, No. 12. – P. 1684. – DOI: 10.3390/healthcare11121684.

6. Четвериков, С. Ф. Подходы к формированию выборки для контроля качества работы искусственного интеллекта в медико-биологических исследованиях / С. Ф. Четвериков, К. М. Арзамасов, А. Е. Андрейченко, В. П. Новик, Т. М. Бобровская, А. В. Владзимирский // *Современные технологии в медицине*. – 2023. – Т. 15, № 2. – С. 19. – DOI: 10.17691/stm2023.15.2.02.

7. Арзамасов, К. М. Критерии применимости компьютерного зрения для профилактических исследований на примере рентгенографии и флюорографии органов грудной клетки / К. М. Арзамасов, С. С. Семенов, Д. Ю. Кокина, Т. М. Бобровская, Н. А. Павлов, Ю. С. Кирпичев, А. Е. Андрейченко, А. В. Владзимирский // *Медицинская физика*. – 2022. – № 4. – С. 56-63. – DOI: 10.52775/1810-200X-2022-96-4-56-63.

8. Арзамасов, К. М. Применение компьютерного зрения для профилактических исследований на примере маммографии // К. М. Арзамасов, Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская, Т. М. Бобровская, С. С. Семенов, С. Ф. Четвериков, Ю. С. Кирпичев., Н. А. Павлов, А. Е. Андрейченко // *Профилактическая медицина*. – 2023. – Т. 26, № 6. – С. 117-123. – DOI: 10.17116/profmed202326061117.

9. Vasilev, Y. Clinical application of radiological AI for pulmonary nodule evaluation: Replicability and susceptibility to the population shift caused by the COVID-19 pandemic / Y. Vasilev, A. Vladzimirsky, K. Arzamasov, O. Omelyanskaya, I. Shulkin, D. Kozikhina, I. Goncharova, R. Reshetnikov, S. Chetverikov, I. Blokhin, T. Bobrovskaya, A. Andreychenko // International Journal of Medical Informatics. – 2023. – No. 178. – 10 p. – DOI: 10.1016/j.ijmedinf.2023.105190.

10. Arzamasov, K. Independent evaluation of the accuracy of 5 artificial intelligence software for detecting lung nodules on chest X-rays / K. Arzamasov, Y. Vasilev, M. Zelenova, L. Pestrenin, Y. Busygina, T. Bobrovskaya, S. Chetverikov, D. Shikhmuradov, A. Pankratov, Y. Kirpichev, V. Sinitsyn, I. Son, O. Omelyanskaya // Quantitative Imaging in Medicine and Surgery. – 2024. – Vol. 14, No. 8. – P. 5288-5303. – DOI: 10.21037/qims-24-160.

Авторские свидетельства на государственную регистрацию баз данных:

1. № 2023621046 MosMedData: КТ с признаками аневризмы брюшного отдела аорты;
2. № 2023621073 MosMedData: РГ с признаками перелома костей плечевого сустава;
3. № 2023624880 MosMedData: ММГ с наличием и отсутствием признаков злокачественных новообразований молочной железы, обогащенный клинической информацией;
4. № 2023621043 MosMedData: КТ с признаками лимфаденопатии;
5. № 2023621044 MosMedData: КТ с признаками образований почек;
6. № 2023621 MosMedData: КТ брюшной полости с признаками остеопороза позвоночника;
7. № 2023621047 MosMedData: КТ с признаками образований печени;
8. № 2023621056 MosMedData: КТ с признаками коронавирусной инфекции (COVID-19);
9. № 2023621074 MosMedData: КТ с признаками мочекаменной болезни
10. № 2023621079 MosMedData: РГ с признаками перелома костей лучезапястного сустава;

11. № 2023621080 MosMedData: КТ с признаками гидроторакса;
12. № 2023621083 MosMedData: КТ с признаками эпикардального жира;
13. № 2023621091 MosMedData: КТ с признаками образований надпочечников;
14. № 2023621102 MosMedData: цРГ с признаками артроза коленного сустава;
15. № 2023621103 MosMedData: КТ с признаками рака легкого;
16. № 2023621170 MosMedData: КТ с признаками бронхоэктатической болезни;
17. № 2023621171 MosMedData: КТ с признаками остеопороза позвоночника;
18. № 2023621204 MosMedData: РГ_ФЛГ с признаками патологий ОГК;
19. № 2023621207 MosMedData: РГ с признаками артроза тазобедренного сустава;
20. № 2023621208 MosMedData: КТ с признаками эмфиземы;
21. № 2023621221 MosMedData: РГ с признаками компрессионного перелома тел позвонков;
22. № 2023621223 MosMedData: МРТ с признаками патологий пояснично-крестцового отдела позвоночника;
23. № 2023621225 MosMedData: цРГ с признаками переломов трубчатых костей;
24. № 2023621226 MosMedData: РГ с признаками синусита;
25. № 2023621227 MosMedData: РГ с признаками остеохондроза;
26. № 2023621228 MosMedData: КТ с признаками саркоидоза легких;
27. № 2023621229 MosMedData: РГ с признаками спондилолистеза;
28. № 2023621233 MosMedData: МРТ с признаками ЗНО головного мозга;
29. № 2023621234 MosMedData: РГ с признаками коронавирусной инфекции (COVID-19);
30. № 2023621235 MosMedData: КТ с признаками коронарного кальция;
31. № 2023621351 MosMedData: КТ с признаками туберкулеза легких;
32. № 2023621926 MosMedData: МРТ с признаками интракраниальных образований;
33. № 2023622421 MosMedData: РГ ОГК с наличием и отсутствием легочных узлов;
34. № 2023621253 MosMedData: НДКТ с признаками рака легкого;

35. № 2024621476 MosMedData: текстовые протоколы КТ грудной полости с наличием и отсутствием признаков расширения легочного ствола, аневризмы аорты, эмфиземы, гидроторакса, компрессионного перелома тел позвонков;
36. № 2023621240 MosMedData: КТ с признаками паракардиального жира;
37. № 2023621246 MosMedData: цРГ с признаками плоскостопия;
38. № 2023621247 MosMedData: ФЛГ с признаками патологий ОГК;
39. № 2023621262 MosMedData: КТ с признаками аневризмы аорты;
40. № 2023621293 MosMedData: РГ с признаками перелома костей тазобедренного сустава;
41. № 2023621254 MosMedData: КТ с отсутствием и наличием признаков расширения легочного ствола;
42. № 2023621305 MosMedData: РГ с признаками артефактов и дефектов укладки;

СПИСОК СОКРАЩЕНИЙ

ГБУЗ «НПКЦ ДиТ ДЗМ»	государственное бюджетное учреждение здравоохранения «Научно-практический клинический центр диагностики и телемедицины департамента здравоохранения города Москвы»
ИИ	искусственный интеллект
НД	набор данных
СИИ	система искусственного интеллекта
ТИИ	технологии искусственного интеллекта
ROC AUC	(receiver operating characteristic area under the curve) площадь под кривой рабочей характеристики приемника

Подписано в печать: 23.10.2025

Объем: 1 п.л.

Тираж 100 экз. Заказ №253

Отпечатано в типографии «КопиЦентр»

443010, г. Самара

Ул. Ульяновская, 13

89023362957, econom-print.ru