ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

ГБУЗ «НАУЧНО-ПРАКТИЧЕСКИЙ КЛИНИЧЕСКИЙ ЦЕНТР ДИАГНОСТИКИ И ТЕЛЕМЕДИЦИНСКИХ ТЕХНОЛОГИЙ ДЕПАРТАМЕНТА ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ»

ПОДГОТОВКА НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ И ТЕСТИРОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Учебно-методическое пособие

Москва
2023

ЦЕНТР ДИАГНОСТИКИ
И ТЕЛЕМЕДИЦИНЫ

ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

СОГЛАСОВАНО

Главный внештатный специалист
Департамента здравоохранения
города Москвы по лучевой и
инструментальной диагностике

_____Ю.А. Васильев

«15» мая 2023 г.

РЕКОМЕНДОВАНО

Экспертным советом по науке
Департамента здравоохранения
города Москвы № 5

«05» _____ 2023 г.

ПОДГОТОВКА НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ И ТЕСТИРОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Учебно-методическое пособие № 22

Москва
2023

УДК 004.896+073.75
ББК 53.6
П 44

Ю. А. Васильев, К. М. Арзамасов, А. В. Владзимирский, О. В. Омелянская, Т. М. Бобровская, Д. Е. Шарова, Н. Ю. Никитин, М. Р. Коденко

П 44 Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта: учебно-методическое пособие / Ю. А. Васильев, К. М. Арзамасов, А. В. Владзимирский [и др.]. – М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. – 108 с.

Рецензенты:

Руднов Н. В. – доктор медицинских наук, профессор, заместитель директора ФГБУ «РНЦРР» Минздрава России по научной работе

Лебедев Г. С. – доктор технических наук, профессор, директор Института цифровой медицины, заведующий кафедрой информационных и интернет-технологий Института цифровой медицины ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет)

Пособие содержит учебную информацию, дополняющую и частично заменяющую учебник, а также способствующую рациональному достижению целей обучения дисциплинам «Общественное здоровье и здравоохранение», «Лучевая диагностика» и «Медицинская информатика» по специальностям 31.05.01 Лечебное дело, 31.05.02 Педиатрия, 31.05.03 Стоматология, 31.08.09 Рентгенология, 30.05.03 Медицинская кибернетика и 30.05.02 Медицинская биофизика. Кроме того, материалы учебно-методического пособия будут полезны обучающимся по образовательной программе бакалавриата, направление подготовки – 09.03.04 Программная инженерия, специальности – 09.04.02 Информационные системы и технологии, 06.004 Специалист по тестированию в области информационных технологий. Представленная информация направлена на приобретение и расширение обучаемыми лицами необходимых компетенций, получение знаний, умений и навыков в области принципов и методологий подготовки набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта.

В пособии излагаются теоретические основы и практические методы создания релевантных, репрезентативных, корректно размеченных наборов данных, используемых для разработки, обученияи валидации программного обеспечения на основе технологии искусственного интеллекта. Пособие может быть использовано разработчиками программного обеспечения на основе технологии искусственного интеллекта при создании и валидации наборов данных.

Предназначено для студентов, магистров, слушателей программ дополнительного профессионального образования.

УДК 004.896+073.75
ББК 53.6

Рекомендовано ученым советом государственного бюджетного учреждения здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» в качестве учебно-методического пособия для студентов, изучающих учебный курс «Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта»

Учебно-методическое пособие разработано в ходе выполнения научно-практического проекта в сфере медицины (№ ЕГИСУ: 122112400040-1) «Эталонные наборы данных для устойчивого развития технологий искусственного интеллекта в медицинской диагностике с целью минимизации долгосрочных последствий пандемии коронавирусной инфекции для здоровья населения города Москвы»

ОГЛАВЛЕНИЕ

ПРИНЯТЫЕ СОКРАЩЕНИЯ И АББРЕВИАТУРЫ

БДТ – базовые диагностические требования
БФТ – базовые функциональные требования
ДЗМ – Департамент здравоохранения города Москва
ЕМИАС – Единая медицинская информационно-аналитическая система
ЕРИС – Единый радиологический информационный сервис
ЗНО – злокачественное новообразование
КТ – компьютерная томография
МИС – медицинская информационная система
МК – медицинская карта
МКБ – Международная классификация болезней
ММГ – маммография
МО – медицинская организация
МРТ – магнитно-резонансная томография
НД – набор данных
ОК – общекультурные компетенции
ОС – операционная система
ОПК – общепрофессиональные компетенции
ПК – персональный компьютер
ПО – программное обеспечение
РГ ОГК – рентгенография органов грудной клетки
РМЖ – рак молочной железы
ТЗ – техническое задание
ТИИ – технологи искусственного интеллекта
УЗИ – ультразвуковое исследование
УИД – уникальный идентификатор
Ф.И.О. – фамилия, имя, отчество
ФИПС – Федеральный институт промышленной собственности
ФС – федеральный справочник
ЭКГ – электрокардиография
ЭНМГ – электронейромиография
ЭЭГ – электроэнцефалография
DICOM – Digital Imaging and Communications in Medicine (медицинский отраслевой стандарт создания, хранения, передачи и визуализации цифровых медицинских изображений и документов обследованных пациентов)

ВВЕДЕНИЕ

Цель данного учебно-методического пособия – приобретение и расширение обучаемыми лицами необходимых компетенций, получение знаний, умений и навыков в области принципов и методологий подготовки набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта.

Задачи:

– изучение общетеоретических вопросов, терминологии, значения в системе здравоохранения технологий искусственного интеллекта и необходимых для их развития наборов данных;

– изучение этапов жизненного цикла набора данных в сфере здравоохранения;

– изучение алгоритма создания набора данных;

– изучение мер по профилактике дефектов и ошибок при создании наборов данных;

– обеспечение уровня компетенций и навыков в соответствии с требованиями профессионального стандарта «Специалист в области организации здравоохранения и общественного здоровья»[1];

– обеспечение уровня компетенций и навыков в соответствии с требованиями профессионального стандарта «Врач-рентгенолог»[2];

– обеспечение уровня компетенций и навыков в соответствии с требованиями профессионального стандарта «Специалист по тестированию в области информационных технологий»[3].

Требования к входным знаниям, компетенциям и умениям для проведения занятий: теоретические знания и практические навыки в соответствии с федеральными государственными образовательными стандартами высшего образования по специальностям 31.05.01 Лечебное дело, 31.05.02 Педиатрия, 31.05.03 Стоматология, 31.08.09 Рентгенология, 30.05.03 Медицинская кибернетика и 30.05.02 Медицинская биофизика, а также дисциплинам образовательной программы бакалавриата по направлению подготовки 09.03.04 Программная

---

[1] Приказ Министерства труда и социальной защиты Российской Федерации от 07.11.2017 № 768н «Об утверждении профессионального стандарта „Специалист в области организации здравоохранения и общественного здоровья"».

[2] Приказ Министерства труда и социальной защиты Российской Федерации от 19.03.2019 № 160н «Об утверждении профессионального стандарта „Врач-рентгенолог"».

[3] Приказ Министерства труда и социальной защиты Российской Федерации от 02.08.2021 № 531н «Об утверждении профессионального стандарта „Специалист по тестированию в области информационных технологий"».

инженерия, специальности 09.04.02 Информационные системы и технологии и 06.004 Специалист по тестированию в области информационных технологий.

Изучение пособия направлено на дальнейшее формирование у обучающихся следующих компетенций:

I. По специальностям 31.05.01 Лечебное дело, 31.05.02 Педиатрия, 31.05.03 Стоматология, 31.08.09 Рентгенология (дисциплина «Общественное здоровье и здравоохранение»):

1. Общекультурных:

– способность к абстрактному мышлению, анализу, синтезу (ОК-1);

– готовность к саморазвитию, самореализации, самообразованию, использованию творческого потенциала (ОК-5).

2. Общепрофессиональных:

– готовность решать стандартные задачи профессиональной деятельности с использованием информационных, библиографических ресурсов, медико-биологической терминологии, информационно-коммуникационных технологий и учетом основных требований информационной безопасности (ОПК-1);

– способность и готовность анализировать результаты собственной деятельности для предотвращения профессиональных ошибок (ОПК-5).

3. Профессиональных:

– способность к применению основных принципов организации и управления в сфере охраны здоровья граждан, в медицинских организациях и их структурных подразделениях (ПК-17);

– готовность к участию во внедрении новых методов и методик, направленных на охрану здоровья граждан (ПК-22).

4. Дополнительной:

– способность организовывать оказание разных видов медицинской помощи с применением допущенных к обращению медицинских изделий на основе технологий искусственного интеллекта;

– способность понимать принципы работы современных информационных технологий, технологий искусственного интеллекта и использовать их для решения задач профессиональной деятельности.

II. По специальности 09.04.02 Информационные системы и технологии:

1. Общекультурных:

– способность совершенствовать и развивать свой интеллектуальный и общекультурный уровень (ОК-1);

– способность к самостоятельному обучению новым методам исследования, к изменению научного и научно-производственного профиля своей профессиональной деятельности (ОК-2);

– использование на практике умений и навыков в организации исследовательских и проектных работ, в управлении коллективом (ОК-4);

– способность к профессиональной эксплуатации современного оборудования и приборов (ОК-7).

2. Общепрофессиональных и профессиональных:

– способность воспринимать математические, естественнонаучные, социально-экономические и профессиональные знания, умение самостоятельно приобретать, развивать и применять их для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте (ОПК-1);

– владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях (ОПК-5);

– умение разрабатывать стратегии проектирования, определять цели проектирования, критерии эффективности, ограничения применимости (ПК-1);

– умение проводить разработку и исследование теоретических и экспериментальных моделей объектов профессиональной деятельности в области медицины (ПК-8).

III. По специальности 09.03.04 Программная инженерия:

1. Универсальных:

– способность осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач (УК-1).

2. Общепрофессиональных:

– способность применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности (ОПК-1);

– способность использовать современные информационные технологии и программные средства, в том числе отечественного производства, при решении задач профессиональной деятельности (ОПК-2);

– способность осуществлять поиск, хранение, обработку и анализ информации из различных источников и баз данных, представлять ее в требуемом формате с использованием информационных, компьютерных и сетевых технологий (ОПК-8).

IV. По специальности 06.004 Специалист по тестированию в области информационных технологий:

1. Общекультурных:

– способность совершенствовать и развивать свой интеллектуальный и общекультурный уровень (ОК-1);

– способность к самостоятельному обучению новым методам исследований, к изменению научного и научно-производственного профиля своей профессиональной деятельности (ОК-2);

– использование на практике умений и навыков в организации исследовательских и проектных работ, управление коллективом (ОК-4);

– способность к профессиональной эксплуатации современного оборудования и приборов (ОК-7).

2. Общепрофессиональных и профессиональных:

– способность применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности (ОПК-1);

– владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях (ОПК-5).

V. По специальности 30.05.03 Медицинская кибернетика и 30.05.02 Медицинская биофизика:

1. Общекультурных:

– способность совершенствовать и развивать свой интеллектуальный и общекультурный уровень (ОК-1);

– способность к самостоятельному обучению новым методам исследований, к изменению научного и научно-производственного профиля своей профессиональной деятельности (ОК-2);

– использование на практике умений и навыков в организации исследовательских и проектных работ, управление коллективом (ОК-4);

– способность к профессиональной эксплуатации современного оборудования и приборов (ОК-7).

2. Общепрофессиональных и профессиональных:

– способность применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности (ОПК-1);

– владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе в глобальных компьютерных сетях (ОПК-5).

В результате изучения материала обучаемый должен знать:

– основную терминологию, базовые принципы юридического регулирования, цели и задачи создания и эксплуатации наборов данных в здравоохранении;

– принципы стандартизации процессов создания и эксплуатации наборов данных в здравоохранении;

– принципы классификации, основные требования к структуре, составу, описанию наборов данных;

– подходы к постановке клинической задачи, решаемой с применением конкретного набора данных;

уметь:

– организовывать процесс подготовки набора данных для сферы здравоохранения;

– организовывать процессы контроля и непрерывного повышения качества при подготовке наборов данных;

– обеспечивать защиту персональных данных;

владеть:

– навыками создания технического задания на набор данных;

– отдельными навыками разметки разных типов биомедицинских данных;

– навыками создания описания набора данных для здравоохранения.

Изучение материала пособия рассчитано на 6 академических часов самостоятельной работы, для его успешного освоения рекомендуется использовать открытые библиотеки наборов данных в сфере здравоохранения: https://mosmed.ai/datasets/; https://ai2.rt-eu.ru/. В целях проверки усвоения информации предусмотрены вопросы для самоконтроля. Для повышения уровня эрудированности и вовлеченности обучаемых в изучение учебного курса опционально рекомендуется подготовка рефератов и докладов-презентаций.

Коллектив авторов выражает благодарность за помощь в подготовке учебно-методического пособия В. П. Новику, Е. Ф. Савкиной, Д. В. Козлову, У. А. Сахащик, Ю. С. Бусыгиной, Е. Г. Бахтеевой.

ОБЩИЕ ПОЛОЖЕНИЯ

В последнее время стали популярными такие слова, как искусственный интеллект, машинное обучение, большие данные (big data). Эти термины входят в повседневное употребление и уже встречаются не только в узконаправленных специализированных областях. Не стала исключением и сфера здравоохранения: автоматизированные системы диагностики, системы распознавания медицинских записей и естественного языка, системы анализа и предсказания событий, автоматической классификации и сверки информации, чат-боты поддержки пациентов, электронная медицинская карта и многое другое – результаты масштабной цифровизации в данной сфере[4,5]. Столь мощный прогресс цифровых технологий в Российской Федерации поддерживается Национальной стратегией развития искусственного интеллекта на период до 2030 года [1].

Искусственный интеллект (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. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений [1].

Технологии искусственного интеллекта (ТИИ) – технологии, основанные на использовании искусственного интеллекта, включая компьютерное зрение, обработку естественного языка, распознавание и синтез речи, интеллектуальную поддержку принятия решений и перспективные методы искусственного интеллекта [1].

В соответствии с Национальной стратегией использование технологий искусственного интеллекта **в социальной сфере** способствует созданию условий для улучшения уровня жизни населения, в том числе за счет **повышения качества услуг в сфере здравоохранения**, включая профилактические обследования, диагностику, основанную на анализе изображений, прогнозирование возникновения и развития заболеваний, подбор оптимальных дозировок лекарственных препаратов, сокращение угроз пандемий, автоматизацию и точность хирургических вмешательств.

---

[4] Гусев А. В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения // Врач и информационные технологии. 2017. №3. С. 92–105. URL: https://www.idmz.ru/jurnali/vrach-i-informatsionnye-tekhnologii/2017/3/perspektivy-neironnykh-setei-i-glubokogo-mashinnogo-obucheniia-v-sozdanii-reshenii-dlia-zdravookhraneniia.

[5] Гусев А. В., Добриднюк С. Л. Искусственный интеллект в медицине и здравоохранении // Информационное общество. 2017. № 4–5. С. 78–93.

Основные факторы развития ТИИ – это увеличение объема доступных данных, в том числе данных, прошедших разметку и структурирование, а также постоянное развитие информационно-телекоммуникационной инфраструктуры для обеспечения доступа к наборам таких данных.

С развитием медицины, повышением ее доступности и повсеместного внедрения цифровых технологий в медицинскую практику[6] отмечается высокий рост количества медицинских данных: клинических, лабораторных и инструментальных[7]. Данные – представление информации в формализованном виде, пригодном для передачи, интерпретации и обработки [2].

Большой объем данных способствует оптимальной организации интересующей сферы (в частности, здравоохранения) для достижения наилучших результатов работы. Данные могут быть использованы для прогнозирования текущих тенденций определенных параметров и будущих событий. В последние годы в медицинской практике активно внедряются электронные медицинские карты и медицинские информационные системы, что приводит к необходимости стандартизации медицинской информации.

Например, результаты лабораторных (патоморфологические исследования, клинические анализы, генетические исследования и т. д.), лучевых (КТ, МРТ, ММГ, УЗИ, рентгенография и т.д.) и сигнальных (ЭКГ, ЭЭГ, ЭНМГ и т.д.) исследований максимально стандартизованы и оцифрованы, что способствует росту количества данных по этим направлениям, инструментам для их обработки (программное обеспечение предназначенное для обработки медицинских данных), передаче и хранению, и, как следствие, развитию ТИИ в этой области[8].

Внедрение ТИИ в сферу здравоохранения позволяет повысить качество предоставляемых услуг [1], а также снизить нагрузку на врачей. Например, при скрининге рака молочной железы требуется «двойное чтение» результатов маммографических исследований, т.е. каждое исследование должно быть просмотрено двумя специалистами.

Однако многочисленные исследования[9] показывают, что одно чтение можно доверить ПО на основе ТИИ, при этом качество скрининга не ухудшается[10]. Другой пример успешного применения ПО на основе ТИИ – пандемия COVID-19: в условиях острой нехватки медицинского персонала применение ТИИ

---

[6] Соболева С. У., Голиков В. В., Тажибов А. А. Информационные технологии в здравоохранении: особенности отраслевого применения // E-Management. State University of Management, 2021. Т. 4, № 2. С. 37–43.

[7] Dash S., Shakyawar S. K., Sharma M., et al. Big data in healthcare: management, analysis and future prospects // J Big Data. SpringerOpen. 2019. Vol. 6, № 1. P. 1–25.

[8] Shakhabov I. V., Melnikov Yu. Yu., Smyshlyaev A. V. Development of digital technologies in healthcare during the COVID-19 pandemic // Scientific Review. Medical Sciences. 2020. №6. P. 66–71.

[9] Henriksen E. L. Carlsen F., Vejborg I. M., et al. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review // Acta radiol. 2019. Vol. 60, № 1. P. 13–18.

[10] Lauritzen A. D., Rodríguez-Ruiz A., von Euler-Chelpin M. C. et al. An Artificial Intelligence–based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload // Radiology. 2022. Vol. 304, № 1. P. 41–49.

позволило уменьшить время обработки заключения КТ[11], а также осуществить сортировку исследований, благодаря чему исследования пациентов в более тяжелом состоянии обрабатывались в первую очередь [3].

Однако для успешного применения ТИИ необходимо создание релевантных, репрезентативных, корректно размеченных наборов данных (НД).

НД используются не только для разработки и обучения ПО на основе ТИИ, но и их валидации, т.е. проверки качества работы ПО. Благодаря Национальной стратегии развития искусственного интеллекта в Российской Федерации стало возможным активное создание и внедрение в повседневную практику таких НД, а также инструментов их хранения, администрирования и использования.

На первый взгляд может показаться, что создание НД – несложный процесс: ведь ежедневно генерируются терабайты данных медицинской информации, а применение МИС позволяет их хранить, передавать и использовать (например, данные лучевой диагностики медицинских организаций ДЗМ хранятся в Едином радиологическом информационном сервисе – ЕРИС ЕМИАС). Тем не менее процесс создания НД (не стоит забывать о том, что они должны быть релевантными, репрезентативными и корректно размеченными) – очень сложный, имеет множество важных аспектов и вовлекает в себя большое количество специалистов, как медицинских (врачи, лаборанты), так и технических (инженеры, разработчики, аналитики и т.д.), а также смежных направлений (биофизики, кибернетики, биоинформатики).

Кроме того, недостаточно просто создать НД, необходимо уделить внимание инфраструктуре и инструментам хранения, использования и управления, таким, например, как библиотеки и реестры. Их основными задачами являются аннотация, интеграция и представление НД для контроля качества, удобного и повсеместного использования, в том числе для ПО на основе ТИИ.

Методологии создания наборов данных для сферы здравоохранения продолжают формироваться и в настоящее время, прежде всего – на основе масштабных научных исследований. Так, в основу настоящего учебно-методического пособия положены результаты «Эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» (mosmed.ai) – крупнейшего в мире проспективного многоцентрового клинического исследования технологий искусственного интеллекта [3].

---

[11] Морозов С. П., Гаврилов А. В., Архипов И. В. [и др.]. Влияние технологий искусственного интеллекта на длительность описаний результатов компьютерной томографии пациентов с COVID-19 в стационарном звене здравоохранения // Профилактическая медицина. 2022. Т. 25, № 1. С. 14–20.

ГЛАВА 1.
НАБОРЫ ДАННЫХ И ПРИНЦИПЫ ИХ КЛАССИФИКАЦИИ

1.1. Основные понятия

Медицинские данные подразделяются на несколько подмножеств, каждое из которых является важным компонентом в обучении, оценке качества ПО на основе ТИИ и используется для других прикладных и фундаментальных задач в сфере искусственного интеллекта для здравоохранения. Каждый компонент (подмножество, набор) данных направлен на решение определенной задачи.

Набор данных (НД) – это совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства Российской Федерации об информации, информационных технологиях и о защите информации, и необходимых для разработки программного обеспечения на основе искусственного интеллекта [1].

Разметка данных – 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 методов машинного обучения [1].

В процессе создания, хранения и использования НД необходимо руководствоваться следующими **нормативно-правовыми актами, межгосударственными и национальными стандартами:**

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

– ГОСТ 34.602-2020. Информационные технологии. Комплекс стандартов на автоматизированные системы;

– ГОСТ 19.201-78. Единая система программной документации. Техническое задание. Требования к содержанию и оформлению;

– ГОСТ 19.101-77. Единая система программной документации. Виды программ и программных документов;

– ГОСТ Р 59921.1-7-2022. Системы искусственного интеллекта в клинической медицине. Алгоритмы анализа медицинских изображений;

– ГОСТ Р 8.736-2011. Государственная система обеспечения единства измерений. Измерения прямые многократные. Методы обработки результатов измерений. Основные положения;

– Федеральный закон «Об информации, информационных технологиях и о защите информации» от 27.07.2006 № 149-ФЗ.

Для обучения, внутренней и внешней валидации, клинико-технических и клинических испытаний технологий искусственного интеллекта применяют **эталонные наборы** данных, под которыми понимают упорядоченную совокупность:

– результатов диагностических исследований одной или нескольких модальностей и/или однотипных медицинских документов;

– сведений о наличии, характере и локализации и т.д. целевых признаков; для текстовых документов – библиотеки ключевых слов, словосочетаний и их критичных сочетаний;

– сведений о верификации (опционально).

Информация о наличии, характере, локализации и т.д. целевых признаков (в том числе в соответствии с Международной классификацией болезней – МКБ) может быть подтверждена объективно – в таком случае набор данных именуется ***верифицированным.***

Размер набора данных (математически – размер выборки) и баланс классов определяются исходя из целей и задач проводимого исследования и требований технического задания на проведение исследований, а также с учетом требований заказчика.

Эталонный набор данных должен быть проверен профильной медицинской научно-исследовательской организацией на предмет полноты и качества содержащейся в нем информации. Рекомендуется при проведении клинических испытаний применять эталонные наборы данных, имеющие государственную регистрацию в качестве базы данных.

Эталонный набор данных для клинических испытаний должен содержать такие сведения (описательного характера) [4]:

– номер свидетельства о государственной регистрации базы данных (рекомендательно);

– характеристика популяции (гендерно-возрастные показатели, этнический состав, регионы проживания и т.д.);

– сведения о медицинских организациях, послуживших источниками для формирования набора данных;

– характеристика исследований: анатомическая область(и), модальность, проекции;

– целевой признак;

– общее количество клинических случаев, исследований, изображений, документов и их распределение по диагностическим группам (в т.ч. «норма»:«патология»);

– сведения о верификации.

Требования к эталонному набору данных [4]:

1. Структура набора данных должна соответствовать поставленной цели его формирования (решаемой клинической задаче).

2. Планируемый размер эталонного набора данных должен быть обоснован в протоколе исследования, исходя из статистических соображений и желаемой точности оценки основных метрик.

3. Разметка должна быть проведена с использованием стандартизированной терминологии – т.н. тезауруса (кодированной библиотеки типовых формулировок, соответствующих нормативно-правовой документации, клиническим рекомендациям или рекомендациям профессиональных врачебных ассоциаций).

4. Подготовка и разметка должны быть проведены техническими и медицинскими специалистами, имеющими соответствующие навыки и компетенции.

Наборы данных для обучения и тестирования алгоритмов искусственного интеллекта можно классифицировать различными способами. Например, выделяют наборы со структурированными, частично структурированными и неструктурированными данными; либо разделяют их по источникам формирования, условиям использования, типам биомедицинских и клинических данных, по временным характеристикам, файловой структуре, наконец, по видам задач, для решения которых наборы сформированы и т.д.

Рекомендуется использовать две классификации: по диагностической ценности (подробнее см. параграф 1.2 «Классификация разметки и наборов данных») и по целевому назначению (подробнее см. параграф 3.1 «Этап инициирования создания набора данных»).

Контрольные вопросы

1. Дайте определение понятию «Набор данных».
2. Дайте определение понятию «Разметка данных».
3. Перечислите нормативно-правовые акты, регулирующие создание набора данных.
4. Что такое эталонный набор данных?
5. Перечислите основные требования к эталонному набору данных.

1.2. Классификация разметки и наборов данных

Под разметкой в контексте классификации медицинских наборов данных понимается установка категориального или визуального признака в данных, выполненная медицинским персоналом и/или врачом-экспертом.

Класс разметки варьируется в зависимости от задачи, поставленной ПО на основе ТИИ, и основывается на методах верификации данных. В таблице 1 представлены принципы классификации методов верификации, разработанные на основе собственного опыта, а также рекомендаций Управления по санитарному надзору за качеством пищевых продуктов и медикаментов (Food and Drug

Administration, FDA [5]). Под верификацией понимают проверку данных на достоверность, правильность и точность. На рисунке 1 изображены методы верификации данных по возрастанию их ценности.

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 рака предстательной железы
Исследование той же модальности в динамике	Для верификации перелома позвонков на компьютерной томографии: признаки перелома позвонков 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

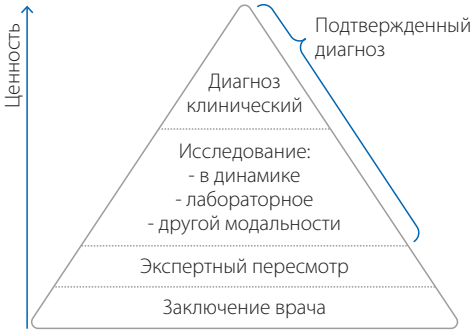


Рисунок 1 – Диаграмма методов верификации НД

Наименьшей ценностью обладает верификация по заключению врача, т.е. вывод о наличии или отсутствии патологии делается на основании заключения врача, описывавшего исследование. Как правило, такой способ разметки используется на первых этапах отбора данных и может быть осуществлен с помощью алгоритмов автоматического анализа текстовых протоколов, например, MedLabel[12]. Следующим по ценности методом верификации является экспертный пересмотр: слепой анализ исследований врачами-экспертами с достижением

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

заданного уровня согласованности их решений (подробно описан в подпараграфе 3.3.2 «Разметка данных»). Следующие две группы методов являются наиболее достоверными и их можно условно назвать «подтвержденный диагноз»: исследование той же модальности в динамике, исследование другой модальности, лабораторное исследование, которые в совокупности с остальными данными медицинской карты дают клинический диагноз. Стоит отметить, что для верификации каждой патологии существует свой метод «золотого стандарта», который позволяет подтвердить диагноз.

На рисунке 2 представлена классификация видов разметки на примере рака молочной железы (РМЖ) с учетом ценности разметки.

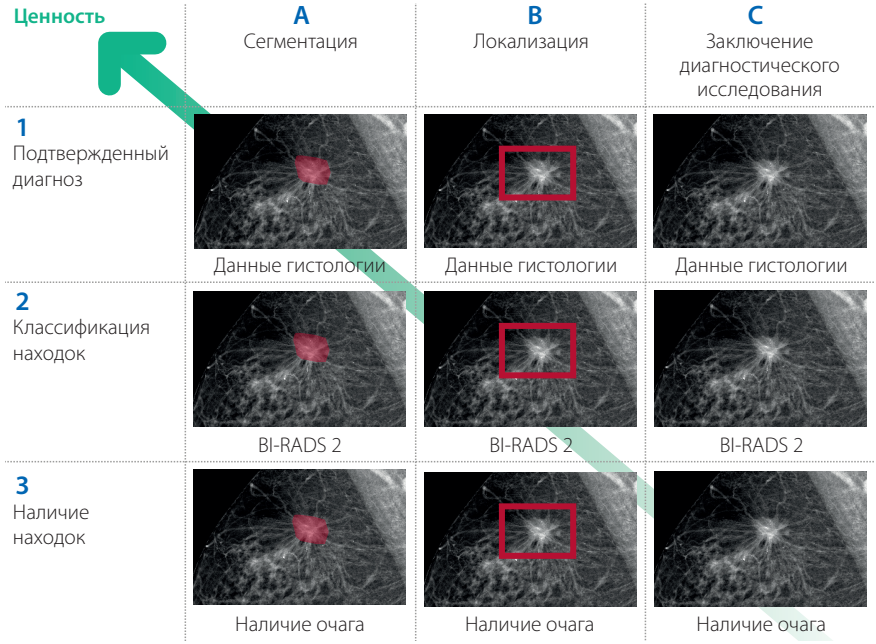


Рисунок 2 – Классификация видов разметки в медицинской диагностике по диагностической ценности

В наиболее общем виде разметка данных может проводиться на основании:

A. Информации об имеющейся целевой патологической находке, представленной на изображении в виде пиксельной маски (оконтуренной области изображения). Дополнительно может содержаться в метаданных (аннотации).

B. Информации об имеющейся целевой патологической находке, представленной в виде координат. Может помещаться в метаданных (в аннотации,

в сводном табличном сопроводительном файле) и/или присутствовать на изображении в виде отметки области расположения простой геометрической фигурой.

C. Информации о наличии/отсутствии целевой патологической находки, содержащейся в метаданных (то есть в аннотации – сопроводительных файлах) и отсутствующей на изображении.

Классификация A, B, C для уровня 3 (обнаружение находки) предполагает вовлечение врачей-экспертов с целью поиска (наличие/отсутствие – C), локализации (B) и сегментации (A)[13].

В случае локализации врачу необходимо обозначить координаты области интереса простой геометрической фигурой, в случае сегментации – обвести контур области интереса, т.е. создать пиксельную маску. Для уровня 2 (классификация находки) необходимо классифицировать находку, используя общепринятые шкалы (например, BI-RADS[14], ASPECTS[15]). Для уровня 1 (подтвержденный диагноз) необходимы данные медицинской карты, позволяющие поставить диагноз.

Классификация отображает взаимосвязь:

- объемов и качества исходных данных;
- трудозатрат на подготовку;
- методик разметки и работы с первичными данными;
- диагностической ценности.

Стоит отметить, что данная классификация применима в случае поиска патологических находок. Для некоторых НД, например, при задаче сегментации анатомической структуры, подтверждение диагноза неприменимо, соответственно данную классификацию использовать нельзя.

Также разметку данных можно разделить на проспективную и ретроспективную, т.е. по времени их получения.

Проспективная разметка аналогично ретроспективной представляет собой сбор элементов в соответствии с поставленной целью, при этом обязательным условием является проведение дополнительных манипуляций с элементами (например, постановка метки начала и окончания события, меток обнаружения признаков, обозначений патологий и т.п.). Этот вид разметки про-

---

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

[14] BI-RADS – Breast Imaging Reporting and Data System – стандартизированная шкала оценки результатов маммографии, УЗИ и МРТ по степени риска наличия злокачественных образований молочной железы: Breast Imaging Reporting & Data System / American College of Radiology [Internet]. [cited 2023 Apr 8]. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads.

[15] ASPECTS (Alberta Stroke Program Early CT Score) – шкала качественной топографической оценки изменений, выявляемых при КТ у пациентов с инсультом головного мозга: Pexman J. H., Barber P. A., Hill M., et al. Use of the Alberta Stroke Program Early CT Score (ASPECTS) for assessing: CT scans in patients with acute stroke // AJNR Am J Neuroradiol. 2001. Vol. 22, №8. P. 1534–1542.

водят с участием обученного медицинского персонала (зачастую квалифицированного врача в субспециализации размечаемого набора данных) путем ручного аннотирования содержания данных или их частей.

Ретроспективная разметка данных представляет собой сбор элементов в соответствии с метаданными, которые отбираются по поставленной цели. Такую разметку проводят путем минимальных трудозатрат: выгрузка данных происходит из медицинской информационной системы, которую может провести инженер (аналитик) без участия врача. При этом для каждого элемента (изображение, сигнальные данные и т.д.) набора данных устанавливают соответствие с медицинской информацией (диагноз, результаты лабораторного тестирования и т.п.).

Также разметка характеризуется следующими параметрами:

1. Уровень разметки: пациент, серия, набор изображений, изображение.

Примеры:

– на уровне пациента: у пациентки с диагнозом злокачественного новообразования (ЗНО) молочной железы разметка проводится на основании маммографии и гистологического исследования;

– на уровне серии (у той же пациентки): маммография, прямая и боковая проекции;

– на уровне изображения: прямая проекция правой молочной железы.

2. Тип разметки: бинарная, мультикласс, мультилейбл.

Примеры:

– бинарная разметка: норма/патология;

– мультиклассовая разметка: норма/патология/технический дефект;

– мультилейбл разметка: лейбл «Признаки эмфиземы легкого», лейбл «Процент поражения легкого».

3. Характер разметки: бинарная, категориальная, регрессионная.

Примеры:

– бинарная: наличие признаков патологии/отсутствие признаков патологии;

– категориальная: категория BI-RADS для маммографии;

– регрессионная: процент поражения легкого при COVID-19.

Контрольные вопросы

1. Какие бывают методы верификации данных?
2. Какие бывают виды разметки данных по диагностической ценности?
3. Как классифицируется разметка данных в зависимости от времени получения данных?
4. Перечислите параметры разметки.
5. Какие бывают уровни разметки данных? Приведите примеры.

ГЛАВА 2.
ЖИЗНЕННЫЙ ЦИКЛ НАБОРОВ МЕДИЦИНСКИХ ДАННЫХ

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

Жизненный цикл данных – последовательность этапов, которую конкретная порция данных проходит от начального этапа создания или получения до момента архивации или удаления [6].

Жизненный цикл наборов данных состоит из следующих этапов:

- инициирования;
- планирования;
- формирования;
- этап регистрации и публикации;
- использования;
- смены версии;
- удаления и архивации.

Последовательность и взаимосвязь этих этапов представлена на рисунке 3.

Рисунок 3 – Жизненный цикл наборов данных

Этап инициирования

Данный этап начинается с момента возникновения потребности или идеи создания НД, поэтому первое, с чем необходимо определиться – это цель их создания. На основании цели создания НД разработана классификация по типам:

I. Проведение тестирований для оценки функционала (функциональное тестирование) и оценки метрик диагностической точности, настройки ПО на основе ТИИ (калибровочное тестирование) [7].

II. «Самотестирование техническое» – проведение самостоятельной

проверки разработчиками способности ПО на основе ТИИ обрабатывать исследования с диагностических устройств разных производителей и моделей [8].

III. «Самотестирование диагностическое» – проведение самостоятельной проверки корректности клинической интерпретации исследований ПО на основе ТИИ [9].

IV. Выполнение клинических испытаний – оценка безопасности и эффективности медицинского изделия [4,10].

V. Выполнение технических испытаний – оценка соответствия характеристик ПО на основе ТИИ требованиям нормативно-правовой, технической и эксплуатационной документации [11].

VI. Проведение разметки текстовых протоколов с помощью программ автоматизированного анализа текстов.

VII. Проведение научных исследований [12].

VIII. Разработка ПО на основе ТИИ: обучение и дообучение [13].

После определения цели создания НД формируются или используются ранее подготовленные базовые диагностические требования (БДТ) и базовые функциональные требования (БФТ) [14]. БДТ – это требования к содержащейся в информации НД, необходимой для решения поставленных задач и достижения цели (модальность исследования, целевая патология, критерии отнесения исследований к классам и т.д.). Процесс создания БДТ описан в главе 3, подпараграф 3.1.1. БФТ – это описание технических особенностей отображения результатов клинических исследований (серия изображений, толщина срезов, окно визуализации и т.д.). Процесс создания БФТ описан в главе 3, подпараграф 3.1.2.

БДТ и БФТ – основные документы для формирования технического задания (ТЗ), которое в свою очередь является основным документом, регламентирующим и структурирующим разработку НД. Процесс создания ТЗ описан в главе 3, подпараграф 3.1.3.

Этап планирования

На этапе планирования определяются сроки подготовки НД, финансовые и людские ресурсы (назначаются исполнители, а именно врачи-разметчики, специалисты, ответственные за сборку НД и формирование сопровождающей документации, руководитель проекта), необходимые для подготовки НД, определяются риски (технические, административные и т.д.), которые могут повлиять на выполнение работы. При определении содержания работ, осуществляемых конкретным специалистом, проводится декомпозиция ТЗ на создание НД и уточняются требования к составу, количеству исследований, типам и способам разметки для каждого из задействованных специалистов (если это необходимо для выполнения работы).

Этап формирования

На данном этапе происходит непосредственно процесс создания НД: сбор данных, их разметка, структурирование, анонимизация, формирование файлов данных, разметки и сопроводительного текстового файла (readme-файла). Все файлы помещаются в хранилище данных. Подробный алгоритм формирования НД описан в главе 3 (параграф 3.3 «Этап формирования набора данных»).

Этап регистрации и публикации

На этапе регистрации вся информация о НД вносится в реестр. Полностью формируется так называемая карточка НД, где указываются все клинические, популяционные, технические параметры, параметры разметки, область применения, а также сформированные название и идентификатор НД.

Завершающим этапом процесса создания НД является его публикация – помещение структурированного набора файлов в отдельную директорию хранилища с регламентированным уровнем доступа.

По уровню доступа НД разделяются на общедоступные (открытые), ограниченного доступа (закрытые) и закрытые с общедоступными примерами. Общедоступные НД размещаются в открытом доступе (так называемые библиотеки НД) и предназначены для использования разработчиками ПО на основе ТИИ для проведения обучения, тестирования и/или валидации своей разработки.

Наборы данных, имеющие ограниченный доступ, предназначены для проведения внутренних исследований или для предоставления третьим лицам на особых условиях.

Регламент предоставления доступа к закрытым НД определяется политикой информационной безопасности и законодательством Российской Федерации.

Этап использования

На данном этапе происходит непосредственное использование НД согласно целям, обозначенным на этапе инициирования.

Этап смены версии

В процессе подготовки набора данных необходимо определить следующие шаги по его сопровождению:

- изменение уровня доступа третьим лицам;
- частота обновления;
- срок поддержки;
- способ утилизации.

Некоторые категории наборов данных подлежат регулярным обновлениям. Изменения могут вноситься как в сопроводительную информацию

(например, при появлении исследований в динамике для верификации), так и касаться самих единиц наборов данных (например, добавление новых случаев в особых эпидемиологических условиях). В этих случаях следует отдельно описать принципы получения новых данных, внесения изменений, в том числе в номер версии. Частота обновления данных оговаривается в ТЗ на подготовку набора данных.

Версия набора данных позволяет оценить вносимые изменения с течением времени. Изменения (включая любую смену версии) должны быть задокументированы, а документация – приложена к набору данных.

При смене версии предлагается использовать обозначения формата А.Б.В, где А – мажорная версия, Б – минорная версия, В – патч-версия[16]:

– мажорная версия увеличивается при изменении значимых параметров набора данных, связанных с клинической задачей, целью, принципами разметки и верификации данных;

– минорная версия увеличивается при замене, добавлении и удалении единиц данных (изображений, текстовых или сигнальных данных и др.) без изменений значимых параметров набора данных (минорная версия Б = 0 при выпуске новой мажорной версии);

– патч-версия увеличивается при внесении корректировок в сопроводительную документацию, исправлении опечаток или ошибок в файлах разметки и верификации, при этом не меняется ни количество, ни качество входных данных (патч-версия В = 0 при выпуске новой минорной или мажорной версии).

Новый набор данных создается при условии полного изменения назначения, цели создания и клинических задач. При внесении изменений в НД ему присваивается новый идентификатор, и он проходит весь жизненный цикл заново.

Сроки поддержки НД определяются ТЗ и/или государственным контрактом на выполнение работ.

Этап удаления/архивации

По мере использования НД может оказаться более неактуальным и может быть скрыт из доступа, заархивирован в длительное хранение без возможности быстрого восстановления. Однако бесследное удаление НД не рекомендуется, так как в будущем может появиться необходимость восстановить источник «потерянных» исследований.

---

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

Контрольные вопросы

1. Опишите основные этапы жизненного цикла данных.
2. В чем заключается самотестирование техническое, самотестирование диагностическое?
3. В чем заключается выполнение технических испытаний?
4. Какие документы готовятся на этапе инициирования создания набора данных?
5. Опишите этап планирования. Какие основные факторы необходимо учитывать при планировании работ по созданию НД?

ГЛАВА 3.
АЛГОРИТМ СОЗДАНИЯ НАБОРА ДАННЫХ

3.1. Этап инициирования создания набора данных

3.1.1. Базовые диагностические требования

Основным содержанием БДТ являются требования к результатам работы ПО на основе ТИИ [14]; они включают следующую информацию:

1) тип исследования;
2) клиническая задача, решаемая ПО на основе ТИИ;
3) критерии классификации исследования;
4) содержание ответа ПО на основе ТИИ;
5) формат представления ответа ПО на основе ТИИ.

Раздел «*тип исследования*» включает в себя данные о модальности исследования и области сканирования. Данная информация представлена в клинических рекомендациях по лечению соответствующей патологии.

Раздел «*клиническая задача, решаемая ПО на основе ТИИ*» определяет целевое назначение НД.

При постановке клинической задачи необходимо учитывать, что ПО на основе ТИИ решает те задачи или часть задач, которые решает врач при описании исследования по данному направлению. Например, при описании ММГ врач-рентгенолог должен оценить:

1) техническое качество выполнения исследования (PGMI[17]);
2) плотность молочной железы по шкале ACR[18];
3) изменения в ткани молочной железы по шкале BI-RADS (выявить и классифицировать).

Соответственно, при постановке клинической задачи для ПО на основе ТИИ может быть решена как отдельно третья задача, так и все три. В зависимости от поставленной задачи для ПО будут формироваться требования для НД. Таким образом, НД может состоять из исследований, классифицированных только по шкале BI-RADS, или может быть дополнен значениями по шкале ACR и PGMI.

---

[17] PGMI – Perfect, Good, Moderate, Inadequate – метод оценки качества клинического изображения при маммографии: National Health Service Breast Cancer Screening Programme. National quality assurance coordinating group for radiography. Quality assurance guidelines for mammography including radiographic quality control. Sheffield: NHSBSP; 2006. Publication No. 63; ISBN 1 84463 028 5.

[18] ACR – American College of Radiology – шкала рентгенологической плотности молочной железы: Breast Imaging Reporting & Data System / American College of Radiology [Internet]. [cited 2023 Apr 8]. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads.

Однозначность и точность формулировки клинической задачи позволяет определить, в частности, является ли выбранный метод диагностики «золотым стандартом» в отношении выбранной патологии либо требуется дополнительная верификация. Так, например, для злокачественных новообразований легкого корректная формулировка клинической задачи звучит следующим образом: «Выявление компьютерно-томографических признаков, коррелирующих с наличием злокачественных новообразований в легких», так как диагноз «ЗНО легкого» не может быть выставлен только по данным компьютерной томографии (КТ). В то же время для диагностики дилатации брюшной аорты необходимо и достаточно данных КТ, поэтому клиническая задача формулируется более конкретно: «Определение расширения брюшного отдела аорты».

Раздел «*критерии классификации исследования*» определяет тип классификации (бинарная либо мультиклассовая) и постулирует критерии отнесения исследования к каждому из классов. Необходимо указать однозначные и непересекающиеся критерии включения в каждый класс, поэтому стоит остановиться на некоторых аспектах терминологии.

Класс – это множество всех объектов с заданным значением метки. В медицинских данных чаще всего встречаются классы «наличие патологии/отсутствие патологии» в случае бинарной классификации или, в случае мультиклассовой классификации, шкалы степени тяжести заболевания (KT-COVID[19], BI-RADS, ASPECTS и т.д.). Еще одним примером мультиклассовой классификации является разделение исследований на группы: наличие патологии/отсутствие патологии/технический дефект. Лейбл (от англ. label – ярлык, этикетка) – название патологического (или нормального) состояния, которое подвергается классификации. Например, в НД компьютерной томографии грудной клетки может быть 2 лейбла: «признаки рака легких» и «признаки коронавирусной инфекции».

Критерии включения могут быть пороговыми, в тех случаях, когда диагностический признак патологии имеет явное численное определение (например, размеры объекта или его рентгеновская плотность), или неявными (например, паттерны поражения легочной ткани, характерные для пациентов с COVID-19). Стоит учесть, что для некоторых патологий диапазоны численных значений диагностических признаков могут перекрываться или быть относительными (например, гиподенсное образование печени – область, рентгеновская плотность которой ниже средней плотности неизменной ткани печени). В этих случаях следует отдельно рассмотреть возможность явного определения класса

---

[19] Prokop M., Van Everdingen W., Van Rees V., et al. KT-COVID – шкала оценки степени поражения легких COVID-19 при КТ ОГК. CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19-Definition and Evaluation // Radiology. 2020. Vol. 296, №2. P. E97–E104. URL: https://doi.org/10.1148/RADIOL.2020201473/ASSET/IMAGES/LARGE/RADIOL.2020201473.FIG4.JPEG.

патологии как совокупности отдельных признаков. При формулировке критериев особое внимание надо уделить доступности их восприятия: определения должны быть лаконичными и не содержать большой объем узкоспециальной терминологии, численные диапазоны величин необходимо сопровождать ссылками на литературные источники.

Формулировка критериев должна проводиться строго в рамках выбранной клинической задачи. При наличии у искомой патологии (иного) верификационного диагностического метода, его можно указать опционально (в разделе дополнительной информации).

В случае если первичной задачей при подготовке НД является не классификация исследований, а их сегментация или детекция, данный раздел формируется исходя из критериев отнесения объектов к целевой области интереса (перечисляют не диагностические, а сегментационные признаки объекта).

Раздел «*содержание ответа ПО на основе ТИИ*» должен отражать обязательные и опциональные требования к результатам работы ИИ. При формировании содержания раздела необходимо учесть целевое назначение ПО и его объективный потенциал. Важно в каждом конкретном случае разграничить задачи врача и ПО на основе ТИИ, учитывая, что общей задачей внедрения ТИИ в диагностику является автоматизация рутинных процессов. Целевое назначение конкретного ПО определяется исходя из клинического сценария диагностики соответствующей патологии.

Пример

Диагностика лимфаденопатии проводится на основании измерения лимфоузлов, следовательно, от ПО на основе ТИИ требуется измерение всех лимфоузлов выбранной области и выделение тех, чьи размеры превышают диагностический порог патологии.

К обязательным требованиям относят представление всех признаков, которые необходимы для однозначной классификации исследования в рамках поставленной клинической задачи. К опциональным требованиям относят функционал ПО, определяющий его потенциал в отношении удобства использования врачом-оператором.

При заполнении данного раздела стоит учитывать, что обе категории требований (обязательные и опциональные) к содержанию ответа ПО на основе ТИИ должны однозначно соответствовать разметке НД. Следовательно, НД должен быть подготовлен так, чтобы любой из указанных в данном разделе пунктов можно было бы верифицировать или оценить.

Необходимо отметить класс диагностических задач, который предполагает косвенный расчет параметров. К таким задачам относят использование раз-

личных шкал или индексов (таких как Genant[20] или Agastson[21]) и других метрик, предполагающих применение специальных формул расчета или таблиц. В этих случаях результаты полуавтоматического анализа изображения обычно слабо коррелируют с субъективной оценкой врача. Например, процент поражения легочной ткани при эмфиземе может быть диффузным (распределенным) в объеме легких, что затрудняет постановку диагноза при визуальной оценке изображения, однако, не влияет на результат автоматического анализа области. При формировании НД для подобных задач стоит заранее определить инструментарий создания достоверной разметки (например, использовать инструменты полуавтоматической сегментации).

Раздел *«формат предоставления ответа ПО на основе ТИИ»* кратко обобщает формат и вид представления каждого из ответов ПО на основе ТИИ. Результаты его работы могут быть в виде одного либо комбинации следующих типов:

– численное значение конкретного диагностического признака (например, размер объекта) либо вероятности отнесения к классу;

– контур либо маска, однозначно устанавливающая локализацию объекта;

– текст для категориальных величин.

Формирование раздела определяется способом постобработки результатов ПО на основе ТИИ. БДТ содержит общую информацию о формате данных; детализация процедуры обмена данными отражена в базовых функциональных требованиях.

Для специалиста, создающего НД, ключевое значение в БДТ имеют наименование, клиническая задача и содержание подготовительного этапа. Главный этап относится к ПО на основе ТИИ на этапе практического использования. Пример заполнения БДТ представлен на рисунке 4.

[20] Genant – полуколичественная оценка и классификация патологических переломов тел позвонков при остеопорозе: Genant H. K., Wu C. Y., van Kuijk C., Nevitt M. C. Vertebral fracture assessment using a semiquantitative technique // J Bone Miner Res. 1993. Vol. 8, №9. P.1137–1148. DOI:10.1002/jbmr.5650080915.

[21] Agastson – метод вычислений степени кальцификации коронарных артерий при проведении КТ: Agatston A.S., Janowitz W.R., Hildner F.J., et al. Quantification of coronary artery calcium using ultrafast computed tomography // J Am Coll Cardiol. 1990. Vol. 15, №4. P. 827–832. DOI:10.1016/0735-1097(90)90282-t.

Базовые диагностические требования к результатам работы ИИ-сервисов для выявления изменений в легких при COVID-19 по данным КТ

Наименование	Клиническая задача, решаемая ИИ-сервисом	Подготовительный этап (ретроспективное исследование) – признаки исследований калибровочного набора данных	Основной этап (проспективное исследование) – признаки, для которых ожидаются положительный и отрицательный результаты работы ИИ-сервиса	Содержание ответа ИИ-сервиса	Формат ответа ИИ-сервиса	Форма предоставления ответа ИИ-сервиса
Компьютерная томография органов грудной клетки	Выявление компьютерно-томографических признаков, коррелирующих с поражением легких при коронавирусной инфекции (COVID-19)	**Есть признаки патологии:** А*. 1. Инфильтрация легочной паренхимы по типу «матового стекла» с обеих сторон, преимущественно периферической локализации, с наличием или в отсутствие инфильтрации легочной паренхимы по типу консолидации с положительным признаком воздушной бронхограммы. 2. Инфильтрация легочной паренхимы по типу «булыжной мостовой» (утолщение междолькового интерстиция на фоне «матового стекла») с обеих сторон, преимущественно периферической локализации, с наличием или в отсутствие инфильтрации легочной паренхимы по типу консолидации с положительным признаком воздушной бронхограммы. Б. (исключительно для подготовительного этапа) 1. Положительные результаты лабораторной верификации коронавирусной инфекции (COVID-19) при помощи ПЦР-тестирования. 2. Установленный диагноз U07.1 (Коронавирусная инфекция, вызванная вирусом COVID-19, вирус идентифицирован). Для отнесения исследования к патологии достаточно одного из признаков. * На нативных изображениях		**Обязательно** – вероятность поражений легких, вызванных COVID-19 (признаки из списка А)	Число	Apache Kafka Message + DICOM SR
				Обязательно – классификация степени тяжести поражения легких по категориям «КТ 0–4» с указанием вероятности отнесения к каждому из классов	Число КТ-0 – вероятность КТ-1 – вероятность КТ-2 – вероятность КТ-3 – вероятность КТ-4 – вероятность	Apache Kafka Message + DICOM SR
				Обязательно – поражение (%) паренхимы отдельно для каждого лёгкого	Число	Apache Kafka Message + DICOM SR
		Нет признаков патологии: нет ни одного признака из перечня А		**Обязательно** – локализация найденных патологических находок	Контур/маска	DICOM

ИСТОЧНИКИ:
1. Лучевая диагностика коронавирусной болезни (COVID-19): организация, методология, интерпретация результатов: методические рекомендации / сост. С. П. Морозов, Д. Н. Проценко, С. В. Сметанина [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. 65. – М. : ГБУЗ «НПКЦДиТ ДЗМ», 2020. – 80 с. – URL: https://tele-med.ai/documents/500/19_ЛУЧЕВАЯ__ДИАГНОСТИКА__КОРОНАВИРУСНОЙ__БОЛЕЗНИ.pdf (дата обращения : 24.05.2021).
2. Министерство Здравоохранения Российской Федерации. Временные методические рекомендации. Профилактика, диагностика и лечение новой коронавирусной инфекции (COVID-19). Версия 15 (22.02.2022)

Рисунок 4 – Пример заполнения базовых диагностических требований. В рамках «Эксперимента по использованию инновационных технологий в области компьютерного зрения в системе здравоохранения города Москвы» ПО на основе ТИИ являлись ИИ-сервисы

3.1.2. Базовые функциональные требования

Основное содержание БФТ представляет собой обязательный набор требований к работе ПО на основе ТИИ для использования в деятельности практикующих врачей-рентгенологов. Документ содержит детализированную информацию по следующим пунктам:

1) параметры исследования: серия, толщина срезов, окно визуализации;

2) подробное описание и требования к формату представления, структуре и содержанию результатов работы ПО на основе ТИИ;

3) подробное описание процедуры обработки данных ПО на основе ТИИ: порядок предоставления, защиты, хранения и удаления данных, обработки исключений (например, формирование сообщений об ошибках работы);

4) словарь ключевых терминов для формирования машиночитаемых отчетов о работе ПО на основе ТИИ.

Раздел «*параметры*» содержит в себе указание серии исследований, требуемых для анализа целевой патологии.

В зависимости от требований, заданных в БФТ, ответственным за подготовку НД сотрудником в специальном программном обеспечении для просмотра медицинских исследований для каждого исследования выбирается «окно визуализации», «серия» изображений и «толщина срезов». На рисунке 5 представлен пример расположения УИД, «окна визуализации», «серии» изображений с указанием «толщины среза».

В таблице 2 представлен пример заполненных базовых функциональных требований, использующихся при подготовке НД. Пункты 2–4 не используются сотрудниками, ответственными за подготовку НД, и более детально изложены в базовых рекомендациях к результатам работы ИИ-сервисов [14].

Для ПО на основе ТИИ единицей набора данных является парная запись входных данных и ожидаемые выходные данные после обработки и анализа входных данных. Ожидаемые выходные данные формируются в процессе разметки, т. е. процесс разметки позволяет сформировать эталонные, «правильные» ответы, по которым затем будут оцениваться ответы от ПО на основе ТИИ в целях их разработки, тестирования и/или пострегистрационного мониторинга.

БДТ и БФТ в совокупности формируют базовый комплект основной документации для разработки комплекта технической документации на формирование НД. Комплект технической документации представляет собой техническое задание и приложения к нему.

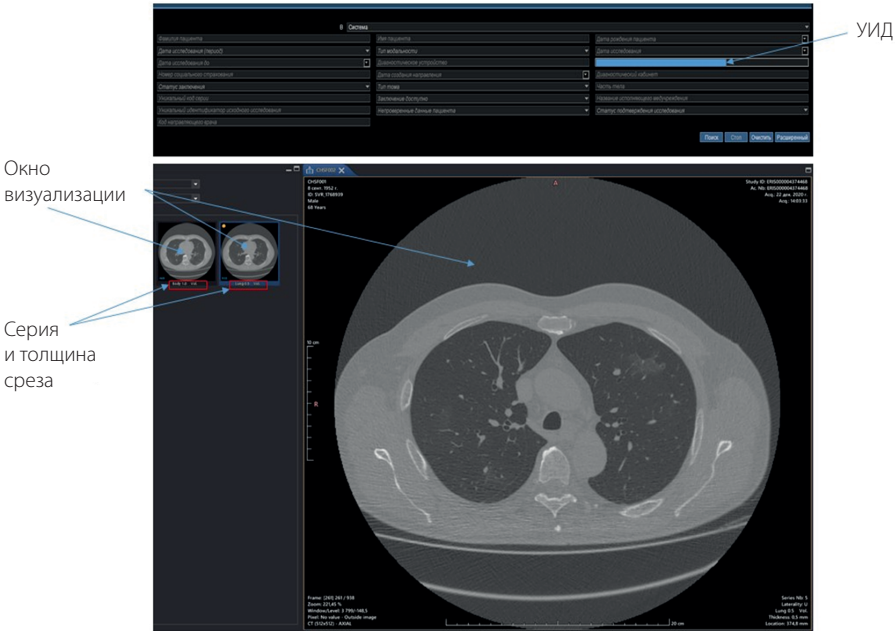


Рисунок 5 – Примеры расположения УИД-исследования, «окон визуализации», «серии» изображений и «толщины среза» в изображениях

Таблица 2 – Пример базовых функциональных требований, содержащих «окно визуализации» (совокупность данных из колонок с наименованием «Область исследований» и «Выбор окна (W/L)»), требования к «толщине среза»

Модальность	Область исследования	Целевая патология	Выбор окна (W/L)	Толщина среза
КТ	Область грудной клетки	COVID-19	Легочное	≤3 мм. Приоритет мини-мальной толщине
		Злокачественные новообразования легких	Легочное	
		Компрессионный перелом тел позвонков	Мягкотканное	
		Свободная жидкость (выпот) в плевральных полостях	Мягкотканное	
		Ишемическая болезнь сердца (коронарный кальций)	Мягкотканное	
		Аневризма грудного отдела аорты с определением диаметра грудной аорты	Мягкотканное	

Продолжение таблицы 2

Модальность	Область исследования	Целевая патология	Выбор окна (W/L)	Толщина среза
		Расширение легочного ствола с определением диаметра легочного ствола	Мягкотканное	
		Увеличенные внутригрудные лимфатические узлы (лимфаденопатия)	Мягкотканное	
		Эмфизема легких	Легочное	

3.1.3. Техническое задание на создание набора данных

Техническое задание на создание НД является основным документом, которым руководствуется сотрудник, ответственный за его создание. ТЗ должно регламентировать все процессы таким образом, чтобы можно было на его основе воспроизвести разработку НД.

Расчет объема НД и баланса классов зависит от цели создания НД, а также определяется заказчиком. Структура будущего НД и дизайн исследования определяются исходя из цели, клинической задачи и пожеланий заказчика, но должны составлять не менее четырех исследований [15].

При формировании ТЗ важно предусмотреть не только общее число исследований, включающееся в НД, но и распределение их по классам. Например, при проведении тестирований для оценки метрик диагностической точности может потребоваться соблюдение баланса классов (признаков). Напротив, НД, предназначенные для дообучения, могут иметь смещенный баланс классов в целях повышения точности при обнаружении отдельных видов патологии.

В калибровочном тестировании программного обеспечения на основе технологии искусственного интеллекта необходимо учитывать, что при различных балансах классов количество исследований, отражающих наибольшую неоднородность значений метрики диагностической точности (наибольшее отклонение от среднего значения), различно. Если целью валидации является установление наихудшего случая поведения значений метрик диагностической точности, то для исследованного ПО на основе ТИИ доля «не нормы» должна составлять 10 %, а количество исследований – 190.

В случае если валидация проводится в условиях ограниченного количества данных, то доля «не нормы» должна составлять 50 %, и количество исследований равняться 70. Следовательно, для изученного ПО на основе ТИИ будет наблюдаться максимальное отклонение от среднего значения метрик диагностической точности.

Стоит отметить, что большое разнообразие медицинских данных не позволяет сделать единый шаблон для ТЗ, однако ниже указаны наиболее важные и часто встречающиеся разделы. Представленный перечень требований не является конечным и может дополняться при наличии специальных требований к создаваемым наборам данных.

Блок 1 «Общие положения» технического задания – представляется общая информация о проекте создания НД.

Пункт 1.0 «Название НД»: указывается предварительное название НД, которое должно отражать модальность (например, *КТ, РГ, МРТ* и т. д.), наименование и признаки целевой патологии (например, признаки артроза тазобедренного сустава), тип (см. классификацию по цели создания) и другие параметры на усмотрение ответственного за НД. Позднее, на этапе формирования, определяются окончательное название и идентификаторы наборов данных (глава 3, параграф 3.4 «Этап регистрации и публикации НД») на основании БДТ, БФТ (см. таблица 2, пункты «Модальность», «Целевая патология» и «Область исследования»).

Пункт 1.1 «Тип НД»: классификация НД в соответствии с целью создания.

Пункт 1.2 «Версия»: трехзначное численное обозначение (например, 1.2.0) A.B.C, где A – мажорная версия, B – минорная версия, C – патч-версия (см. главу 2 «Жизненный цикл наборов медицинских данных»).

Пункт 1.3 «Заказчик»: государственная организация, частная компания и др.

Пункт 1.4 «Источник финансирования»: источники финансирования, использующиеся для оплаты труда специалистов и иных расходов, возникающих при создании НД (например, *государственный контракт №123-б, за счет собственных средств* и т.д.).

Пункт 1.5 «Исполнитель»: организация, выполняющая подготовку набора данных.

Пункт 1.6 «Ф.И.О. сотрудника, ответственного от Исполнителя»: фамилия, имя, отчество руководителя проекта, ответственного за создание НД. Руководитель проекта назначается приказом руководителя организации, в которой создается НД.

Пункт 1.7 «Основание для создания НД»: юридическое основание для создания НД, а именно реквизиты приказа организации, в которой создается НД; реквизиты государственного контракта, на основании которого создается данный НД, и т.д.

Пункт 1.8 «Термины, определения и сокращения»: применяемые в документе обозначения и сокращения для наименования процедур и др.

Пункт 1.9 «Ключевые слова»: ключевые слова, имеющие ассоциативную связь с разработанным НД и обычно включающие наименование модальности, патологии, название НД.

Пункт 1.10 «Языки»: наименование языка, на котором представляется текстовая информация в НД (например, *Русский, English* и т.д.).

Пункт 1.11 «Список авторов»: перечень лиц участников проекта создания НД.

Пункт 1.12 «Дополнительная информация»: любая необходимая уточняющая информация, не указанная в полях 1.1–1.10.

Блок 2 «Назначение и цель создания НД».

Пункт 2.1 «Назначение и область применения, целевая аудитория»: раскрывает предполагаемый сценарий использования и назначение НД в практике.

Пункт 2.2 «Цель»: раскрывает функционал НД.

Блок 3 «Выгрузка из BI или других источников» (если применимо) – дана детализация по критериям и порядку отбора данных.

Наименование процедуры: наименование процедуры клинического исследования – данный пункт наследуется из БДТ, часть «Наименование», и уточняется по пунктам «Подготовительный этап», «Основной этап».

Параметры исследования: требования, полностью заимствованные из пункта «Выбор окна» и «Толщина среза» БФТ. Допускается уточнение.

Тип МО: тип медицинской организации, в которой получены исходные данные, используемые при создании НД.

Дополнительные ограничения: дополнительные ограничения, устанавливаемые заказчиком при создании НД (например, «Верхняя граница возраста пациента» и/или «Нижняя граница возраста пациента»). Требования могут дублироваться из пунктов «Подготовительный этап», «Основной этап» БДТ и уточняются на этапе формирования ТЗ.

Период сбора: временной интервал, за который проводились клинические исследования пациентов в учреждениях здравоохранения. Могут содержаться в БДТ, в пунктах «Подготовительный этап» и «Основной этап».

Количество исследований: общее количество уникальных исследований, которое необходимо обработать для отбора целевого числа исследований (оценочное значение).

Блок 4 «Выгрузка из МИС» – описаны требования к параметрам запросов для МИС (например, ЕРИС ЕМИАС) или других источников данных.

Пункт 4.1 «Формирование первичной выборки исследований. Перечень ключевых слов для каждого класса» содержит требования к данным, необходимым для получения информации, содержащейся в МИС. Должны присутствовать УИД пациента, содержание исследования, заключение врача. Устанавливается перечень ключевых слов, соответствующих норме (обозначается «0»), и перечень ключевых слов, соответствующих патологии (обозначается «1»). Ключевые

слова могут быть отражены в БДТ, в пунктах «Подготовительный этап» и «Основной этап».

Ключевые слова для осуществления отбора исследований (не путать с ключевыми словами-тегами из пункта 1.9 блока 1) являются специфическими медицинскими терминами, указывающими на наличие той или иной патологии), словосочетания, численные значения из которых формируются параметрами запроса, для осуществления выгрузки УИД пациентов из МИС.

Подпункт 4.1.1 «Необходимость включения в класс „Без патологии" исследований с патологическими изменениями, от которых следует дифференцировать целевую патологию» может принимать значения «да» или «нет». При их наличии следует перечислить патологии дифференциального ряда.

Пункт 4.2 «Необходимость отбора медицинским специалистом исследований согласно пункту 4.1 по текстовым протоколам медицинских описаний и заключений» содержит один из двух вариантов: «да» или «нет». Если пункт содержит «да», то устанавливается лицо ответственное за проведение проверки данных, и лицо, которое утверждает результаты проверки. Если содержится «нет», то проверка не выполняется.

Пункт 4.3 «Список УИД-исследований для включения в НД, отобранных медицинским специалистом. Предварительный отбор на основании предразметки текстовых протоколов медицинских описаний и заключений» состоит из количества исследований, содержащих патологию, и количества исследований, содержащих норму. Описываются требования к патологическим исследованиям, которые могут заимствоваться из БДТ, пункты «Подготовительный этап» и «Основной этап».

Пункт 4.4 «Требования для формирования подвыборок в НД»: требования количества данных, содержащих норму, и содержащих патологию, в подвыборках. Требования для подвыборок могут быть заимствованы из БДТ (пункты «Подготовительный этап» и «Основной этап»).

Блок 5 «Обработка данных ТЗ на создание НД» – требования к обработке результатов выгрузки из МИС. В случае если выгрузка осуществлялась из ЕРИС ЕМИАС, то это – требования к обработке DICOM-файлов.

Пункт 5.1 «Обезличивание DICOM-исследований»: описывается, какими средствами производится обезличивание DICOM-исследований.

Пункт 5.2 «Проверка неповрежденности тегов»: один из двух вариантов «да» или «нет», отвечает на вопрос «Требуется ли проверка тегов DICOM-исследования на наличие повреждений?».

Пункт 5.3 «Дополнительная защита интеллектуальной собственности»: дополнительные требования защиты интеллектуальной собственности техническими средствами.

Блок 6 «Разметка ТЗ на создание НД» – требования к разметке изображений, содержащихся в DICOM-файле. Под разметкой понимается установка категориального или визуального признака в данных, выполненная экспертом. При разметке изображений требования устанавливаются для визуального признака.

Пункт 6.1 «Необходимость разметки с привлечением медицинских специалистов по направлению»: требования привлечения профильного медицинского специалиста к процессу разметки изображений. Пункт может содержать один из двух вариантов ответов: «да» или «нет».

Пункт 6.2 «Количество и уровень медицинского персонала, привлекаемого к разметке»: число медицинских специалистов, привлекаемых для разметки изображения, их профиль и уровень квалификации. Данный пункт заполняется в случае, если в пункте 6.1 установлено требование привлечения медицинских специалистов («да»).

Пункт 6.3 «Необходимость численных измерений параметров целевой патологии»: требования к необходимости измерений пространственных (средних размеров – длина, ширина, высота и т.д.), временных (частота сердцебиения, амплитуда и т.д.) и/или плотностных характеристик патологии. Данный пункт содержит один из двух вариантов: «да» или «нет».

Пункт 6.4 «Критерии согласованности разметчиков»: критерии согласованности разметчиков при проведении измерений патологии. Устанавливаются критерии, при которых измеренные разметчиками величины считаются совпадающими, и критерии значительных расхождений при проведении измерений. При значительных расхождениях в измерениях устанавливается необходимость третьего (экспертного) мнения.

Данный пункт заполняется, если в пункте 6.3 установлен ответ «да».

Пункт 6.5 «Требования, предъявляемые к разметчикам»: требования, предъявляемые к медицинскому персоналу, участвующему в разметке изображений. Наличие специальных сертификатов, стажа работы в данной области медицины, сертификата об окончании курсов повышения квалификации и т. д.

Подпункт 6.5.1 «Требования, предъявляемые к экспертам»: требования, предъявляемые к экспертам, участвующим в разметке исследования. Наличие медицинской практики по данной специализации в течение определенного количества лет, регулярное прохождение курсов повышения квалификации, наличие ученых степеней и званий в данной научной области исследования и т.д.

Пункт 6.6 «Требования к техническому состоянию разметки»: требования к техническому состоянию объекта, подвергаемого разметке. В случае если размечаемый объект не удовлетворяет требованиям данного пункта, то разметчик признает это исследование «браком», помечает его как брак и помещает в отдельное хранилище с сохранением всех идентификационных данных исследования.

Пункт 6.7 «Требования к разметке целевой патологии»: требования к проведению процесса разметки и условия, при которых привлекается эксперт. Описывается баланс классов (количество исследований, содержащих норму, и количество исследований, содержащих патологию). Устанавливаются правила совместной работы привлеченных к разметке медицинских работников (независимая разметка нормы и патологии, каждый привлеченный специалист просматривает исследования самостоятельно и т.д.) и правила привлечения эксперта при разметке исследований (эксперт привлекается только для просмотра спорных исследований или просматривает все исследования независимо). При заполнении данного пункта используются требования, установленные в БФТ, (пункты «Выбор окна», «Целевая патология»), и БДТ (пункты «Подготовительный этап» и «Основной этап»).

Подпункт 6.7.1 «Инструмент и метод, используемый для первичной разметки по целевой патологии»: наименование программного обеспечения, используемого при проведении разметки исследований медицинскими специалистами и экспертами.

Пункт 6.8 «Необходимость разметки по дополнительным параметрам»: указание на необходимость проведения дополнительной разметки. Данный пункт заполняется на основании требований БДТ (пункты «Предварительный этап» и «Основной этап»). Если дополнительные параметры не указаны, то в пункте прописывается «нет»

Пункт 6.8.1 «Инструмент и метод первичной разметки по дополнительным параметрам (нецелевой патологии»): наименование ПО, используемого при проведении первичной разметки по дополнительным параметрам. Допускается использование различного ПО, применяемого по основным параметрам и по дополнительным, при условии отсутствия потери качества разметки в исследованиях.

Пункт 6.9 «Ф.И.О. экспертов и медицинского персонала, привлекаемых для разметки и верификации разметки»: фамилия, имя, отчество медицинских работников, участвующих в разметке исследований по целевым патологиям, нецелевым патологиям; экспертов и Ф.И.О. лица, ответственного за работу группы.

Блок 7 «Дополнительная обработка размеченного НД» ТЗ на создание НД – дополнительные требования к обработке размеченного НД. Указываются необходимость изменения баланса представленных классов («да» или «нет»), дополнительные условия, накладываемые на балансы классов, требуемый объем памяти для хранения созданного НД и др.

Пункт 7.1 «Необходимость балансировки НД»: требования к проведению дополнительной балансировки классов, присутствующих в НД. Может принимать одно из двух значений: «да» или «нет».

Пункт 7.2 «Тип балансировки НД»: заполняется, если в пункте 7.1 установлено «да»; содержит количество норм, патологий и (если такие имеются) количество подтипов патологий и подтипов норм, необходимых для содержания в созданном НД.

Пункт 7.3 «Метод балансировки НД»: заполняется, если в пункте 7.1 установлено «да»; содержит количество исследований, на которые необходимо уменьшить основной класс.

Пункт 7.4 «Состав итогового набора данных»: количество исследований, содержащихся в конечном наборе данных, предъявляемом заказчику. Описывается количество исследований, содержащихся в каждом классе и подклассе.

Пункт 7.5 «Требуемый объем памяти для хранения НД»: требования к объему памяти, необходимому для хранения и обработки исследований.

В таблице 3 представлен пример типового ТЗ на подготовку набора данных с привлечением врачей-рентгенологов.

Таблица 3 – Пример заполнения технического задания на создание НД MosMedData РГ ОГК с признаками артефактов и дефектов укладки, тип VII

1 Общие положения	
1.1. Тип (для БД)	VII Проведение научных исследований
1.2. Версия	1.0.0
1.3. Заказчик	-
1.4. Источник финансирования	Грантовая деятельность
1.5. Исполнитель	ГБУЗ «НПКЦ ДиТ ДЗМ»
1.6. Ф.И.О. ответственного от Исполнителя	Иванов И. И.
1.7. Основание для создания НД	Соглашение с …от…
1.8. Термины, определения и сокращения	Рентгенография органов грудной клетки (РГ ОГК)
1.9. Ключевые слова	Рентгенография органов грудной клетки, технические замечания, нарушение методики сканирования, наличие артефактов, посторонние предметы
1.10. Языки	Русский
1.11. Список авторов НД	Иванов И. И., Петров П. П.
1.12. Дополнительная информация	-
2. Назначение и цель создания НД	
2.1. Назначение и область применения, целевая аудитория	Создаваемый НД предназначен для разработки и валидации ПО на основе ТИИ для оценки технического качества РГ ОГК

Продолжение таблицы 3

2.2. Цель	Оценка технического качества РГ ОГК
***3. Выгрузка из BI или других источников (если применимо)**	
Выгрузка из BI Study Instance УИД-исследований для взрослых пациентов (старше 18 лет) по наименованию процедуры, типу медицинской организации	**Наименование процедуры:** рентгенография органов грудной клетки, рентгенография органов грудной клетки обзорная
	Параметры исследования: исследование РГ ОГК в стандартных прямой и боковой проекциях
	Тип МО: все
	Дополнительные ограничения (при наличии): нет
	Период сбора: 2021
	Количество: 1000 уникальных УИД
4. Выгрузка из ЕРИС ЕМИАС	
4.1. Формирование первичной выборки исследований. Перечень ключевых слов для каждого класса	Таблица УИД с колонками, содержащими: описание, заключение врача, наличие (1) или отсутствие (0) следующих признаков: 1. Видимые анатомические структуры: 1.1. Контуры нечеткие. 1.2. Размытое изображение. 1.3. Срезана область исследования. 2. Укладка: 2.1. Критерии укладки: 2.1.1. Ротация. 2.1.2. Наклон. 2.1.3. Излишнее сгибание. 2.1.4. Излишнее разгибание. 2.1.5. Неправильный угол центрального луча. 2.2. Критерии положения пациента: 2.2.1. Позиция пациента противоречит методике. 2.2.2. Снимок перевернут (неправильное расположение органа на рентгенограмме). 3. ЦЛ (центральный луч) и диафрагмирование: 3.1. Не диафрагмирована область исследования. 3.2. ЦЛ не по центру кассеты. 3.3. Неправильный угол наклона ЦЛ. 4. Нарушение технических условий экспозиции: 4.1. Плохая контрастность. 4.2. Неравномерная оптическая плотность. 4.3. Зерно. 4.4. Движение. 5. Маркировка: 5.1. Не указана сторона исследуемой области. 5.3. Не указаны дата и время снимка. 6. Артефакты: 6.1. Устранимые артефакты:

Продолжение таблицы 3

	6.1.1. Украшения. 6.1.2. Одежда. 6.1.3. Посторонние предметы. 6.1.4. Наложение других частей тела. 6.1.5. Буква на области исследования. 6.2. Неустранимые артефакты: 6.2.1. Импланты, протезы. 6.2.2. ЭКС. 6.2.3. Порт, катетеры, помпы. 6.2.4. Дробь, пуля, металлическая стружка. 6.2.5. Артефакты от оборудования (полосы, наводки и т. д.). 7. ЭЭД (эффективная эквивалентная доза): 7.1. Не указана. 7.2. Значение «ноль». 7.3. Очень большое значение. 8. Номенклатура: 8.1. Выставленная услуга не соответствует области исследования. 9. Нет снимка. 10. Другое
4.1.1. Необходимость включения в класс «Без патологии» исследований с патологическими изменениями, с которыми следует дифференцировать целевую патологию	Нет
4.2. Необходимость отбора медицинским специалистом исследований, согласно пункту 4.1, по текстовым протоколам медицинских описаний и заключений	Да Проверку проводит заведующий отделом экспертизы и качества
4.3. Список Study Instance УИД-исследований для включения в НД, отобранных медицинским специалистом. Предварительный отбор на основании предразметки текстовых протоколов медицинских описаний и заключений	Должно быть отобрано: 1. **С техническими замечаниями:** не менее 500 исследований. 2. **Без технических замечаний:** не менее 500 исследований
	Критерии отбора: 1. **С техническими замечаниями:** нарушение укладки нефункциональное, нарушение укладки нефункциональное (комментарий), нарушение укладки функциональное, нарушение укладки функциональное (комментарий), нарушение методики, нарушение методики (комментарий), устранимые артефакты (от элементов одежды, неправильных действий лаборанта, движения, дыхания и т. д.), устранимые артефакты (от элементов одежды, неправильных действий

Продолжение таблицы 3

	лаборанта, движения, дыхания и т. д.) (комментарий), неустранимые артефакты (связанные с оборудованием, физиологией человека, от протезов, имплантов), неустранимые артефакты (связанные с оборудованием, физиологией человека, от протезов, имплантов) (комментарий), не подлежит дальнейшей оценке (нет протокола описания, нет изображений или их кол-во недостаточно для объективной оценки), не подлежит дальнейшей оценке (нет протокола описания, нет изображений или их кол-ва недостаточно для объективной оценки) (комментарий). 2. **Без технических замечаний:** без признаков по п.1
4.4. Формирование подвыборки	Из списка УИД по п. 4.3 отобрать: 1. **С техническими замечаниями:** не менее 400 исследований. 2. **Без технических замечаний:** не менее 400 исследований
5. Обработка данных	
5.1. Обезличивание DICOM-исследований	Выполняется с помощью стандартных средств
5.2. Проверка неповрежденности тегов	Нет
5.3. Дополнительная защита интеллектуальной собственности	Нет
6. Разметка	
6.1. Необходимость разметки с привлечением медицинских специалистов по направлению	Да
6.2. Если в п. 6.1 выбран ответ «да», количество и уровень медицинского персонала, привлекаемого к разметке	Определяется в процессе проведения разметки, составляет от 2 до 3: два разметчика проводят первичную разметку, третий разметчик (эксперт) привлекается в случае необходимости разрешения несогласия между разметчиками
6.3. Необходимость численных измерений параметров целевой патологии	Да
6.4. Если в п. 6.3 выбран ответ «да», критерий согласованности разметчиков	По умолчанию: третий разметчик (эксперт) привлекается только в случае, если имеется расхождение в численных измерениях параметра целевого дефекта
6.5.1. Требования, предъявляемые к разметчикам	Действующий сертификат по специальности «Рентгенология»
6.5.2. Требования, предъявляемые к разметчикам-экспертам	Стаж работы более 5 лет

Продолжение таблицы 3

6.6. Требования к техническому состоянию разметки (проводится медицинским специалистом в процессе разметки)	Требуется подтверждение, что изображение пригодно для включения в разметку и последующего включения в НД с технической точки зрения
6.7. Требования к разметке целевой патологии	Результаты разметки должны включать все позиции, относящиеся к разметке технического качества изображений. Разметка проводится согласно инструкции. Общая схема разметки: 1. Исходная выборка по каждому классу делится на 2 части ((200 + 200) исследований «с техническими замечаниями» и (200 + 200) исследований «без технических замечаний»). 2. Каждый разметчик независимо просматривает 400 исследований (по 200 исследований из каждого класса), пока не будет отобрано 800 исследований (по 400 исследований из каждого класса). В случае необходимости измерений для каждого исследования создается и сохраняется ключевое изображение. 3. Разметчики обмениваются размеченными наборами данных (по 400 исследований) и проводят верификацию первичной разметки (по ключевому изображению, в случае наличия измерений). 4. В случае если общее число исследований одного класса, для которых мнение разметчиков согласовано, не соответствует требуемому количеству, выполняется следующее: 4.1 Для исследований класса «С патологией» привлекается третий разметчик (эксперт), присваивающий итоговую категорию исследованиям, для которых мнение первого и второго разметчика разошлись. Для исследований класса «Без технических замечаний» разметчики повторяют п. 2 и п. 3 для оставшихся (не просмотренных ранее) исследований из выгрузки
6.7.1. Инструмент и метод, используемый для первичной разметки по целевой патологии	Agfa (DICOM viewer ЕРИС ЕМИАС)
6.8. Необходимость разметки по дополнительным параметрам (нецелевой патологии) тип разметки: бинарная классификация (наличие (1) или отсутствие (0) патологии в исследовании)	Диагностические расхождения
6.8.1. Инструмент и метод первичной разметки по дополнительным параметрам (нецелевой патологии)	Agfa (DICOM viewer ЕРИС ЕМИАС)

Продолжение таблицы 3

6.9. Ф.И.О. экспертов и врачей-рентгенологов, привлекаемых для разметки и верификации разметки	
7. Дополнительная обработка размеченного НД	
7.1. Необходимость балансировки	Да
7.2. Если в п. 7.1 выбран ответ «да», тип балансировки	Распределение гарантирует одинаковую представленность примеров, относящихся ко всем выделяемым в НД классам
7.3. Если в п. 7.1 выбран ответ «да», метод балансировки	Уменьшение числа примеров мажоритарного класса
7.4. Состав итогового НД	Не менее 400 примеров с целевой находкой. Не менее 400 примеров без целевой находки
7.5. Требуемый объем памяти для хранения НД	XX Гб

Контрольные вопросы

1. Какие разделы входят в базовые диагностические требования?
2. Что включает в себя раздел «Критерии классификации исследования»?
3. Какие разделы входят в базовые функциональные требования?
4. Какие основные разделы входят в техническое задание?
5. Дайте определение понятию «разметка».

3.2. Этап планирования создания набора данных

Планирование проведения работ по созданию НД осуществляется на основании ТЗ, утвержденного заказчиком и согласованного исполнителем. Предварительное планирование может выполняться на этапе подготовки ТЗ.

Планирование проведения работ по подготовке набора данных осуществляет лицо, ответственное за ведение проекта, назначенное приказом директора (главного врача).

Основными составляющими любого проекта являются:

1. Сроки выполнения проекта.
2. Объем работ выполняемого проекта.
3. Финансирование проекта.
4. Содержание проекта.
5. Качество результатов проекта.
6. Риски, возникающие при выполнении проекта.

Каждый руководитель проекта может самостоятельно определять методологию планирования работ на основании принятых в организации норматив-

ных документов, требований ГОСТ и законодательства Российской Федерации. В данном учебно-методическом пособии приводится один из вариантов планирования, основанный на контроле сроков выполнения работ по подготовке НД.

Специалисты, участвующие в создании НД

При подготовке набора данных требуется координация большого количества специалистов, каждый из которых выполняет ряд определенных функций в соответствии с ролью. В зависимости от ТЗ несколько специалистов могут играть одну роль и наоборот: например, разметчиков должно быть минимум двое, а роль руководителя и менеджера может играть один специалист. Основные роли представлены в таблице 4.

Таблица 4 – Роли и функции специалистов, участвующих в создании набора данных

Роль	Функции
Руководитель	Организация процесса подготовки НД
Менеджер	Взаимодействие с заказчиком и сопровождение договорных отношений, контроль выполнения этапов
Аналитик	Составление и согласование ТЗ, таблицы и инструкции для разметки, ведение реестра
Технический специалист	Сборка, выгрузка, анонимизация НД. Формирование README-файла и сопроводительной документации
Разметчик	Разметка НД
Эксперт	Валидация разметки, согласование ТЗ, подготовка БДТ, БФТ

Исследования, проведенные в государственном бюджетном учреждении здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»[22], показали, что количество разметчиков, работающих с данными, должно составлять не менее двух человек. Количество экспертов, задействованных в пересмотре разметки, – не менее одного человека.

Расчет сроков выполнения работ

Первичное внимание уделяется следующим требованиям ТЗ:

1. Сроки выполнения работ по созданию НД.
2. Количество медицинского персонала, в том числе экспертов, подключаемых к созданию НД.

[22] Кульберг Н. С., Гусев М. А., Решетников Р. В. [и др.]. Методология и инструментарий создания обучающих выборок для систем искусственного интеллекта по распознаванию рака легкого на КТ-изображениях // Здравоохранение Российской Федерации. 2020. Т. 64, № 6. С. 343–350

3. Уровень задействования медицинского персонала, в том числе экспертов, в подготовке НД по другим ТЗ, по непосредственной специализации.

4. Ориентировочное время, затрачиваемое одним медицинским специалистом на разметку одного исследования.

5. Ориентировочное время, затрачиваемое экспертом на разметку одного исследования.

На основании описанных данных осуществляется расчет «чистого» времени, затрачиваемого каждым медицинским специалистом и каждым экспертом, задействованным в разметке набора данных по уравнению (1):

$$T_{НД} = ((t^{1}_{m,э} * N) * 1.2)/8, \quad (1)$$

где $t^{1}_{m,э}$ – ориентировочное время, затрачиваемое медицинским работником или экспертом на разметку одного исследования; N – число исследований, которое необходимо разметить; 1,2 – 20 %-коэффициент, позволяющий сделать запасы на риски; $T_{НД}$ – «чистое» время, затрачиваемое каждым медицинским специалистом или каждым экспертом на разметку набора данных (размерность, дни).

Уравнение (1) дает приблизительную оценку времени, не является точным и исчерпывающим и может изменяться в зависимости от принятых в организации нормативных документов.

Ориентировочное время, затрачиваемое медицинским специалистом или экспертом на разметку одного исследования, может быть установлено двумя путями:

1) на основании опроса медицинских специалистов и экспертов («Сколько времени потребуется на выполнение данной работы?»);

2) на основании хронометрических исследований, разметки медицинским специалистом и/или экспертом.

Полученное время разметки набора данных медицинским персоналом наносится на диаграмму Ганта. Пример диаграммы приведен на рисунке 6.

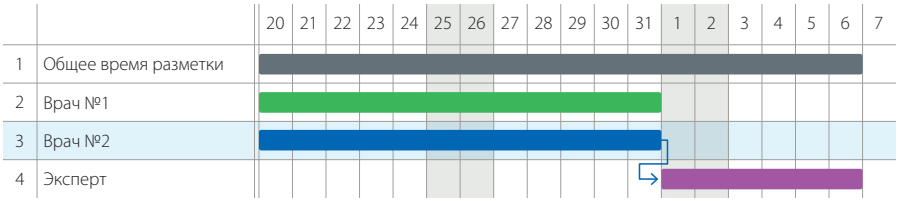


Рисунок 6 – Пример диаграммы Ганта для двух врачей, независимо друг от друга размечающих исследования, и привлечения эксперта для анализа спорных случаев

Полученная диаграмма показывает, сколько «чистых рабочих дней» выпадает на нерабочие дни. В случае примера, приведенного на рисунке 6, у врачей это 25-е и 26-е число, у эксперта – 1-е и 2-е число. Эти дни добавляются к общему расчету для получения конечного рабочего дня и завершения работ по разметке исследований врачами. Те же действия повторяются для получения конечного рабочего дня завершения просмотра размеченных исследований экспертом. После учета нерабочих дней лицо, ответственное за планирование проекта, получает «чистое время» разметки данных, затрачиваемое медицинским персоналом и экспертом.

После получения «чистого времени» разметки на диаграмму Ганта наносится период отпусков медицинских специалистов и экспертов, задействованных в подготовке набора данных. Если период отпусков пересекается с периодом выполнения работ, то:

1. Если приоритет подготовки НД «высокий» (приоритетность проводимых работ отражена в требованиях ТЗ), а количество медицинского и экспертного персонала, способного провести разметку исследований, достаточное (нет дефицита кадров), то такой эксперт или медицинский специалист заменяется; в противном случае (при наличии дефицита кадров) предусматриваются меры стимулирования или сдвигаются сроки выполнения работ, имеющих более низкий приоритет (средний или низкий).
2. Если приоритет подготовки НД «средний» и количество медицинского и экспертного персонала, способного провести разметку исследований, достаточное, то сроки выполнения разметки исследования сдвигаются относительно проектов с «высоким» приоритетом.
3. Если приоритет подготовки НД «низкий» и количество медицинского и экспертного персонала, способного провести разметку исследований, достаточное, то сроки выполнения разметки сдвигаются относительно работ с большим приоритетом.

После определения сроков и конкретных дат проведения работ по разметке исследований к полученным срокам добавляются сроки работ специалистов, готовящих сопроводительную документацию к набору данных.

По окончании расчета сроков выполнения работ проводится сопоставление полученных теоретических сроков и требований заказчика. В случае точного совпадения сроков или их превышения предпринимаются меры административного характера, направленные на сокращение сроков выполнения работ (привлечение субподрядчиков, привлечение персонала с более высокой квалификацией, меры финансового стимулирования и т. д.).

Основными рисками, возникающими на этапе разметки исследований, являются[23]:

1. Изменение технического задания на подготовку набора данных.
2. Временная нетрудоспособность медицинских работников и/или экспертов.
3. Возникновение форс-мажорных обстоятельств.
4. Действия административного персонала вне организации, выполняющей подготовку НД, препятствующие осуществлению работ.
5. Преступные действия в отношении сотрудников организации и/или организации, в целом препятствующие проведению работ по подготовке НД.
6. Действия лиц внутри организации, не заинтересованных в успешном выполнении проекта.

Данные риски должны быть учтены при планировании работ.

Расчет экономических показателей проекта

Расчет экономических показателей проекта создания НД осуществляется специалистами планово-экономического отдела предприятия, выполняющего работы по подготовке наборов данных, на основании внутренних документов организации и законодательства Российской Федерации. При наличии возможности и отсутствии административных ограничений в финансовой части проекта делаются запасы на риски в объеме 20 % от стоимости выполнения работ.

Контрольные вопросы

1. Опишите основные составляющие проекта.
2. Напишите формулу расчета времени, затрачиваемого сотрудником на разметку данных.
3. Опишите способы приоритизации группы проектов.
4. Опишите основные риски, возникающие при проведении проекта.
5. Какие «запасы на риски» принято создавать при планировании работ?

3.3. Этап формирования набора данных

На этапе формирования происходит процесс создания НД, а именно:

1. Сбор данных.
2. Разметка данных.
3. Структурирование данных.
4. Анонимизация данных.

[23] A Guide to the Project Management Body of Knowledge (PMBOK ® Guide), 2000. URL: http://www.cs.bilkent.edu.tr/~cagatay/cs413/PMBOK.pdf (дата обращения: 21.04.2023).

5. Формирование файлов данных и разметки.
6. Создание сопроводительного текстового файла (readme-файла).

3.3.1. Сбор данных

Основанием для сбора данных является ТЗ на подготовку НД. Источником данных могут служить пациенты, фантомы, а также данные могут быть синтетическими (сгенерированными математическими моделями). Чаще всего источником данных являются пациенты, а сами данные могут быть получены непосредственно в ходе исследования или выгружены из МИС, например, ЕРИС ЕМИАС. Ниже описан процесс сбора данных путем их выгрузки из ЕРИС BI (инструмент, обеспечивающий перевод транзакционной информации об исследованиях в ЕРИС ЕМИАС в человекочитаемую форму и работу с этой информацией).

Этапы сбора данных:

1. Выгрузка уникальных идентификаторов исследований.
2. Выгрузка медицинских исследований в формате DICOM-файлов по уникальным идентификаторам исследований.

Выгрузка уникальных идентификаторов исследований при формировании наборов данных осуществляется путем формирования запроса к базе данных ЕРИС BI, содержащего определенные условия. Такими условиями могут быть: перечень медицинских организаций, даты проведения исследований, названия процедур, модальности и т. д. (параметры условий запроса определяются соответствующими пунктами ТЗ). Для удобства работы пользователя используется интерфейс (рисунок 7), позволяющий формировать и применять наборы фильтров.

В данной системе доступные фильтры по характеристикам исследований в базе. В таблице 5 приведены самые распространенные фильтры, применяемые в системе ЕРИС BI.

Таблица 5 – Наиболее распространенные фильтры в системе ЕРИС BI

Название фильтра	Расшифровка названия фильтра	Примеры значений фильтра
Календарь	Время (проведения) исследования	Год, месяц, квартал, число, неделя, час, минута
МО	Медицинская организация	Название, адрес и т. д.
Вид учреждения	Вид медицинского учреждения	Взрослое или детское
Тип учреждения	Тип медицинского учреждения	Стационарное, амбулаторное, специальное и т. д.
Устройства	Характеристики диагностического устройства	Название, модель, производитель и т. д.
Тип услуги	Модальность исследования	КТ, МРТ, УЗИ и т. д.
Наименование процедуры	Название проведенной процедуры в данном исследовании	«Компьютерная томография грудной полости», «Rg-графия турецкого седла в 2-х проекциях», «Rg-графия трубчатых костей» и т. д.

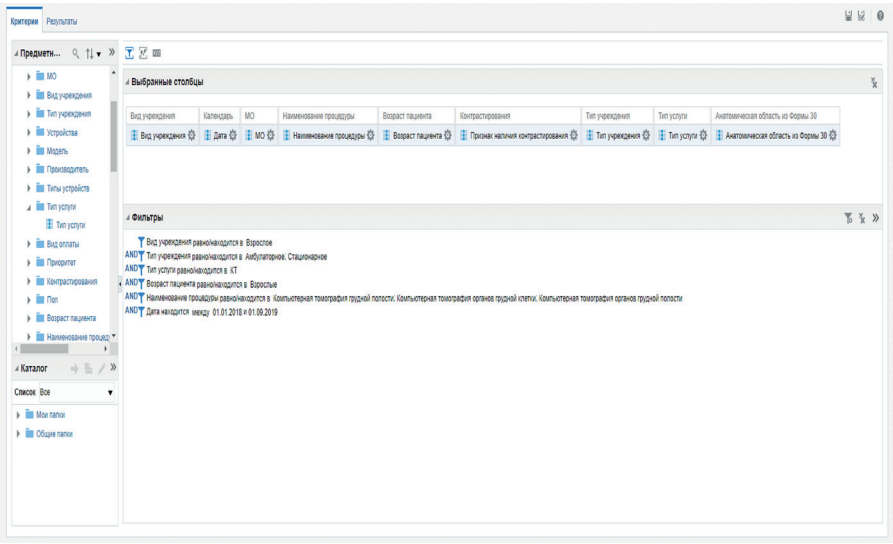


Рисунок 7 – Интерфейс ЕРИС BI с примерами фильтров

После выбора фильтров из системы ЕРИС BI можно получить список уникальных идентификаторов исследований, удовлетворяющих характеристикам фильтров. Полученный список используется для выгрузки медицинских исследований в формате DICOM-файлов на локальный компьютер.

Для загрузки медицинских данных лучевой диагностики из ЕРИС BI на локальное рабочее место используется специальный программный модуль. Он позволяет загружать на автоматизированное рабочее место медицинские исследования, содержащие различные модальности: рентгенограммы, флюорограммы, результаты магнитно-резонансной томографии, компьютерной томографии, позитронно-эмиссионной томографии.

Входными данными для модуля служат уникальные идентификаторы исследований. В модуле реализовано три варианта получения медицинских данных, отличающихся по объему и характеру содержащейся в них информации:

1) описательные данные;

2) протокол диагноза;

3) медицинские изображения в формате DICOM.

Описательные данные исследования содержат краткую информацию о пациенте (пол, возраст, дата рождения), проведенном исследовании (модальность и наименование выполненной процедуры) и медицинской организации, проводящей исследование. Протокол диагноза содержит описание и диагноз, поставленный врачом-рентгенологом в формате pdf.

Медицинские изображения – это набор файлов в формате DICOM, получаемых с диагностического устройства (аппарата МРТ, КТ и т. д.). Необходимый объем и содержание данных определяются по целевому назначению и требованиям технического задания на НД. По окончании выгрузки данных на локальный компьютер медицинский персонал и эксперты, задействованные в подготовке НД, приступают к процессу разметки.

3.3.2. Разметка данных

Для проведения анализа заключений и описаний, полученных при сборе данных, можно использовать готовое или разработать собственное специальное программное обеспечение, позволяющее на основании ключевых слов отбирать исследования с искомой целевой патологией[24]. Ключевые слова устанавливаются требованиями ТЗ.

Если разметка данных осуществляется на основании диагностического заключения и носит характер ретроспективной, то привлечение медицинских специалистов необязательно. В остальных случаях разметка НД проводится только с участием медицинских специалистов и/или медицинских экспертов.

---

[24] Кокина Д. Ю., Гомболевский В. А., Арзамасов К. М. [и др.]. Возможности и ограничения использования инструментов машинной обработки текстов в лучевой диагностике // Digital Diagnostics. 2022. Т. 3, № 4. С. 374–383. DOI: https://doi.org/10.17816/DD101099.

Медицинский персонал и эксперты, принимающие участие в разметке данных, могут быть отобраны по результатам предварительного тестирования. Количество разметчиков, работающих с данными, должно составлять не менее двух человек, экспертов – не менее одного человека. В приложении А в качестве примера приведено описание платформы, разработанной в ГБУЗ «НПКЦ ДиТ ДЗМ» для проведения тестирования медицинского персонала и экспертов, участвующих в разметке.

Требования к персоналу, выполняющему разметку

Персонал, осуществляющий деятельность, влияющую на качество подготовки набора данных, должен быть компетентным, иметь соответствующее образование, подготовку, навыки и опыт согласно ГОСТ ISO 13485-2017 «Изделия медицинские. Системы менеджмента качества. Требования для целей регулирования». Организация должна документировать процесс(ы), определяющий(е) компетентность персонала, проведение обучения, обеспечение информированности персонала.

Разметчиков необходимо подбирать по нескольким критериям:

– компетентность в области конкретных типов данных: изображения, текстовые данные или сигнальные (ЭКГ, ЭЭГ, спирометрия и т. д.), количественные данные (частота сердечных сокращений, артериальное давление, спирометрия и др.), бинарные данные (например, да/нет);

– уровень сложности планируемой разметки и/или аннотирования: первичная разметка (сегментирование) или экспертная; детализация на уровне классов или подклассов, установление связи с метаданными, определение вероятных исходов (прогнозирования);

– успешное прохождение предварительного тестирования (платформа тестирования – см. приложение А).

Требования к экспертам

В экспертную группу должны входить специалисты, имеющие большой опыт работы с определенным типом наборов данных. Как правило, предъявляют требование к опыту работы от трех лет.

Эксперты должны обладать опытом в областях, соответствующих решаемым задачам. При подборе их следует учитывать наличие конфликтов интересов, которые могут стать существенным препятствием для получения объективного суждения.

Требования к экспертам, привлекаемым к подготовке набора данных, должны быть документированы в рамках системы менеджмента качества.

Пример

Разметка НД на основании медицинских изображений с локализацией исследуемых структур (например, патологий) проводится только профильными медицинскими специалистами и/или профильными медицинскими экспертами.

При экспертной разметке требуются минимум 2 двое разметчиков. В случае недостижения консенсуса подключается третий (эксперт). При проведении разметки оконтуриванием, в зависимости от цели и необходимой точности, возможны сценарии:

– *оконтуривает разметчик, далее эксперт проверяет, при необходимости корректирует;*

– *для задач, связанных с высокой точностью оконтуривания, требуется минимум двое разметчиков. Финальная разметка определяется по объединению контуров или их пересечению. Далее подключается эксперт, осуществляющий валидацию, при необходимости – корректировку контуров. Обязательно требуется инструктаж и тестирование разметчиков перед допуском к работе.*

В приложении В представлен перечень программного обеспечения, которое может быть использовано при проведении разметки медицинских изображений.

3.3.3. Структурирование данных

Структурирование данных – это процесс разделения данных по отдельным критериям на группы, имеющие между собой логические связи.

После завершения разметки все данные собираются в единую структуру (структура может быть организована в виде файла электронной таблицы), содержащую:

1. Номер исследования по порядку.
2. УИД исследования.
3. Кластер медицинской организации: амбулаторный (А) или стационарный (С).
4. Наличие (1) или отсутствие (0) целевой патологии.
5. Наличие (1) или отсутствие (0) сопутствующих патологий.
6. Размеры (если требуется по Т3) основной патологии.
7. Размеры (если требуется по Т3) сопутствующей патологии.
8. Ссылка на местоположение файлов исследований (локальное хранилище, серверное хранилище и т. д.).
9. Номер исследовательского просмотра.
10. Указание на наличие (1) или отсутствие (0) брака в исследовании.
11. Комментарии к исследованию.

Пункты с 4 по 11 повторяются для каждого врача, просмотревшего каждое исследование.

В следующей вкладке файла электронной таблицы представляются:

1. Номер исследования по порядку.
2. УИД исследования.

3. Исходное изображение исследования (без разметки).
4. Изображение, содержащее разметку (если разметка исследования проводилась несколькими медицинскими сотрудниками и экспертами, то приводится каждое изображение).
5. Комментарии.
6. Ссылка на местоположение файлов исследований (локальное хранилище, серверное хранилище и т.д.).

Данные могут структурироваться на основании различных признаков и целей создания НД, отраженных в ТЗ на создание НД, и должны учитываться ответственными лицами при структурировании НД.

Структурированный набор данных является информацией закрытого типа, т. к. может содержать персональные данные пациентов, что не позволяет применять его в коммерческих или исследовательских целях. Для получения возможности широкого использования НД проводится «анонимизация» (обезличивание) набора данных.

3.3.4. Анонимизация (обезличивание) набора данных

При подготовке НД используются данные пациентов (за исключением данных, полученных с фантомов и синтезированных данных), которым оказана медицинская услуга. Перед получением медицинской услуги пациенты должны быть проинформированы о том, что их данные могут быть использованы для подготовки НД, и подписать информированное добровольное согласие.

Персональные данные в медицинских наборах данных должны быть удалены или анонимизированы. В нашей практике применяется анонимизация для самотестирования ПО на основе ТИИ (самотестирование), которые в дальнейшем размещаются в открытом доступе. Далее подробнее описан процесс анонимизации данных.

Методические рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 года № 996 «Об утверждении требований и методов по обезличиванию персональных данных» определяют процесс «обезличивание персональных данных» как действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных. Медицинские исследования могут содержать:

- личные данные пациентов;
- данные о врачах или медицинских работниках, проводящих исследования или другие действия для лечения пациента;
- информацию о медицинской организации, проводящей исследования;
- другие сведения.

Обезличивание персональных данных всех лиц, участвующих в процессе лечения, является обязательным.

DICOM-файл медицинского исследования – объектно-ориентированный файл с теговой организацией для представления кадра изображения (или серии кадров) и сопровождающей или управляющей информации в виде DICOM-тегов. Данный файл имеет древовидную структуру. К персональным данным относятся 11 атрибутов, но также имеются расширения группы тегов[25]. Существуют различные варианты анонимизации DICOM-изображений[26]. Исследователи рекомендуют расширенный список из 50 тегов, подлежащих удалению или изменению (таблица 6). Стандартные методы анонимизации реализуются с помощью замены или уничтожения тегов приватной информации. Состав тегов может меняться в зависимости от целей.

Таблица 6 – Расширенный список тегов для анонимизации

№	Tag ID	Tag Name
1	0008,0020	StudyDate
2	0008,0021	SeriesDate
3	0008,0022	AcquisitionDate
4	0008,0023	ContentDate
5	0008,0024	OverlayDate
6	0008,0025	CurveDate
7	0008,002A	AcquisitionDatetime
8	0008,0030	StudyTime
9	0008,0031	SeriesTime
10	0008,0032	AcquisitionTime
11	0008,0033	ContentTime
12	0008,0034	OverlayTime
13	0008,0035	CurveTime
14	0008,0050	AccessionNumber
15	0008,0080	InstitutionName
16	0008,0081	InstitutionAddress
17	0008,0090	ReferringPhysiciansName

[25] DICOM: [сайт]. United States, 2021. URL: http://dicom.nema.org/Medical/dicom/ 2016d/output/chtml/part03/sect_C.2.2.html (дата обращения: 15.04.2023).

[26] Aryanto K. Y. E., Oudkerk M., van Ooijen P. M. A. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy // European radiology. 2015. Т. 25, №. 12. P. 3685–3695.

Продолжение таблицы 6

№	Tag ID	Tag Name
18	0008,0092	ReferringPhysiciansAddress
19	0008,0094	ReferringPhysiciansTelephoneNumber
20	0008,0096	ReferringPhysicianIDSequence
21	0008,1040	InstitutionalDepartmentName
22	0008,1048	PhysicianOfRecord
23	0008,1049	PhysicianOfRecordIDSequence
24	0008,1050	PerformingPhysiciansName
25	0008,1052	PerformingPhysicianIDSequence
26	0008,1060	NameOfPhysicianReadingStudy
27	0008,1062	PhysicianReadingStudyIDSequence
28	0008,1070	OperatorsName
29	0010,0010	PatientsName
30	0010,0020	PatientID
31	0010,0021	IssuerOfPatientID
32	0010,0030	PatientsBirthDate
33	0010,0032	PatientsBirthTime
34	0010,0040	PatientsSex
35	0010,1000	OtherPatientIDs
36	0010,1001	OtherPatientNames
37	0010,1005	PatientsBirthName
38	0010,1010	PatientsAge
39	0010,1040	PatientsAddress
40	0010,1060	PatientsMothersBirthName
41	0010,2150	CountryOfResidence
42	0010,2152	RegionOfResidence
43	0010,2154	PatientsTelephoneNumbers
44	0020,0010	StudyID
45	0038,0300	CurrentPatientLocation
46	0038,0400	PatientsInstitutionResidence
47	0040,A120	DateTime
48	0040,A121	Date
49	0040,A122	Time
50	0040,A123	PersonName

Задача анонимизации DICOM-файлов актуальна для многих исследовательских и научных групп, создающих наборы медицинских данных. По этой причине в свободном доступе имеются отдельные алгоритмы или ПО для решения данной задачи. Примером такого ПО является KitwareMedical[27].

Данное ПО предусматривает 9 действий над группами тегов. Группы тегов и действия над ними прописаны во вспомогательном файле данного программного продукта. Группы тегов для каждого действия можно изменять. Достоинствами являются открытый программный код и простота пользования для небольших объемов данных, недостатками – аварийное завершение работы для некоторых типов файлов, доработка программного кода для работы с набором исследований.

Для устранения вышеуказанных недостатков нами был разработан модуль анонимизации на языке Python, в котором в том числе была реализована поддержка набора сценариев анонимизации. По умолчанию используется вариант анонимизации согласно таблице атрибутов уровня конфиденциальности[28].

Тип используемого варианта анонимизации также записывается в обрабатываемом файле в тег «[0x0012, 0x0063]». В данном варианте действия проводятся над 433 тегами. Осуществляемые действия на каждом теге – это одна из операций: удаление, очистка, замена, замена с генерированием идентификатора. Реализация этих действий над соответствующими группами тегов позволяет значительно снизить риск распространения приватной информации в обрабатываемых DICOM-файлах. Модуль анонимизации поддерживает пакетную обработку по спискам исследований.

Дополнительные меры по защите персональных данных

Несмотря на то, что при формировании наборов данных исследования анонимизируются, остается риск раскрытия приватной информации о пациентах и медицинском персонале. И в этих условиях проблема защиты персональных данных встает очень остро. Стандартные методы анонимизации реализуются с помощью замены или уничтожения тегов приватной информации. Однако часть персональных данных может быть закодирована непосредственно в пикселях изображений (например, в дозовых отчетах, вторичных объемных реконструкциях и др.). Для этих целей специально создан тег с «прожигаемыми» персональными данными, но он не всегда заполняется, и данные могут оставаться.

---

[27] GitHub – KitwareMedical/dicom-anonymizer: Tool to anonymize DICOM files according to the DICOM standard: [сайт]. United States, 2021. URL: https://github.com/KitwareMedical/dicom-anonymizer (дата обращения: 15.04.2023).

[28] DICOM: [сайт]. United States, 2021. URL: https://dicom.nema.org/medical/ dicom/current/output/html/part15.html#table_E.1-1 (дата обращения: 15.04.2023).

Удаление приватной информации, внедренной в изображение, опирается на методы оптического распознавания текста (Optical character recognition, OCR). На рисунке 8 представлен пример медицинского изображения с внедренной приватной информацией.

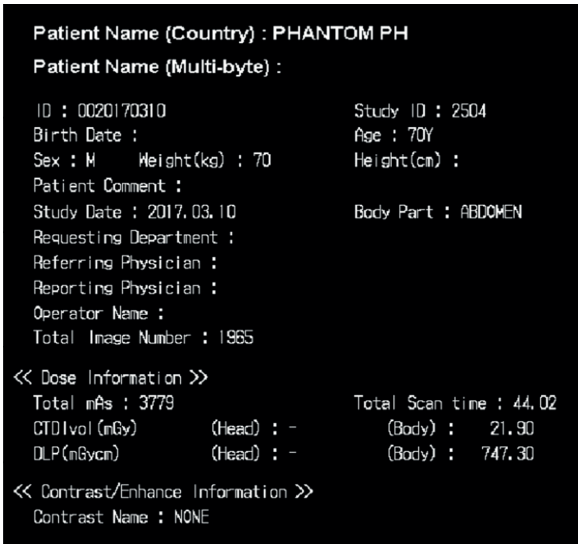


Рисунок 8 – Пример медицинского изображения с внедренной приватной информацией

Для анонимизации внедренного в изображения текста разработан модуль локализации и маскирования соответствующих участков изображения. После проведения анонимизации файлов медицинских исследований производится формирование файлов данных и разметки.

3.3.5. Формирование файлов данных и разметки

Под формированием файлов данных и разметки понимается размещение анонимизированных файлов медицинских исследований по папкам, каждой папке присваивается имя, соответствующее УИД медицинского исследования. В корневом каталоге размещается файл электронной таблицы, содержащий:

1. УИД медицинского исследования.
2. Окончательное заключение по исследованию (патология (1) или норма (0)).
3. Гендерные, возрастные признаки и т.д. (в зависимости от требований ТЗ).

По окончании формирования файлов данных и разметки формируется сопроводительный файл (readme-файл).

3.3.6. Сопроводительный readme-файл

README является распространенной формой документирования содержимого и структуры наборов данных, позволяет другим исследователям быстро найти тезисную информацию. Сопроводительный файл полезен и для разработчиков, так как позволяет осуществлять оперативный поиск, вносить изменения и облегчает долгосрочное хранение. Для сохранения единого формата документирования README целесообразно использовать инструменты автоматизации и структурные шаблоны.

Документация экспериментов, проектов, результатов анализа данных или информационных ресурсов может храниться в различных формах метаданных[29]:

– файл README (далее – README) – это текстовый файл, расположенный в папке, связанной с научным, техническим или программным проектом, который описывает его содержимое и структуру[30]. README также используется для документирования наборов данных;

– словарь данных: определяет и описывает элементы набора данных, чтобы их можно было однозначно интерпретировать и использовать позднее;

– протокол (методология): описывает процедуру или методы, используемые при реализации исследовательского проекта или эксперимента;

– лабораторные записи (чаще электронные): используются для документирования гипотез, экспериментов, анализов и интерпретаций экспериментов.

Формы документирования README:

– обычный текстовый файл с именем README, ReadMe, READ.ME, README.TXT[31], README.PDF, README.1ST;

– текстовый файл, написанный на языке разметки Markdown, с названием README.md и аналогичными[32].

Далее представлены требования к README для наборов данных медицинской диагностики. Важной особенностью сбора электронных медицинских данных, получаемых, как правило, в ходе рутинной клинической практики, является

---

[29] Harvard Biomedical Data Management: [сайт]. United States, 2022. URL: https://datamanagement.hms.harvard.edu (дата обращения: 20.09.2022)

[30] Guide to writing «readme» style metadata. Research Data Management Service Group: [сайт]. United States, 2022. URL: https://data.research.cornell.edu/content/readme (дата обращения: 20.09.2022).

[31] Raymond E. S. The New Hacker's Dictionary. MIT Press, 1996. P. 378–79.

[32] Abdelhafith O. README.md: History and Components // Medium. 2015. August 13. URL: https://medium.com/@NSomar/readme-md-history-and-components-a365aff07f10 (дата обращения: 27.04.2023).

их неоднородность, а также большое разнообразие форматов представления. Процесс сбора медицинских данных в единую структуру, имеющую необходимый набор характеристик и подходящую для дальнейших манипуляций с помощью алгоритмов или исследователями, еще не стандартизирован.

Одной из актуальных проблем в области искусственного интеллекта, нацеленного на решение клинических задач, остается отсутствие качественного README в папке с набором данных. Это является причиной неверной интерпретации и дальнейшего использования содержимого папки. Таким образом, еще на этапе планирования или подготовки набора данных следует создать документацию в формате README. При этом целесообразно применять стандартизированные и заранее разработанные шаблоны, что позволит сохранить единый формат сопроводительной документации для всех наборов данных, сформированных одной исследовательской организацией.

Одним из требований для допуска набора медицинских диагностических данных к публикации является наличие в корневой директории файлов README_XX.md и README_XX.pdf, где XX – буквенное обозначение языка файла согласно стандарту, ISO 639-1[33]. Файлы содержат следующие обязательные разделы: титульный лист, дисклеймер, общую информацию, аффилиацию и участников, структуру набора данных, обзор данных, правила использования и распространения, информацию о версии. Файлы README_EN.md и README_RU.md включают общую информацию о наборе данных на языке разметки и в формате Markdown на английском и русском языках соответственно. Файлы README_EN.pdf и README_RU.pdf содержат общую информацию о наборе данных в формате PDF на английском и русском языках соответственно.

В случае внесения изменений в НД по причине обнаружения ошибок или изменения содержимого необходимо документировать изменение его версии, что также отражается в README. Для удобства поиска и обращения к основной информации о НД, а также упрощения процессов создания README он должен соответствовать разработанному структурному шаблону. В «шапке» располагаются логотип учреждения, название, тип набора данных, краткая информация о содержимом. Далее следует письменный отказ от ответственности (дисклеймер) с указанием целей, для которых предназначен данный набор, периода сбора данных, диагностической модальности и краткой информации по лицензионным ограничениям. Тело README содержит всю информацию, необходимую для получения, интерпретации, повторного использования, распространения согласно лицензии, контактную информацию для связи с авторами,

[33] Code for the Representation of the Names of Languages. From ISO 639, revised 1989 // Cover Pages. URL: http://xml.coverpages.org/iso639a.html (дата обращения: 27.04.2023).

аффилиацию, а также структуру и обзор данных, ссылки для цитирования. Пример структурированного шаблона README для наборов медицинских данных представлен в приложении Б.

Экспертные научно-практические организации, которые специализируются в области медицинской диагностики, имеют возможность формировать наборы данных в непрерывном режиме для использования научным и информационно-технологическим сообществом. Для ускорения формирования и стандартизации предоставления качественных наборов данных необходимо применять структурные шаблоны и инструменты автоматизации. В сети присутствует множество шаблонов формирования документации в формате README онлайн или программные интерфейсы приложений, однако они не учитывают локальных особенностей представления данных либо не содержат всех необходимых разделов[34]. Кроме того, в связи с обеспечением информационной безопасности на данный момент нет возможности использовать внешние программные интерфейсы приложений.

Таким образом, одной из задач по оптимизации является получение README для НД в полуавтоматическом режиме. Для решения данной задачи возможно создание программных инструментов, которые позволят, используя разработанный шаблон, формировать однородные файлы README для всех НД. В дальнейшем инструмент можно интегрировать в единую систему полуавтоматического формирования наборов данных.

Контрольные вопросы

1. Опишите основные составляющие этапа формирования набора данных.
2. Опишите диаграмму методов верификации. Какие бывают методы верификации данных?
3. Какая из областей диаграммы обладает наименьшей ценностью? Какие – наибольшей? Почему?
4. Какой из типов разметки может выполняться специалистом, не имеющим медицинского образования? Почему? При каких условиях?
5. Дайте определение понятию «разметка данных».

3.4. Этап регистрации и публикации набора данных

Процесс создания НД – непростая задача, для решения которой необходимо привлечение большого числа специалистов, кроме того, каждый НД обладает огромным количеством различных характеристик начиная от классифи-

[34] Creating a README for your dataset (Quick Guide): [сайт]. United States, 2022. URL: https://zenodo.org/record/4058972#.Y1-6kjPP0uV (дата обращения: 31.10.2022).

каций и заканчивая техническими параметрами. Помимо этого, существует еще ряд проблем, которые требуют внимания:

1. Бурный рост количества данных и, как следствие, увеличение числа ПО на основе ТИИ, что в свою очередь приводит к созданию большого количества НД для решения различных задач. Например, при внешней валидации ПО на основе ТИИ требуется создание как минимум двух НД с той же спецификацией, но без повторения элементов НД: для первичного тестирования и (при необходимости) вторичного (вариант НД). Большое количество НД требует очень четкого, упорядоченного регламента процессов работы с ними на всех этапах жизненного цикла с целью упрощения их поиска и применения.

2. Необходимость создания единых стандартов представления информации о НД, ее структуризации, централизации и контроля качества. Существуют всевозможные справочники и словари, призванные стандартизировать и структурировать медицинские понятия и обеспечить семантическое отношение между терминами для удобства представления данных и обеспечения электронного обмена медицинской информацией: современная процедурная терминология (Current Procedural Terminology)[35], систематизированная медицинская номенклатура – клинические термины (SNOMED CT)[36], логические идентификаторы наблюдений, имена и коды (LOINC Logical Observation Identifiers Names and Codes)[37], словарь радиологических терминов RadLex[38]. Однако они имеют ряд ограничений[39], и многие пренебрегают их применением, зачастую ограничиваясь лишь использованием Международной классификации болезней (МКБ-10).

Кроме того, при публикации каждого готового НД необходимо приложить соответствующую документацию в виде сопроводительного текстового файла (readme-файл), в котором описаны основные параметры. На сегодняшний день единых стандартов такой документации не существует. Зачастую в readme-файлах отсутствует важная информация, которая могла бы позволить конечному пользователю принять решение о применимости данного НД в его задачах. Или, наоборот, такой файл может содержать избыточную, несистематизированную информацию, что также затрудняет процесс поиска необходимых данных. Базовая структура сопроводительного текстового файла описана в параграфе 3.3 «Этап формирования набора данных».

---

[35] UMLS Metathesaurus – CPT (CPT – Current Procedural Terminology) – Metadata [Electronic resource]. URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT/metadata.html (дата обращения: 19.04.2023).

[36] SNOMED – Home / SNOMED International. URL: https://www.snomed.org (accessed: 19.04.2023).

[37] Logical Observation Identifiers Names and Codes. URL: https://loinc.org (дата обращения: 19.04.2023).

[38] RadLex Term Browser. URL: http://radlex.org (дата обращения: 19.04.2023).

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

3. Нерациональность использования НД. Разметка результатов одного диагностического исследования, а тем более создание полноценного набора данных – это дорогостоящая и трудозатратная процедура, поэтому необходимо обеспечить долгосрочное, надежное и централизованное хранение данных с целью их возможного повторного использования для других задач, в том числе другими исследователями[40]. «Разумная бережливость» – один из принципов развития и использования ПО на основе ТИИ [1], однако многие медицинские и научные учреждения, имея качественные и актуальные НД для решения задач в рамках машинного обучения, часто используют локальные специальные схемы кодирования, которые ограничивают повторное использование, в том числе и другими организациями[41].

Вышеперечисленные проблемы решаются путем стандартизации процесса сбора, обработки и проверки набора данных на всех этапах жизненного цикла, чтобы упростить их использование и повысить эффективность для исследований в области машинного обучения. Для этих целей рекомендуется ведение реестра НД. Он формируется согласно этапам жизненного цикла НД и имеет 5 разделов (инициация, планирование, карточка НД, смена версии, использование), а также регламентирует процессы идентификации НД.

Прежде чем осветить вопросы структуры и полей реестра, необходимо остановиться на вопросе формирования названий и идентификации НД. Правильно сформированное название отражает максимальное количество информации о НД, что позволяет потенциальному пользователю максимально быстро и удобно принять решение о применимости НД в его задачах, а также существенно облегчает поиск и систематизацию данных. Наименования можно условно поделить на «названия» и «идентификаторы». Названия обычно представляются в публичном поле и должны быть максимально удобны и читабельны для широкого круга пользователей. Идентификаторы могут использоваться как в открытых публикациях, так и во внутренних процессах работы учреждения, и должны однозначно устанавливать НД в каждом конкретном случае.

Возможны следующие варианты номенклатуры НД:

1. Рабочее название. На этапе инициирования жизненного цикла будущему НД в первую очередь присваивается порядковый номер согласно реестру, а также рабочее название или, если необходимо, комментарии в свободной форме. Это позволяет отслеживать процесс внесения информации в реестр и контролировать качество работ по формированию НД.

---

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

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

2. Следующие три названия формируются уже при регистрации НД (карточка НД): внутренний (рисунок 9) и публичный (рисунок 10) идентификаторы и публичное название.

3.4.1. Внутренний идентификатор

Внутренний идентификатор уникален, необходим для однозначной идентификации набора данных, наименования файла разметки и предназначен исключительно для внутреннего пользования.

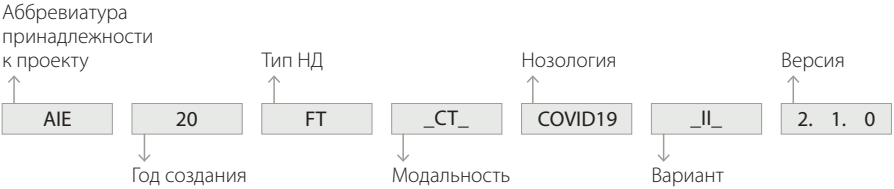


Рисунок 9 – Правила формирования внутреннего идентификатора НД

Название набора данных включает в себя следующие составляющие:

– аббревиатура, позволяющая идентифицировать принадлежность к какому-либо проекту (например, когда в рамках одного учреждения ведется подготовка наборов данных по разным проектам);
– последние 2 цифры года создания НД;
– назначение НД (исходя из типа, таблица 7, идентификатор);
– модальность (формируется согласно стандарту DICOM, таблица 8);
– целевая патология (нозология, «внутренний код»);
– вариант;
– версия;
– наличие разметки (только для наименования файла с разметкой – full_markup, в реестре не указывается).

Таблица 7 – Типы НД по их назначению

Тип	Идентификатор	Назначение НД	
I	FT/CT	Functional testing / calibration testing	Проведение функционального/ калибровочного тестирования
II	TST	Technical Self-test	Проведение селф-теста технического
III	DST	Diagnostic Self-test	Проведение селф-теста диагностического
IV	CIT	Clinical Test	Выполнение клинических испытаний
V	TT	Technical Test	Выполнение технических испытаний
VI	MedLabel	MedLabel	Проведение разметки текстовых протоколов для обучения MedLabel
VII	SS	Scientific Study	Проведение научных исследований
VIII	AID	Artificial Intelligence Design	Разработка ИИ

Таблица 8 – Основные типы (модальности) медицинских изображений, поддерживаемых стандартом DICOM

Идентификатор	Модальность	
EPS	Cardiac Electrophysiology	Электрофизиология сердца
CR	Computed Radiography	Компьютерная рентгенография
CT	Computed Tomography	Компьютерная томография
DX	Digital Radiography	Цифровая рентгенография
ECG	Electrocardiography	Электрокардиография
ES	Endoscopy	Эндоскопия
XC	External-camera Photography	Наружная фотография
IVUS	Intravascular Ultrasound	Внутрисосудистый ультразвук
MR	Magnetic Resonance	Магнитно-резонансная томография
MG	Mammography	Маммография
NM	Nuclear Medicine	Ядерная медицина
OP	Ophthalmic Photography	Офтальмологическая фотография
PX	Panoramic X-Ray	Панорамная рентгенография
PT	Positron emission tomography	Позитронно-эмиссионная томография
RF	Radiofluoroscopy	Рентгенофлюороскопия
RG	Radiographic imaging	Рентгенография
US	Ultrasound	Ультразвуковая диагностика
XA	X-Ray Angiography	Рентгеновская ангиография
BI	Biomagnetic imaging	Биомагнитные изображения

Продолжение таблицы 8

Идентификатор	Модальность	
CD	Color flow Doppler	Цветовое доплеровское картирование
ST	Single-photon emission computed tomography (SPECT)	Однофотонная эмиссионная компьютерная томография
TG	Thermography	Термография
AU	Audio	Аудиозаписи
SR	SR Document	Документ структурированного отчета
SMR	Stereometric Relationship	Стереометрическое взаимодействие
SC	Secondary Capture	Вторичный захват
OT	Other	Другое

3.4.2. Публичный идентификатор

Публичный идентификатор уникален, необходим для публичного использования и однозначного отображения файлов с открытым доступом (например, при формировании библиотек НД). Включает следующие составляющие (рисунок 10):

- наименование организации сбора НД;
- модальность;
- целевая патология (нозология, «внутренний код»);
- тип НД (таблица 7);
- порядковый номер версии НД (формируется согласно реестру и обеспечивает уникальность идентификатора).

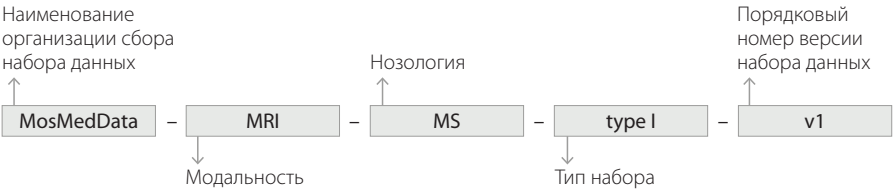


Рисунок 10 – Правила формирования публичного идентификатора НД

3.4.3. Публичное название (полное)

Публичное название (рисунок 11) не является уникальным, необходимо для использования в различных публикациях, библиотеках НД и открытых источниках, а также при регистрации интеллектуальной собственности и формировании readme-файла. Включает в себя:

– наименование организации сбора НД,
– модальность,
– локализацию,
– целевую патологию,
– тип НД (таблица 7).

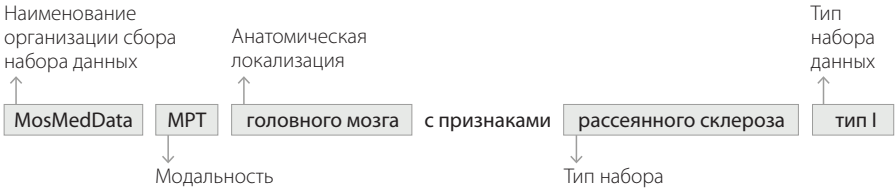


Рисунок 11 – Правила формирования публичного названия набора данных

Единый стандарт идентификации позволяет, только исходя из названия, понять то, какой организацией НД был подготовлен, исследования какой модальности и нозологии в него включены, а также его назначение, что способствует упорядочению, наглядному представлению данных и удобному обращению с ними. Также в зависимости от направлений работы учреждения, в котором создаются НД, возможны изменения структуры названия, а предложенные нами варианты можно использовать в качестве базы для формирования собственной схемы наименования.

3.4.4. Реестр как инструмент контроля качества набора данных

Как было описано выше, ведение реестра позволяет систематизировать данные, но для чего это необходимо? Помимо существенного упрощения процессов хранения и использования, реестр позволяет отслеживать факторы, влияющие на качество работы на всех этапах жизненного цикла НД. Среди них отмечают следующие: управление, непосредственно качество данных, конфиденциальность и безопасность [16].

1. Под управлением подразумевается организационная основа всех процессов жизненного цикла НД, обеспечение ресурсов (финансовых, человеческих и технических) для их функционирования [16]. Формирование набора данных должно быть спланировано и подвержено мониторингу и управлению для обеспечения соответствия качества.

Работой группы может руководить назначенный ответственным сотрудник, который не принимает участия в разметке и/или аннотировании, но будет регулировать срочность, очередность и объем работы между экспертами.

Его обязанностью также является формирование рабочей группы для обеспечения объективности и достоверности результата.

В процессе разработки и применения верифицированного набора данных внедряется система менеджмента качества, представляющая собой организационную структуру, функции, процедуры, процессы и ресурсы, необходимые для скоординированной деятельности по руководству и управлению данным процессом применительно к качеству.

Применение надлежащих принципов управления должно обеспечивать четкий и легкий способ сбора данных на всех этапах. С этой целью создается реестр НД и сопутствующая документация: правила, регламентирующие процесс создания, сопровождения, хранения и использования НД, и подробная инструкция по заполнению реестра со всеми необходимыми ссылками и справочниками. Помимо этого, в управление входит процесс контроля качества на всех этапах жизненного цикла, что обеспечивается и ведением реестра, который позволяет выявлять ошибки при создании НД, контролировать сроки выполнения работ, следить за процессом использования НД, а также оперативно генерировать справки и отчеты по заданным параметрам.

2. Качество данных: для получения медицинских данных требуется пройти весь процесс от начала регистрации пациента, проведения диагностики и исследования, в которых участвует медицинский персонал и используется медицинское оборудование, до составления заключения и передачи этих данных в единую систему хранения медицинских данных. Должны быть применены методы оценки качества набора данных, по которому будет производиться разметка:

– проверка отсутствия пропусков элементов в наборе данных;

– проверка отсутствия некорректных элементов для решения поставленных задач;

– проверка соответствия качества элементов набора данных рекомендованным критериям профессионального медицинского сообщества.

Подготовленные наборы данных могут быть структурированы посредством выделения признаков в соответствии с поставленной задачей.

В процессе структурирования снижают размерность набора данных, оставляя достаточный список атрибутов для точного и полного описания элементов, что будет способствовать последующему обобщению шагов и проведению качественной разметки (аннотации) данных.

Фильтрация набора данных позволяет уменьшить затраты на их разметку за счет исключения данных, не соответствующих заданным параметрам. Процедура контроля качества включает нахождение, предотвращение и устранение проблем, связанных с качеством наборов данных. Фильтрацию и контроль качества наборов данных возможно осуществлять с помощью визуального контроля, специальных инструментов (например, DICOM-валидаторов), а также с исполь-

зованием ПО на основе ТИИ (например, для автоматической оценки качества изображения).

Также следует учитывать сложнейшую структуру организации данных с рядом специфических параметров, которые необходимо фильтровать для использования ИИ. Одной из основных сложностей, связанных с подготовкой НД для машинного обучения и валидации ПО на основе ТИИ, является разнообразие характеристик, меняющихся от задачи к задаче, поэтому особое внимание необходимо уделять вопросам стандартизации и структуризации.

Качество самого реестра определяется не только результативностью использования данных (количество и качество публикаций, тестирований, разработок, запросов и т.д.)[42], но и «точностью» и «полнотой». Точность – степень, в которой зарегистрированные данные соответствуют истине, а полнота – степень, в которой все необходимые данные, которые могли бы быть зарегистрированы, действительно были зарегистрированы [16]. Надлежащее качество реестра обеспечивается процессами управления, описанными выше.

3. Вопросы конфиденциальности и безопасности данных связаны не только с защитой интеллектуальной собственности (которой является НД), но и с вопросами неприкосновенности частной жизни, персональных данных, врачебной тайны, что достигается за счет анонимизации данных. Все меры информационной безопасности должны регламентироваться действующим законодательством[43].

Таким образом, реестр представляет из себя таблицу, в строках которой находятся наборы данных, а в столбцах – их параметры. Нами была разработана структура реестра, параметры которого распределены согласно этапам жизненного цикла НД. Подробное содержание полей реестра приведено в примере реестра НД.

1. Параметры этапов инициирования и планирования обеспечивают процессы управления: они позволяют распределять ресурсы, отслеживать сроки выполнения работ, по необходимости обращаться к ответственным лицам для решения возникающих вопросов, контролировать процесс планирования и сбора НД, а также формировать справки о ходе работы для отчетности.

2. Параметры карточки НД заполняются на этапе регистрации. Это самый большой раздел реестра, который содержит структурированное описание НД, и из него формируются readme-файл и карточка НД в библиотеке НД.

---

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

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

На этом же этапе присваиваются идентификаторы и формируются названия. Данные карточки включают также коды различных справочников. Для реестра НД в медицинской диагностике мы предлагаем использовать следующие справочники:

– Федеральный справочник анатомической локализации[44];
– Федеральный справочник инструментальных диагностических исследований[45];
– МКБ -10[46];
– Номенклатура медицинских услуг[47];
– RadLex;
– LOINC;
– SNOMED.

Кроме того, в данном разделе предлагается поле «внутренний код», которое формируется исходя из целевой патологии и участвует в формировании идентификаторов.

3. Раздел «смена версии» имеет только один параметр, в котором указывается, сформирован ли НД на базе другого или направлен на утилизацию.

4. Параметры раздела использования НД необходимы для отслеживания информации о его применении. Как правило, в медицинской организации ведутся отдельные журналы и документы для фиксирования информации о тестированиях на различных платформах, о научном сотрудничестве, публикациях, доступе для разработчиков и другое. Данная информация хранится разрозненно, и при необходимости ее получения для отчетов или других целей требуется координация деятельности многих сотрудников. Во избежание этого ссылки на такие журналы и другая информация по использованию (публикации, сотрудничества) также фиксируются в реестре.

Также в этом разделе указывается информация о регистрации НД в ФИПС (Федеральный институт промышленной собственности): необходимость и статус регистрации. Для обеспечения централизованного хранения и оперативного доступа к подробным, максимально структурированным данным указываются ссылка на readme-файл, формат хранения файла и ссылки на место хранения НД с разметкой и без. В таблице 9 представлен пример содержания реестра набора данных.

---

[44] Справочник нормативно-справочной информации. Анатомические локализации. URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477/version/4.3 (дата обращения: 19.04.2022).

[45] Федеральный справочник инструментальных диагностических исследований. URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1471/version/2.15 (дата обращения: 19.04.2023).

[46] Международная классификация болезней 10-го пересмотра (МКБ-10) URL: https://mkb-10.com/ (дата обращения: 19.04.2023).

[47] Приказ Министерства здравоохранения Российской Федерации от 13.10.2017 № 804н «Об утверждении номенклатуры медицинских услуг». URL: http://publication.pravo.gov.ru/Document/View/0001201711080036 (дата обращения: 19.04.2023).

Таблица 9 – Пример содержания реестра НД

Раздел реестра	Краткое описание
Инициирование	
Рабочее название/комментарий	В свободной форме предварительное название НД
Заказчик/контактное лицо	Ф.И.О.
Тип НД	Согласно назначению НД (таблица 7)
Ответственный за формирование БДТ	Ф.И.О.
Дата начала работы над БДТ	-
Дата утверждения БДТ	-
Ссылка на БДТ	-
Планирование	
Ссылка на ТЗ на НД	Ссылка на место хранения файла с ТЗ
Планируемая дата начала подготовки НД	-
Планируемая дата завершения подготовки НД	-
Актуальный статус	На каком этапе работ находится НД на момент актуализации реестра. По завершении выполнения всех работ, т.е. после публикации НД, устанавливается статус «готов»
Дата смены статуса	Дата на момент актуализации статуса
Комментарий к статусу	-
Разметчики	Ф.И.О. специалистов, ответственных за разметку данных
Ответственный	Ф.И.О. сотрудника, ответственного за создание НД
Карточка НД	
Идентификация	
Год	Год создания НД или смены его версии
Внутренний идентификатор	Уникальный идентификатор для внутреннего использования
Публичный идентификатор	Уникальный идентификатор для публикаций в открытом доступе
Публичное наименование (полное)	Полное название НД на русском языке для публикаций в открытом доступе, readme-файла и др.
Версия ДС	Версия НД в формате А.Б.В
Условия доступа	Открытый/закрытый/ограниченный
Публичное наименование (полное)	Полное название НД на русском языке для публикаций в открытом доступе, readme-файла и др.

Продолжение таблицы 9

Раздел реестра	Краткое описание
Клинические параметры	
Модальность	Аббревиатура модальности согласно стандарту DICOM
Уникальный идентификатор анатомической локализации	Согласно федеральному справочнику (ФС) инструментальных диагностических исследований
Код RadLex	Код анатомической локализации целевой области согласно справочнику RadLex
Код LOINC	Код анатомической локализации целевой области согласно справочнику LOINC
Код SNOMED_CT (код анатомической области и нозологии)	Код анатомической локализации целевой области и код целевой патологии согласно справочнику SNOMED_CT
Наименование анатомической локализации (русское) согласно ФС инструментальных исследований	Полное наименование анатомической локализации целевой области
Наименование анатомической локализации (русское) согласно ФС анатомических локализаций	Полное наименование анатомической локализации целевой области
Наименование анатомической локализации (английское) согласно RadLex	Полное наименование анатомической локализации целевой области
Наименование анатомической локализации (английское) согласно ФС анатомических локализаций	Полное наименование анатомической локализации целевой области
Внутренний код	Код формируется на английском языке исходя из целевой патологии согласно МКБ-10
Название нозологии	На русском языке согласно МКБ-10
Код МКБ-10 направляющего диагноза	-
Код МКБ-10	Код МКБ-10 целевой патологии
Код услуги	Согласно номенклатуре медицинских услуг
Критерии включения/ невключения пациента	Критерии, по которым принималось решение о включении/невключении обследуемого в НД
Популяционные параметры	
Претестовая вероятность патологии	Частота встречаемости целевой патологии
в популяции	
Возраст (мин., лет)	Возраст самого младшего обследуемого в НД
Возраст (макс., лет)	Возраст самого старшего обследуемого в НД
Возраст (средний, лет)	Среднее значение возраста в НД

Продолжение таблицы 9

Раздел реестра	Краткое описание
Возраст (медиана, лет)	Медианное значение возраста в НД
Пол (М)	Количество лиц мужского пола в НД
Пол (Ж)	Количество лиц женского пола в НД
Пол (не определено)	Количество лиц, данные о поле которых отсутствуют
География сбора	Названия медицинских организаций, в которых происходил сбор данных или, при отсутствии такой информации, район, округ, географический субъект, в котором происходил сбор данных
Период сбора (начало)	Дата проведения самого раннего исследования в НД
Период сбора (конец)	Дата проведения самого позднего исследования в НД
Поток	Тип медицинских организаций, в которых происходил сбор данных: амбулаторный/ стационарный/ специализированный
Эпидемиологическая обстановка	Состояние распространенности инфекционной болезни людей на территории сбора данных на момент сбора.
Источник данных	Фантомные, синтетические, пациенты
Назначение (область применения)	
Клиническая/ практическая/ научная задача создания НД	-
Направление Эксперимента	Согласно приказу, порядку и условиям проведения Эксперимента
Вид тестирования	Для НД, предназначенных для тестирования: калибровочное, функциональное, селф-тест
Вариант	Для НД, предназначенных для тестирования вариант: I, II, III и т. д.
Параметры разметки	
Способы предразметки	Способ предварительного отбора информации в НД: вручную или с использованием какого-либо алгоритма (анализатора текстовых протоколов Medlabel)
Уровень разметки	Уровень, на котором происходит разметка данных: пациент, исследование, серия, изображение
Тип разметки	Бинарная, мультикласс, мультилейбл
Количество лейблов	Лейбл – название патологического (или нормального) состояния, которое подвергается классификации
Названия лейблов	
Характер разметки	Бинарная, категориальная, регрессионная
Количество классов	Класс – множество всех объектов с заданным значением метки, например, с патологией/без патологии
Названия классов	

Продолжение таблицы 9

Раздел реестра	Краткое описание
Количество по классам	Количество единиц набора данных в каждом классе
Класс разметки	Классификация НД по разметке
Метод валидации/верификации	Метод, с помощью которого верифицировались данные
Количество специалистов	Количество специалистов, участвующих в разметке одной единицы НД
Опыт специалистов	Стаж работы врачей и экспертов, участвующих в разметке
Временной промежуток между входными данными и данными верификации	Для данных, верифицированных с помощью других методов диагностики или исследований в динамике
Используемая при верификации информация из МК пациента	Для данных, верифицированных с помощью других методов диагностики или исследований в динамике
Критерии отнесения к классам	По каким критериям каждая единица НД относилась к тому или иному классу
Технические параметры	
Критерии включения/невключения исследования в НД	Критерии, по которым исследование включалось или не включалось с НД
Протоколы и условия сбора данных	Протоколы проведения исследований, включенных в НД, а также особые условия их сбора, если они были
Единичная запись НД:	Объект, подаваемый на вход модели МО, и результат разметки, получаемый от модели МО
Объект разметки	
Результат разметки	
Форматы записи НД:	Формат объекта, подаваемого на вход модели МО, и результата разметки, получаемой от модели МО
Объект разметки	
Результат разметки	
Количество записей НД	Количество единиц НД
Общий объем НД (Гб)	-
Количество уникальных источников	Количество диагностических устройств, с которых собирались данные
Перечень моделей и производителей	Модели и производители устройств, с которых собирались данные
Степень анонимизации	Способ анонимизации данных
Комментарий к НД	-
Смена версии/утилизация НД	
Смена версии/утилизация НД	Указывается, был ли НД сформирован в результате смены версии другого НД, или информация об утилизации НД

Продолжение таблицы 9

Раздел реестра	Краткое описание
Использование НД	
Актуальная версия для Эксперимента	Возможность использования НД в Эксперименте
Тестирование	Ссылки на журнал тестирования и необходимые комментарии
Научное сотрудничество	Данные о научных сотрудничествах, в рамках которых использовался НД
Научная статья	Данные о публикациях, в рамках которых использовался НД
Другое	Другие данные об использовании НД
Доступ для разработчиков	Ссылки, по которым доступен НД
Необходимость регистрации	Да/нет
Статус регистрации	На каком этапе регистрации находится НД на момент актуализации реестра
Ссылка на readme	Ссылка на место хранения readme-файла
Формат хранения НД	Формат, в котором НД находится в хранилище
Место хранения НД с разметкой	Ссылка на файл с разметкой
Место хранения НД без разметки	Ссылка на файл без разметки

3.4.5. Библиотеки наборов данных

Для широкого использования НД, экономии ресурсов и стимуляции развития ТИИ необходима публикация НД в открытых источниках. Согласно принципам поддержки конкуренции между организациями, осуществляющими деятельность в области ИИ, для продвижения научных исследований в этой области необходимо развитие исследовательской инфраструктуры и обеспечение доступа к НД посредством создания общедоступных платформ для их хранения, а также разработка унифицированных методологий описания, сбора и разметки данных и механизмов их контроля [1]. Для этого создаются библиотеки, в которых НД представлены в виде каталога карточек. В данном каталоге вся информация должна быть стандартизирована и отображена в наглядной форме. Очень часто библиотеки НД представлены в неструктурированном виде, а иногда и просто в виде разрозненного списка ссылок (например: https://github.com, https://www.kaggle.com). Чтобы исследователи, разработчики и компании легко и быстро могли найти и оценить применимость конкретного НД для их задач, необходимо создание удобных библиотек, с простым, интуитивно понятным интерфейсом.

На рисунке 12 в качестве примера представлен фрагмент библиотеки ГБУЗ «НПКЦ ДиТ ДЗМ»: https://mosmed.ai/datasets. Здесь данные максимально структурированы, имеются фильтры и поиск; карточки НД – наглядные и стандартизированные (см. рисунок 13). Стоит отметить, что создание такой библиотеки существенно упрощается благодаря реестру, т.к. их поля однозначно сопоставляются. Еще сильнее упростить процесс заполнения библиотеки могут инструменты автоматизации при выгрузке информации из реестра в библиотеку.

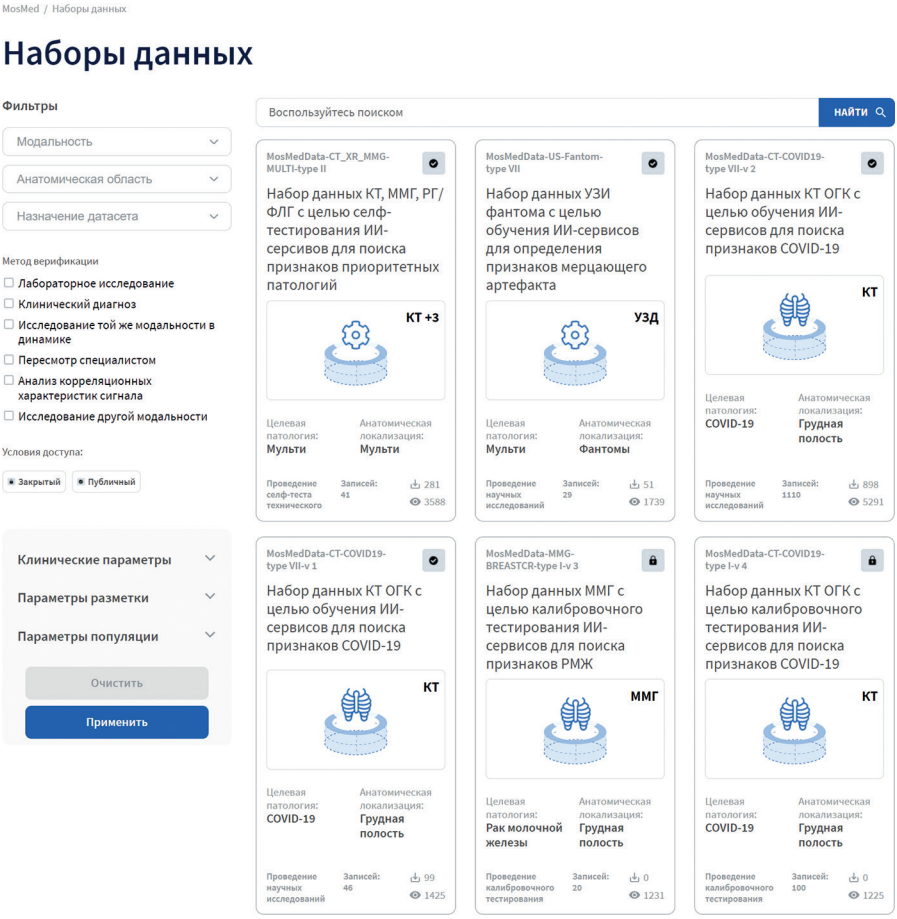


Рисунок 12 – Библиотека наборов данных (https://mosmed.ai/datasets)

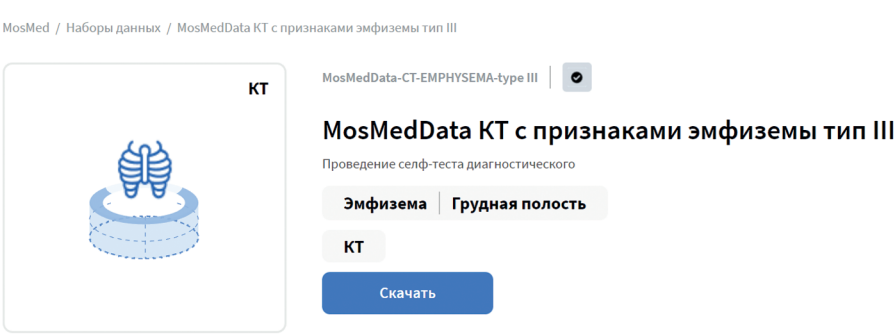


Рисунок 13 – Карточка набора данных в библиотеке

Таким образом ведение реестра позволяет:

– стандартизировать информацию о НД;

– обеспечить централизацию хранения, удобный и быстрый доступ ко всей информации о НД;

– формировать отчеты и справки для регуляции и повышения эффективности деятельности медицинской или научной организации по подготовке НД;

– обеспечить прозрачность, надежность и воспроизводимость разработок в сфере ИИ;

– оперативно сформировать унифицированные и наглядные карточки НД в библиотеках, позволяя пользователю принимать решение о применимости НД для его задач, минуя изучение сопроводительной документации.

Этапы использования, этап смены версии и этап архивации и удаления подробно описаны в главе 2 «Жизненный цикл наборов медицинских данных».

Контрольные вопросы

1. Опишите основные проблемы, возникающие при создании набора данных.
2. Опишите пути решения проблем, возникающих при создании набора данных.
3. Что такое реестр?
4. Какие основные разделы содержит реестр?
5. Какие названия рекомендуется использовать для набора данных?

ГЛАВА 4.
ОШИБКИ ПРИ ПОДГОТОВКЕ НАБОРА ДАННЫХ

При формировании набора данных на каждом этапе могут возникнуть определенные трудности и ошибки[48, 49], которые потенциально могут привести к дефектам в конечном наборе. В таблице 10 рассматриваются ошибки и их потенциальные решения на каждом этапе подготовки НД.

Таблица 10 – Ошибки, возникающие при подготовке набора данных, и способы их устранения

№	Процесс	Возможные ошибки/ трудности	Потенциальные решения
1	Формулирование клинической и практической задачи, которую потенциально можно решить с помощью ПО на основе ТИИ	– Пересмотр задач на последующих этапах	– Привлечение междисциплинарной команды для формулировки задач
2	Написание технических требований	– Корректировка на последующих этапах	– Привлечение междисциплинарной команды, более внимательная проработка ТЗ
3	Выгрузка (согласно ТЗ) исследований из МИС с текстовыми протоколами	– Долгая выгрузка исследований (технический аспект); – возникающие риски могут быть обнаружены лишь на более поздних этапах	– Оптимизация оборудования
4	Фильтрация исследований по ключевым словам (согласно ТЗ) в текстовых протоколах	– Неизбирательная фильтрация; – нехватка исследований – после фильтрации	– Оптимизация инструмента фильтрации, привлечение специалиста для подбора ключевых слов; – выгрузка большего количества исследований

[48] Krajnc D., Spielvogel C. P., Grahovac M., et al. Automated data preparation for in vivo tumor characterization with machine learning // Front Oncol. 2022. №12. DOI: 10.3389/fonc.2022.1017911.

[49] Diaz O., Kushibar K., Osuala R., et al. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools // Physica Medica. 2021. Vol. 83. P. 25–37. URL: https://doi.org/10.1016/j.ejmp.2021.02.007.

Продолжение таблицы 10

5	Просмотр исследований экспертами	– В медицинские изображения внедрены персональные данные пациента, которые нельзя удалить или анонимизировать; – при анонимизации обнаруживается, что описание присутствует, а изображения исследования в системе нет или оно не загружается	– Контроль «вшитых» персональных данных (проверка отсутствия текста на изображении); – исключение из набора данных исследований
6	Проверка заполняемости таблицы	– Пропущенные значения или столбцы	– Таблица возвращается на 4-й этап
7	Составление таблиц с разметкой (согласно ТЗ)	– Недостаточное количество исследований для формирования таблиц; – пересечение исследований (попадание 1 исследования в 2 набора)	– Увеличение выборки или повторение 3-го этапа; – инструмент проверки на дубликаты
8	Заполнение реестра	– При заполнении реестра недостаточно информации из ТЗ и БДТ (например, не указан код МКБ целевой патологии); – при формировании популяционных параметров обнаруживаются пациенты, не удовлетворяющие критериям отбора в ТЗ (например, границам возраста – младше 18 лет)	– Контроль корректности заполнения и корректировка структуры ТЗ и БДТ, автоматизация заполнения реестра; – проверка критериев включения/исключения на более ранних этапах
9	Составление сопроводительного README-файла	– Отсутствие какой-либо информации, необходимой для README; – ошибки при заполнении; – внесение корректировок после создания файла	– Внесение в реестр недостающих параметров; – автоматизация создания README; – утверждение всех параметров до момента создания README

В таблице 10 продемонстрирован общий порядок формирования набора данных, однако в зависимости от поставленной задачи может меняться количество этапов, их порядок, и соответственно будут изменяться возникающие трудности, ошибки, характерные для конкретной модальности и патологии.

Контрольные вопросы

1. Какие основные ошибки бывают при подготовке набора данных?
2. Опишите пути устранения ошибок, возникающих при написании технического задания.
3. Опишите основные ошибки, возникающие при просмотре исследований экспертами.
4. Какие основные ошибки возникают при составлении сопроводительного README-файла?
5. Опишите основные ошибки и пути их решения при заполнении реестра набора данных.

ГЛАВА 5.
ПРИМЕР СОЗДАНИЯ НАБОРА ДАННЫХ

На примере сбора набора данных РГ ОГК с легочными узлами (в таблице 11) подробно описаны процесс, возникшие трудности и способы их преодоления.

Таблица 11 – Процесс формирования набора данных РГ ОГК с легочными узлами

№	Процесс	Описание процесса	Возникшие трудности
1	**Формулирование клинической и практической задачи, которую потенциально можно решить с помощью ПО на основе ТИИ**	Рентгенография органов грудной клетки – одно из наиболее распространенных исследований в лучевой диагностике, которое позволяет визуализировать легочные узлы. При их обнаружении требуется дифференциальная диагностика с ЗНО легких, в том числе с раком легкого. **Клиническая задача:** Увеличить долю диагностики новообразований легких на начальных стадиях, что поможет увеличить шансы наступления благоприятного исхода для пациентов. **Практические задачи:** 1) оценить диагностическую точность ПО на основе ТИИ для выявления легочных узлов на РГ ОГК; 2) определить диагностическую точность и согласованность врачей-рентгенологов при работе с рентгенологическими исследованиями (целевая патология – легочные узлы); 3) оценить целесообразность использования ПО на основе ТИИ для увеличения диагностической точности выявления легочных узлов на РГ ОГК	Нет
2	**Описание технических требований**	Подготовить набор данных из 100 исследований РГ органов грудной клетки: – 50 исследований с разным уровнем выраженности патологических изменений (наличие визуализируемых патологических изменений на РГ, подтвержденных на КТ ОГК); – 50 исследований с отсутствием патологических изменений на РГ: 1) 25 исследований – «легкая» норма (на РГ ОГК отсутствие изменений, с подтверждением отсутствия патологических изменений на КТ ОГК); 2) 25 исследований – «сложная» норма (на РГ ОГК наличие изменений, симулирующих патологию, с подтверждением отсутствия патологических изменений на КТ ОГК)	В технические требования стоило заложить запасные исследования (+10 % для каждой группы), чтобы при обнаружении дефектов можно было осуществить замену, а не выполнять заново все этапы

Продолжение таблицы 11

№	Процесс	Описание процесса			Возникшие трудности
3	**Выгрузка (согласно ТЗ) исследований из ЕРИС ЕМИАС**	Был выгружен перечень пар исследований пациентов с КТ ОГК и РГ из ЕРИС ЕМИАС в формате csv, где присутствовали столбцы: 1) с Ф.И.О. пациента и датой его рождения (для упрощения идентификации); 2) дата проведения РГ ОГК; 3) дата проведения КТ ОГК. Были отобраны пары исследований, разница между КТ и РГ ≤14 дней. КТ ОГК использовалась в качестве метода верификации			Нет
4	**Дополнение выгрузки исследований текстовыми протоколами**	Этот этап проводится только после этапа 3 (предварительного отбора) из-за большого объема первоначальной выгрузки			Нет
5	**Фильтрация по целевой модальности**	В столбце «услуга» были выбраны КТ ОГК и РГ ОГК соответственно			Для некоторых исследований тип медицинской услуги изначально был указан некорректно
6	**Фильтрация текстовых протоколов исследований**	Патология	«Сложная» Норма	«Легкая» норма	
		Отбор протоколов, содержащих ключевые слова для классов «патология» и «сложная норма» или с отсутствием ключевых слов для класса «легкая норма»			
7	**Вычитка текстовых протоколов**	**Дополнительный отбор по словам:** консультация онколога, КТ0, периферическое образование, очаговые изменения, признаки единичных очагов, многочисленных очагов, солидные узлы, новообразования **Исключаем:** резекция легкого, туберкулёма, плевральный выпот, пневмония, киста, фиброз,	Не проводилась	Не проводилась	В процессе вычитки протоколов проводилась корректировка ТЗ, возникали вопросы о включении/ невключении, например, исследований с признаками туберкулезных изменений. Необходимо участие врача-рентгенолога

Продолжение таблицы 11

№	Процесс	Описание процесса			Возникшие трудности
		посттуберкулезные изменения, плеврит, не выявлено, норма, признаки центральных образований легких			
		Предварительно отобранные исследования были направлены врачам-рентгенологам для разметки			
8	**Визуальный просмотр исследований 3 экспертами**	**Включаем:** солидный узел визуализируется на РГ ОГК. **Исключаем:** образование меньше 6 мм, больше 30 мм (по КТ ОГК) (рис.1)	на РГ ОГК наличие изменений, симулирующих патологию, с подтверждением отсутствия патологических изменений на КТ ОГК	на РГ ОГК отсутствие изменений, с подтверждением отсутствия патологических изменений на КТ ОГК	Некоторые исследования отсутствовали в системе, их пришлось исключать
		56 исследований	30 исследований	30 исследований	
9	**Заполнение таблицы**	Составление итоговой таблицы с УИД: 100 исследований РГ ОГК и мультиклассовой разметкой (патология, легкая норма, сложная норма)			Исключение пациентов младше 18 лет на момент исследования, использование запасных исследований
10	**Выгрузка анонимизированных исследований в DICOM-форматеи их пересмотр**	Поскольку алгоритмы анализируют лишь прямую проекцию РГ ОГК, вручную производился просмотр и удаление боковых проекций			В исследованиях также присутствовали серии, обработанные ПО на основе ТИИ. Для высокого качества выгрузки необходим пересмотр
11	**Заполнение документации**	Внесение информации в реестр. Формирование README-файла			Нет
12	**Регистрация РИД**	Заполнение регистрационных форм			Нет

Продолжение таблицы 11

Итог	Был получен набор 100 РГ ОГК: – 50 исследований с патологическими изменениями; – 25 исследований с «легкой» нормой; – 25 исследований со «сложной» нормой; – таблица с мультиклассовой разметкой	Нет

На рисунке 14 представлен пример визуального соответствия легочного узла на РГ ОГК и на КТ ОГК.

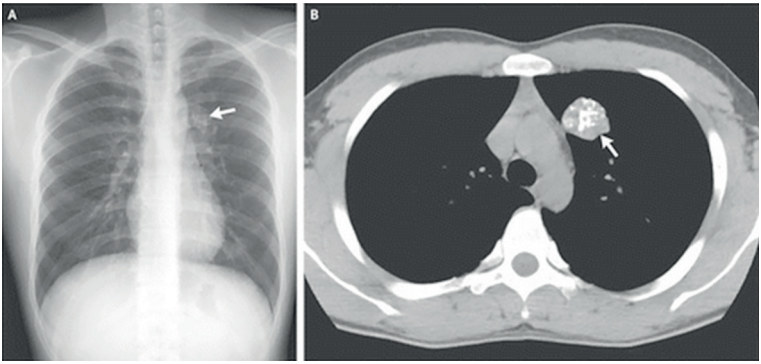


Рисунок 14 – Пример визуального соответствия легочного узла на РГ ОГК с КТ ОГК

ЗАКЛЮЧЕНИЕ

Систематизация накопленного научно-практического опыта создания собственных, валидации сторонних, а также применения наборов данных для тестирования алгоритмов на основе ТИИ позволила создать данное учебно-методическое пособие. Его цель – описать процесс и методологию подготовки качественных, репрезентативных, эталонных НД, которые, в зависимости от цели создания НД, позволят создавать, обучать ПО на основе ТИИ, проводить его качественную оценку и решать другие разнообразные научные задачи. От качества НД зависит качество обученного на нем ПО на основе ТИИ.

В настоящем пособии представлены успешные практики управления жизненным циклом НД, освоив которые, можно создавать качественные НД. Для эффективного освоения материала рекомендуется подготовить НД самостоятельно, используя исследования, находящиеся в свободном доступе, или синтетические данные (https://mosmed.ai/datasets; https://ai2.rt-eu.ru).

В данном учебно-методическом пособии приведены основные этапы создания НД на примере лучевых исследований. Описанный подход позволяет подготовить НД для любого направления инструментальной диагностики, однако при этом возможен ряд трудностей и ограничений, связанных со спецификой и метода диагностики, и МИС, в которой находятся эти данные. Предложенные методы могут быть также экстраполированы и на лабораторную, патоморфологическую, инструментальную диагностику, однако тоже требуют корректировки с учетом специфики этого направления.

СПИСОК ЛИТЕРАТУРЫ

1. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/563441794.

2. ГОСТ Р 52653-2006. Информационно-коммуникационные технологии в образовании. Термины и определения // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200053103.

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

4. Клинические испытания программного обеспечения на основе интеллектуальных технологий (лучевая диагностика) / сост. С. П. Морозов, А. В. Владзимирский, В. Г. Кляшторный [и др.]. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2019.

5. Food and Drug Administration (FDA). Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in-Premarket Notification (510(k)) Submissions Guidance for Industry and FDA Staff Preface Public Comment. 2012. URL: https://www.fda.gov/media/77642/download (дата обращения: 21.04.2023)

6. Lo A. W. Data rich Reinventing Capitalism in the Age of Big Data. Viktor. Mayer-Schönberger and Thomas Ramge Basic Books, 2018. 283 p. // Science. 2018. Vol. 359, № 6376.

7. ГОСТ Р 59921.1–2022. Системы искусственного интеллекта в клинической медицине. Ч.1. Клиническая оценка // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200189205.

8. Регламент подготовки наборов данных с описанием подходов к формированию репрезентативной выборки данных. Ч. 1 / сост. С. П. Морозов, А. В. Владзимирский, А. Е. Андрейченко [и др.]: методические рекомендации. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2022. 40 с.

9. ГОСТ Р 59921.3–2022. Системы искусственного интеллекта в клинической медицине. Ч. 3. Управление изменениями в системах искусственного интеллекта с непрерывным обучением // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200181992.

10. ГОСТ Р 59921.4–2022. Системы искусственного интеллекта в клинической медицине. Ч. 4. Оценка и контроль эксплуатационных параметров // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200181993.

11. ГОСТ Р 59921.5–2022. Системы искусственного интеллекта в клинической медицине. Ч. 5. Требования к структуре и порядку применения набора

данных для обучения и тестирования алгоритмов. М.: Российский институт стандартизации, 2022.

12. ГОСТ Р 59921.6–2022. Системы искусственного интеллекта в клинической медицине. Ч. 6. Общие требования к эксплуатации // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200182011.

13. ГОСТ Р 59921.7 – 2022 Системы искусственного интеллекта в клинической медицине. Ч. 7. Алгоритмы анализа медицинских изображений. Методы испытаний. Общие требования // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200193728.

14. Базовые рекомендации к работе сервисов искусственного интеллекта для лучевой диагностики: методические рекомендации / сост. С. П. Морозов, Л. Р. Абуладзе, А. Е. Андрейченко [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». Вып. 119. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2022. 68 с.

15. ГОСТ Р 8.736-2011. Государственная система обеспечения единства измерений. Измерения прямые многократные. Методы обработки результатов измерений. Основные положения // Электронный фонд правовых и нормативно-технических документов. URL: https://docs.cntd.ru/document/1200089016.

16. Methodological guidelines and recommendations for efficient and rational governance of patient registries / ed. by M. Zaletel, M. Kralj. Ljubljana, 2015. URL: https://ec.europa.eu.

Приложение А

ПЛАТФОРМА ПРЕДВАРИТЕЛЬНОГО ТЕСТИРОВАНИЯ МЕДИЦИНСКИХ СПЕЦИАЛИСТОВ И ЭКСПЕРТОВ

Веб-интерфейс предназначен для определения диагностической точности рентгенологов. В основе графического интерфейса лежит Osimis Viewer, позволяющий визуализировать медицинские изображения в формате DICOM.

Дополнительным функционалом настоящей платформы является возможность осуществления независимой и автоматизированной экспертной разметки исследований на наличие патологии.

Основные этапы взаимодействия с платформой:

1. Врач-рентгенолог регистрируется на платформе через форму регистрации (все данные записываются в базу данных Users).

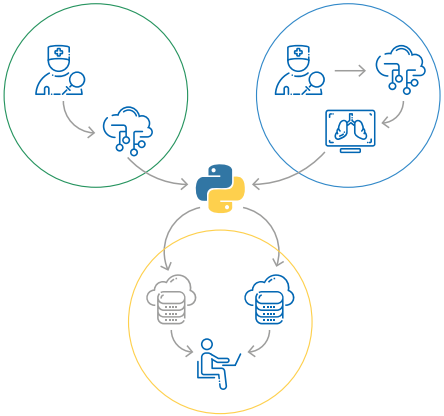
2. После регистрации эксперт авторизуется на платформе и последовательно открывает исследования. Необходимо определить наличие патологии на медицинском изображении из установленного перечня:

- Определенно без патологии.
- Возможно без патологии.
- Затрудняюсь ответить.
- Возможно с патологией.
- Определенно с патологией.

(Представленный перечень может быть изменен в зависимости от задачи.)

3. После оценки экспертом всех исследований в базе данных выполняется автоматический расчет метрик диагностической точности, а также затраченного на разметку времени. По завершении данные сохраняются и направляются врачу.

Для того чтобы начать взаимодействовать с платформой, необходимо зарегистрироваться (рисунки А.1, А.2).



Врач-рентгенолог

Инженер

Backend/Python – с POST и Get запросами в/из базы данных

OsimisViewer – просмотрщик DICOM файлов

Веб платформа

БД с регистрационными данными и расчитанными AUC для каждого зарегистрированного пользователя

БД с метриками точности по каждому исследованию в формате DICOM

*Все БД выгружаются по ссылке напрямую из MONGO express в формате теблицы csv

Рисунок А.1 – Регистрация на платформе

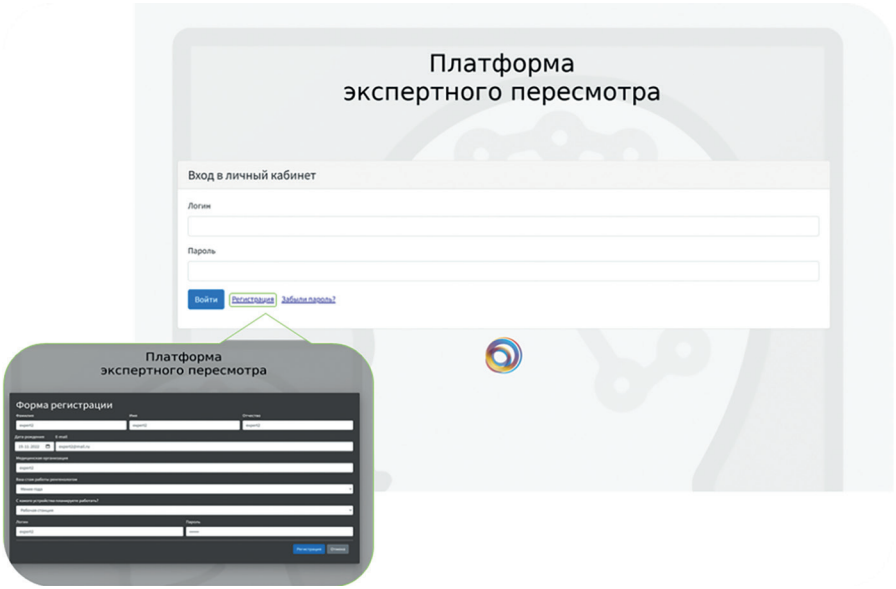


Рисунок А.2 – Окно регистрации и входа на платформу

1. Откройте окно «Регистрация» в меню титульной страницы платформы.
2. Заполните все поля своими персональными данными, а также создайте себе логин и пароль для входа в систему.
3. После подтверждения данных ваша учетная запись будет создана, а за вами – закреплен набор исследований для разметки.
4. Для того чтобы начать разметку, введите созданный вами логин и пароль в соответствующие поля формы входа.
5. Вы допущены к работе на платформе, можете переходить на следующий этап.

После аутентификации пользователь оказывается на главной странице платформы. Здесь описаны все основные положения, требования к обеспечению, правила разметки и тип подгруженных исследований (рисунок А.3).

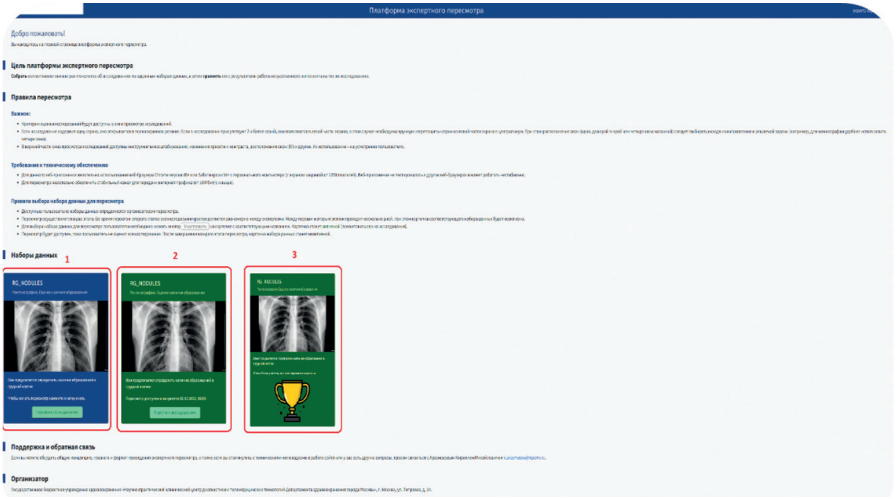


Рисунок А.3 – Окно домашней страницы с описанием функционала и требований по работе с платформой

Также на странице отображается статус работы пользователя с платформой:

1. Открытое окно к полному набору данных.
2. Переход к неразмеченной части исследований.
3. Разметка исследований завершена.

Для того чтобы начать оценку исследований на наличие патологии, необходимо нажать на кнопку «Перейти к исследованиям».

Страница с основным функционалом платформы – разметчик исследований.

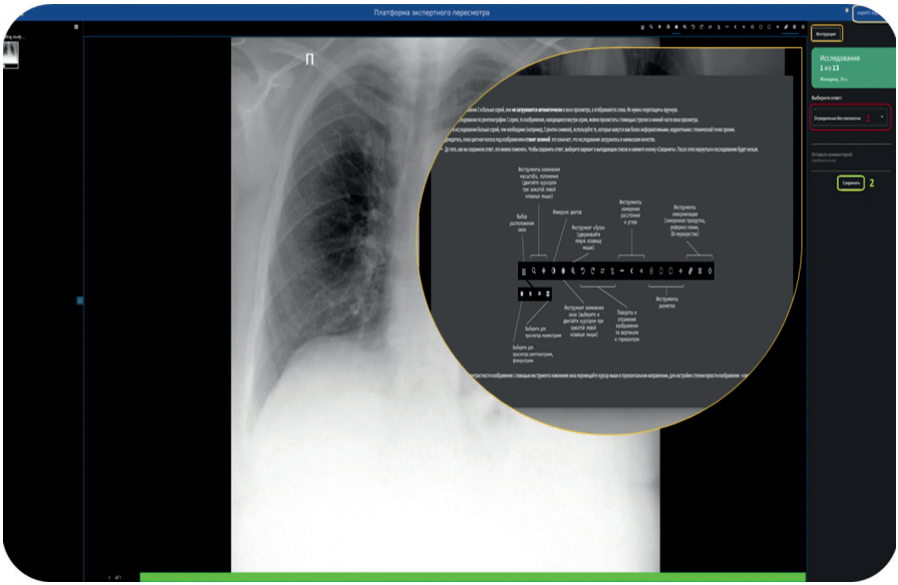


Рисунок А.4 – Окно платформы для разметки исследований

Инструкция:

– Если в исследовании 2-й и более серии они не загружаются автоматически в окно просмотра, а отображаются слева, то их нужно «перетащить» вручную.

– Если в исследовании по рентгенографии 1 серия, то изображения, находящиеся внутри серии, можно пролистать с помощью стрелок в нижней части окна просмотра.

– Если в исследовании больше серий, чем необходимо (например, 5 изображений), используйте те, которые кажутся вам более информативными, корректными с технической точки зрения.

– Дождитесь, пока цветная полоса под изображением станет зеленой: это означает, что исследование загрузилось в наивысшем качестве.

– Для настройки контрастности изображения с помощью инструмента изменения окна перемещайте курсор мыши в горизонтальном направлении, для настройки степени яркости изображения – в вертикальном направлении.

– До того, как вы сохранили ответ, его можно поменять. Чтобы сохранить ответ, выберите вариант в выпадающем списке и нажмите кнопку «Сохранить». После этого вернуться к исследованию будет нельзя.

– Также в этом окне есть кнопки выхода и возвращения во вкладку

«Домашняя страница» в левом и правом углах страницы. Полная инструкция по работе с разметчиком – кнопка «Инструкция».

– Для выгрузки данных о пользователях и разметки набора данных необходимо перейти по закрытому адресу в базу данных (доступ есть только у администратора платформы). В вкладке reader_study находятся 4 базы данных (рисунок А.5).

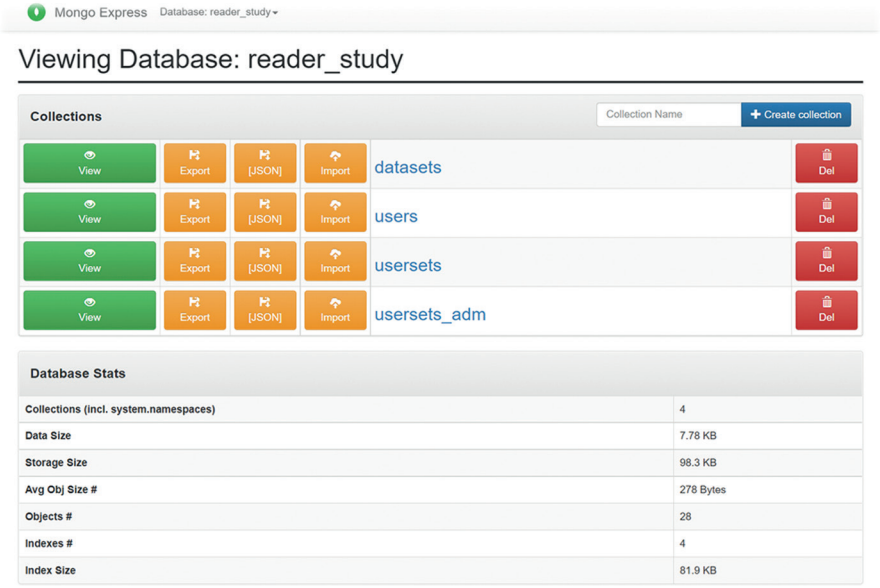


Рисунок А.5 – Структура сформированных баз данных

– Для получения персональных данных пользователя и их рассчитанных параметров AUC необходимо экспортировать базу данных Users (данные выгружаются в формате .cvs).

– Для выгрузки суммарной разметки исследований необходимо экспортировать базу данных Usersets (данные выгружаются в формате .cvs).

Приложение Б

ПРИМЕР СТРУКТУРЫ README-ФАЙЛА

ЦЕНТР ДИАГНОСТИКИ
И ТЕЛЕМЕДИЦИНЫ

MosMedData МРТ с признаками интракраниальных образований тип III

> Набор данных содержит результаты магнитно-резонансной томографии головы с признаками интракраниальных образований, а также без признаков (норма). Данные исследования были собраны в отделениях лучевой диагностики лечебных учреждений города Москвы в период с 11.11.2020 по 14.06.2022.

Disclaimer

> Набор данных предназначен для следующих целей:
>
> - разработка, дообучение и тестирование программных продуктов (использующих в том числе методы компьютерного зрения), выявляющих признаки, характерные для интракраниальных образований;
> - информирование медицинского сообщества и общественности в целом.
>
> Лицензия **позволяет свободно делиться (обмениваться)** набором данных, то есть копировать и распространять материал на любом носителе и в любом формате, **при обязательном соблюдении следующих условий**:
>
> - указано авторство, а именно:
> - авторы;
> - их организации;
> - правообладатель (копирайт);
> - постоянная ссылка на оригинальный набор данных.
> - указана ссылка на лицензию.

Лицензия **запрещает**, в том числе:

> - использовать набор данных в коммерческих целях;
> - распространять переработанный, преобразованный набор данных или новые наборы данных, созданные на основе этого набора;
> - накладывать ограничения поверх существующих ограничений, указанных в лицензии, например:
> - предоставлять платный доступ к набору данных;
> - искусственно сдерживать распространение набора данных техническими методами.

Общая информация

Название набора данных

MosMedData МРТ с признаками интракраниальных образований тип III

Внутренний код

BRAINCR

Классы разметки

3-C

Ключевые слова

MosMedData, тестирование, искусственный интеллект, МРТ, образование, опухоль, рак

Язык

Английский, русский

Финансирование

Внутреннее финансирование

Версия набора данных

1.0

Постоянная ссылка

https://mosmed.ai/datasets/

Дата публикации

15.12.2022

Аффилиация и авторы

Авторы

- Васильев Ю.А. [1]
- Владзимирский А.В. [1]

Аффилиация

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

Структура набора данных

```
.
|-- dataset_registry.xlsx
|-- README_EN.md
|-- README_RU.md
|-- README_EN.pdf
|-- README_RU.pdf
`-- studies
    |--studyUID_X
    |    |-- seriesUID_X
    |    |    |-- UID_X.dcm
    |    |    |-- UID_X.dcm
    |    |    `-- ...
    |    |
    |    `-- seriesUID_X
    |        |-- UID_X.dcm
    |        |-- UID_X.dcm
    |        `-- ...
    |--studyUID_X
    |    |-- seriesUID_X
    |    |    |-- UID_X.dcm
    |    |    |-- UID_X.dcm
    |    |    `-- ...
    |    |
    |    `-- seriesUID_X
    |        |-- UID_X.dcm
    |        |-- UID_X.dcm
    |        `-- ...
    `--...
```

- README_EN.md и README_RU.md содержат общую информацию о наборе данных в формате Markdown на английском и русском языках соответственно; та же информация в формате PDF представлена в README_EN.pdf и README_RU.pdf.
- dataset_registry.xlsx содержит перечень исследований, включенных в набор данных, путь к соответствующему файлу.
- В директории studies находятся директории studyUID-X в каждой из которых содержатся исследования в формате DICOM

Обзор данных

Параметр	Значение
Количество исследований, ед.	6
Количество пациентов, чел.	6
Распределение по полу, чел. (М/ Ж)	2/ 4
Распределение по возрасту, лет (мин./ медиана/ макс.)	33/ 50/ 69
Распределение по классам, ед. (С патологией/ Без патологии)	3/ 3

Правила использования и распространения

Лицензия

Цитирование

Рекомендованная форма для цитирования:

Наборы данных Центр диагностики и телемедицины ДЗМ [Электронный ресурс]. – 2022. – URL: https://mosmed.ai/datasets/

Распространение

Данный набор данных не должен распространяться без указания:

- авторов;
- аффилиаций;
- правообладателя (копирайта);
- постоянной ссылки на оригинальный набор данных;
- ссылки на лицензию.

Приложение В

ИНСТРУМЕНТЫ РАЗМЕТКИ

Инструменты разметки можно условно разделить на инструменты для **локализации** исследуемых структур и **сегментации**.

Для локализации исследуемых структур можно использовать следующее программное обеспечение:

1. Supervisely (коммерческое ПО производства эстонской компании Supervisely OÜ).
2. CVAT (открытое программное обеспечение с клиент-серверной архитектурой).

Программное обеспечение Supervisely является универсальным программным обеспечением, позволяющим производить разметку медицинских изображений. Основные характеристики данного ПО представлены в таблице В.1.

Таблица В.1 – Основные характеристики ПО Supervisely

Параметр	Данные
Тип ПО	Коммерческое программное обеспечение
Ссылки	https://supervise.ly/
Поддерживаемые платформы	Сервер - Linux, клиент – любая ОС с браузером
Расширяемость, возможность добавления новых модулей	Есть
Возможность локальной установки	Есть в расширенной версии
Вид приложения:	Web
Менеджмент процесса разметки	Да
Уровень сложности установки	Для бесплатной версии установка не требуется

Основные поддерживаемые форматы изображений: DICOM и NRRD. Загрузка медицинских изображений происходит через модуль загрузки типов DICOM и NRRD с локального компьютера. Области сегментации сохраняются в виде архива, содержащего исходное медицинское изображение в формате NRRD, внутренний файл Supervisely – с сохранением информации об объектах, растровый файл – с локализацией типа NRRD.

Области сегментации в Supervisely представляют собой объекты на основе классов. Первоначально создаются классы, в основе которых тип сегментации Polygon, Rectangle и Bitmap. При проведении разметки медицинских

изображений с локализацией используются два класса Polygon, Rectangle. Благодаря такой структуре сохраняется не только растровая информация, но и векторная информация об областях. Сохраненные области сегментации-полигоны или прямоугольники могут быть в дальнейшем отредактированы изменением положения точек контуров или добавлением к контурам новых точек.

При работе с медицинскими изображениями Supervisely имеет стандартные и пользовательские настройки окон просмотра. В Supervisely отсутствует 3D-визуализация областей сегментации и исходного исследования, но присутствует окно перспективы. Возможно использование до 16 окон просмотра медицинских изображений одновременно.

ПО Supervisely предоставляет пользователю различные инструменты ручной и полуавтоматической разметки медицинских изображений.

Выделяют следующие достоинства и недостатки ПО Supervisely:

Достоинства	Платформа позволяет создавать собственные настраиваемые решения для выполнения задач компьютерного зрения. Инструмент удобной разметки может быть эффективно использован в случае, когда размечаемая структура находится в небольшом количестве срезов и обладает достаточно отделяемой структурой. Удобный способ хранения сегментаций позволяет в дальнейшем, если понадобится, менять полигоны (например, если один эксперт делает правки в размеченной другим экспертом области сегментации).
Недостатки	Отсутствие 3D-инструментов. Стоимость расширенного продукта. Не всегда удобно применять инструмент «умной разметки».

Открытое программное обеспечение CVAT является ПО с клиент-серверной архитектурой, позволяющей работать с данными, размещенными на серверах CVAT и на локальных ПК. Позволяет проводить ручную и полуавтоматическую разметку изображений и видео. Обеспечивает совместную работу экспертов над аннотацией изображений и содержит функционал для менеджмента процессом разметки. В таблице В.2 представлены общие характеристики CVAT.

Таблица В.2 – Общие характеристики ПО CVAT

Параметр	Данные
Тип ПО	Бесплатное программное обеспечение с открытым исходным кодом
Ссылки	https://www.cvat.ai/
Поддерживаемые платформы	Любая ОС
Расширяемость, возможность добавления новых модулей	Нет
Возможность локальной установки	Есть
Вид приложения:	Web
Менеджмент процесса разметки	Есть
Уровень сложности установки	Высокий

CVAT не позволяет работать с медицинскими изображениями в формате DICOM-файлов напрямую. Пользователям CVAT необходимо предварительно переводить изображения DICOM в графический формат PNG посредством самостоятельно написанного скрипта.

Загрузка изображений в CVAT может быть осуществлена различными вариантами:

1. С локального компьютера.
2. Различные варианты удаленных хранилищ. Разметка может быть сохранена в различных форматах в зависимости от задачи (ключевые точки, детекция объектов, сегментация) и от формата сохранения. Например, в формате CVAT for images сохраняются в виде объекта XML, где для каждого исходного изображения сохраняются объекты, аналогичные объектам при создании разметки (эллипсы, полилинии, полигоны). Данные типа MASK SHAPE внутри XML сохраняются в формате RLE.

Существуют другие форматы для сохранения данных сегментации: Segmentation mask 1.1 (маски сохраняются в виде файлов PNG), TFRecord, ICDAR Segmentation и другие. Формат выбирается в зависимости от дальнейшего использования НД.

CVAT предоставляет пользователю различные инструменты ручной и автоматизированной разметки изображений. В результате анализа CVAT были установлены следующие достоинства и недостатки:

Достоинства Удобный интерфейс для 2D-методов ручной разметки.

Недостатки Сложность установки, невозможность работы непосредственно с форматами медицинских изображений и вытекающие из этого другие трудности, невозможность менять окна отображения динамически, просматривать интенсивности и т.д.

Инструменты разметки для сегментации изображений.

Самым сложным и самым трудоемким типом разметки является попиксельная локализация (сегментация) медицинских изображений. Особенно это касается таких модальностей исследований, как КТ и МРТ, с большим количеством снимков.

Для проведения сегментации медицинских изображений наиболее популярными платформами являются:

- 3D Slicer,
- ITK Snap,
- Supervisely,
- MITK,
- MedSeg,
- CVAT.

Каждая из платформ имеет свои положительные и отрицательные стороны: так 3D-SLICER является программным продуктом широкого профиля, который позволяет обрабатывать многомерные медицинские изображения и обладает наиболее обширным функционалом по сравнению с другими программными продуктами. Общие характеристики 3D-SLICER представлены в таблице В.3.

Таблица В.3 – Общие характеристики 3D-SLICER

Параметр	Данные
Тип ПО	Бесплатное программное обеспечение с открытым исходным кодом
Ссылки	https://www.slicer.org/
Поддерживаемые платформы	Любая ОС
Виды медицинских данных	МРТ, КТ, УЗИ, РГ, ядерная медицина и микроскопия
Расширяемость, возможность добавления новых модулей	Есть
Возможность локальной установки	Есть
Вид приложения:	Desktop

Продолжение таблицы В.3

Менеджмент процесса разметки	Нет
Уровень сложности установки	Низкий

3D-SLICER поддерживает широкий спектр форматов изображений, в том числе: DICOM, NRRD, NIfTI, MetaImage, Analyze, VTK и др.

3D-SLICER позволяет работать с медицинскими изображениями, размещенными локально и на удаленных серверах (хранилищах). 3D-SLICER обладает широкими возможностями визуализации медицинских изображений, а именно:

– Стандартные и пользовательские настройки окон просмотра изображений.
– Множество комбинаций окон просмотра.
– 3D-визуализация областей сегментации и исходного исследования.
– Возможность поворота осей.
– Настройка контраста по области.

3D-SLICER позволяет создавать несколько файлов сегментаций на одно исходное изображение, при этом каждая сегментация может содержать несколько слоев (соответствующих определенным меткам).

3D-SLICER обладает широким набором инструментов для проведения ручной, полуавтоматической и автоматической сегментации изображений.

Результаты сегментации изображений могут быть сохранены в виде поверхности-сетки (STL, OBJ), в виде маски (NRRD, NIfTI) и других форматах. Доступно сохранение меток в формате JSON, что позволяет сохранять различные сегментации в разные файлы.

В процессе исследования ПО 3D-SLICER были выявлены следующие достоинства и недостатки:

Достоинства — Наиболее полный функционал, простота установки программы и расширений, модульность, многие методы работают и не требуют дополнительной настройки параметров.

Недостатки — Средний порог сложности вхождения для пользователя, нестабильность работы

ITK Snap – программный продукт, специально разработанный для сегментации структур в многомерных медицинских изображениях. Общие характеристики ITK Snap представлены в таблице В. 4.

Таблица В. 4 – Общие характеристики ITK Snap

Параметр	Данные
Тип ПО	Бесплатное программное обеспечение с открытым исходным кодом
Ссылки	http://www.itksnap.org/pmwiki/pmwiki.php
Поддерживаемые платформы	Любая ОС
Расширяемость, возможность добавления новых модулей	нет
Возможность локальной установки	есть
Вид приложения:	Desktop
Менеджмент процесса разметки	нет
Уровень сложности установки	Низкий

ITK Snap поддерживает следующие форматы изображений: DICOM, GE, GIPL, VTK, MetaImage, NIfTI, NRRD, Analyze, MINC и др. Для просмотра и первичной корректировки медицинских изображений доступен следующий функционал:

– Изменение контраста происходит через задание диапазонов. Можно задать функцию преобразования из единиц HU в интенсивность цвета. Нет стандартных окон для преобразования в интенсивности цвета.

– Варианты просмотра: отдельная проекция, 3D-окно или все окна одновременно.

– Возможность 3D-визуализации исходного изображения и областей сегментации.

– ITK Snap позволяет работать с изображениями, размещенными только локально. При проведении сегментации изображений доступен следующий функционал:

– несколько сегментаций на одно изображение, каждая сегментация содержит несколько слоев с метками. Список меток одинаковый для всех сегментаций;

– один воксель может содержать одну метку;

– можно просматривать на одном изображении только одну сегментацию.

Сегментация изображения может производиться в ручном и полуавтоматическом режимах.

ITK Snap позволяет сохранять область сегментации в формате поверхности-сетки (STL), в растровых форматах NRRD, NIfTI, VTK, GIPL, MetaImage, Analyze и др. Сохранение разных сегментаций происходит в разные файлы.

В процессе исследования специального программного обеспечения были выявлены следующие достоинства и недостатки ITK Snap:

Достоинства — Понятный интерфейс, не перегруженный не относящимся к сегментации функционалом. Наличие 3D-окна визуализации исходного изображения и областей сегментации.

Недостатки — Методы полуавтоматической сегментации требуют настройки неочевидных параметров. Не всегда удается добиться качественной разметки.

Программный продукт Supervisely позволяет проводить сегментацию изображений. Обзорное рассмотрение данного ПО было представлено в части, посвященной локализации структур на изображении.

MEDSEG – программный продукт с простым ограниченным функционалом и ручными инструментами. Содержит множество AI моделей различной эффективности для сегментации различных органов и некоторых патологий. Общие характеристики MEDSEG представлены в таблице В. 5.

Таблица В.5 – Общие характеристики ПО MEDSEG

Параметр	Данные
Тип ПО	Бесплатное программное обеспечение
Ссылки	https://www.medseg.ai/
Поддерживаемые платформы	Любая ОС
Расширяемость, возможность добавления новых модулей	Разработчики MedSeg готовы обучить новые модели на ваших данных, при условии, что данные будут выложены в открытый доступ
Возможность локальной установки	Нет. Разработчиками позиционируется, что вся обработка происходит на вашем локальном компьютере, и данные не уходят за пределы вашего ПК
Вид приложения:	Web
Менеджмент процесса разметки	Да
Уровень сложности установки	Установка не требуется

MEDSEG поддерживает следующие форматы медицинских изображений: NIfTI, DICOM (набор файлов для исследований, содержащих большое количество срезов). Для просмотра изображений предоставляется простой интерфейс

с возможностью переключения между окнами просмотра медицинского изображения.

В ПО MEDSEG загрузка исследований для сегментации производится с локального компьютера. Сохранение областей сегментаций – в растровом формате (NIfTI).

В процессе исследования ПО MEDSEG, применяемого для разметки медицинских изображений, были выявлены следующие достоинства и недостатки:

Достоинства Программа обладает простым интерфейсом. Некоторые модели автоматической сегментации неплохо справляются со своими задачами.

Недостатки Существуют алгоритмы только под определенные задачи, в основном связанные с сегментацией органов. Если ПК, где запускается Medseg, не имеет GPU, выполнение сегментации может быть затруднено. Моделей для сегментации патологий мало.

MITK (Medical Imaging Interaction Toolkit) – программный продукт на основе библиотеки с открытым исходным кодом. По набору функционала и интерфейсу имеет сходство с 3D Slicer. Общие характеристики ПО MITK представлены в таблице В. 6.

Таблица В. 6 – Общие характеристики MITK

Параметр	Данные
Тип ПО	Бесплатное программное обеспечение с открытым исходным кодом
Ссылки	http://www.mitk.org/
Поддерживаемые платформы	Любая ОС
Расширяемость, возможность добавления новых модулей	Есть
Возможность локальной установки	Есть
Вид приложения:	Desktop
Менеджмент процесса разметки	Нет
Уровень сложности установки	Легкий

MITK поддерживает 2D-images/3D-volumes-форматы исследований: DICOM, NRRD, GIPL, NIfTI, OBJ, STL, VTK и другие. MITK предоставляет следующие возможности визуализации:

- стандартные и пользовательские настройки окон просмотра;
- 3D-визуализация областей сегментации и исходного исследования;
- изменение контраста;
- разные варианты комбинации окон просмотра.

MITK обеспечивает загрузку файлов формата DCM через DICOM-браузер, они могут быть загружены из удаленных хранилищ или с локального компьютера. Файлы остальных форматов могут быть загружены с локального компьютера. MITK позволяет производить разметку медицинских изображений ручным и полуавтоматическим способами.

ПО разрешает одновременно загружать несколько исходных изображений (созданы слои с сегментациями для каждого из них). По окончании работы сегментированное изображение сохраняется послойно, либо все слои данного исследования сохраняются в форматах NRRD, NIfTI и других стандартных форматах.

В результате изучения ПО были выявлены следующие достоинства и недостатки MITK.

Достоинства	Достаточно понятный и неперегруженный лишними функциями интерфейс без ущерба для функциональности. Можно рассматривать данную платформу как дополнение к платформе 3D Slicer.
Недостатки	Не очень удобным является то, что в отличие от 3D Slicer в полуавтоматических методах Fast Matching, Region Growing требуется гибкая настройка неочевидных параметров. В аналогичных методах в ITK Snap формируются вспомогательные графики и так называемые Speed Image-образы, отражающие поле сил, действующих на активные контуры, помогающие эксперту настраивать параметры исходя из сути алгоритма. Нестабильность работы программы.

Каждое медицинское учреждение, задействованное в подготовке НД, может самостоятельно выбирать ПО для проведения разметки медицинских изображений. Однако осуществленные нами исследования показывают, что наиболее универсальным и удобным для проведения разметки медицинских изображений инструментом является 3D Slicer.

ДЛЯ ЗАМЕТОК

Ю. А. Васильев, К. М. Арзамасов, А. В. Владзимирский,
О. В. Омелянская, Т. М. Бобровская, Д. Е. Шарова,
Н. Ю. Никитин, М. Р. Коденко

ПОДГОТОВКА НАБОРА ДАННЫХ ДЛЯ ОБУЧЕНИЯ И ТЕСТИРОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Учебно-методическое пособие

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технический редактор А.И. Овчарова
Компьютерная верстка Е.Д. Бугаенко

ГБУЗ «НПКЦ ДиТ ДЗМ»
127051, г. Москва, ул. Петровка, д. 24, стр. 1

+7 (495) 276-04-36 npcmr@zdrav.mos.ru telemedai.ru