ЭМУ 411КИ

ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

СОГЛАСОВАНО

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

_____ С.П. Морозов

«29» октября 2021 г.

РЕКОМЕНДОВАНО

Экспертным советом по науке Департамента здравоохранения города Москвы № 1

«21» 04 2022 г.

РЕГЛАМЕНТ ПОДГОТОВКИ НАБОРОВ ДАННЫХ С ОПИСАНИЕМ ПОДХОДОВ К ФОРМИРОВАНИЮ РЕПРЕЗЕНТАТИВНОЙ ВЫБОРКИ ДАННЫХ

Часть 1

Методические рекомендации № 1

Москва
2021

УДК 615.84+616-71
ББК 5с51
М 80

Серия «Лучшие практики лучевой и инструментальной диагностики»

Основана в 2021 году

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

Составители:
Морозов С.П. – д.м.н., профессор, главный внештатный специалист по лучевой и инструментальной диагностике ДЗМ и Минздрава России по ЦФО РФ, директор ГБУЗ «НПКЦ ДиТ ДЗМ»
Владзимирский А.В. – д.м.н., заместитель директора по научной работе ГБУЗ «НПКЦ ДиТ ДЗМ»
Андрейченко А.Е. – к.ф.-м.н., руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»
Ахмад Е.С. – младший научный сотрудник сектора стандартизации и контроля качества ГБУЗ «НПКЦ ДиТ ДЗМ»
Блохин И.А. – младший научный сотрудник сектора исследований в лучевой диагностике ГБУЗ «НПКЦ ДиТ ДЗМ»
Гомболевский В.А. – к.м.н., директор ключевых исследовательских программ АНО «Институт искусственного интеллекта»
Зинченко В.В. – начальник сектора клинических и технических испытаний ГБУЗ «НПКЦ ДиТ ДЗМ»
Кульберг Н.С. – к.ф.-м.н., руководитель отдела разработки средств медицинской визуализации ГБУЗ «НПКЦ ДиТ ДЗМ»
Новик В.П. – научный сотрудник отдела разработки средств медицинской визуализации ГБУЗ «НПКЦ ДиТ ДЗМ»
Павлов Н.А. – руководитель проекта ГБУЗ «НПКЦ ДиТ ДЗМ»

М 80 Регламент подготовки наборов данных с описанием подходов к формированию репрезентативной выборки данных. Часть 1: методические рекомендации / сост. С.П. Морозов, А.В. Владзимирский, А.Е. Андрейченко [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. 103. – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», 2021. – 40 с.

Рецензенты:
Кремнева Елена Игоревна – к.м.н., старший научный сотрудник отделения лучевой диагностики ФГБНУ «Научный центр неврологии»
Мищенко Андрей Владимирович – д.м.н., заместитель главного врача ГБУЗ «ГКОБ №1 ДЗМ» по медицинской части

Методические рекомендации предназначены для врачей любых специальностей, организующих и непосредственно проводящих разметку медицинских наборов данных.

Методические рекомендации разработаны в ходе выполнения научно-исследовательской работы «Научное обоснование методологии применения и способов оценки качества интеллектуальных технологий (искусственного интеллекта) в диагностике»

Данный документ является собственностью Департамента здравоохранения города Москвы, не подлежит тиражированию и распространению без соответствующего разрешения

© Департамент здравоохранения города Москвы, 2021
© ГБУЗ «НПКЦ ДиТ ДЗМ», 2021
© Коллектив авторов, 2021

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ

1. ГОСТ ISO 13485-2017 «Системы менеджмента качества. Требования для целей регулирования».
2. Постановление Правительства Москвы от 21.11.2019 № 1543-ПП «О проведении эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы».
3. Приказ Департамента здравоохранения города Москвы от 26.01.2021 №51 «Об утверждении Порядка и условий проведения эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» (в ред. от 30.04.2021 № 413, в ред. от 23.06.2021 г. № 588).
4. Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор) г. Москвы от 05.09.2013 № 996 «Об утверждении требований и методов по обезличиванию персональных данных».
5. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации».
6. Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации».
7. Федеральный закон от 27.07.2006 № 152-ФЗ «О персональных данных».

ОПРЕДЕЛЕНИЯ

В настоящем документе применены следующие термины с соответствующими определениями:

1. **Анонимизация** (обезличивание) – действия, в результате которых удаляется связь между совокупностью идентифицирующих данных и субъектом данных. Необратимо: все атрибуты из записи удаляются либо изменяются таким образом, что выполнить идентификацию субъекта невозможно (необратимое обезличивание).

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

3. **ИИ-сервис** – специальное программное обеспечение на основе алгоритмов искусственного интеллекта (компьютерного зрения) для решения определенной медико-диагностической задачи в лучевой диагностике.

4. **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 то, в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

5. **Метаданные** – информация о наборе данных, являющаяся средством для классификации, упорядочения и описания данных.

6. **Набор данных** – упорядоченная совокупность данных и соответствующих им метаданных, организованных по определенным правилам.

7. **Обратный процесс** – деобезличивание (обратная персонификация) – действия, в результате которых обезличенные данные принимают вид, позволяющий определить их принадлежность конкретному субъекту персональных данных – становятся персональными данными.

8. **Псевдонимизация** – особый случай обезличивания, при котором, помимо удаления прямой связи с субъектом данных, создается связь между конкретной совокупностью характеристик этого субъекта и одним или несколькими псевдонимами.

9. **Разметка (аннотация) 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 технологий искусственного интеллекта.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем документе применены следующие обозначения и сокращения:

1. **ГБУЗ «НПКЦ ДиТ ДЗМ»** – государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»
2. **DICOM** – англ. Digital Imaging and Communications in Medicine (цифровые изображения и передача данных в медицине).
3. **ДЗМ** – Департамент здравоохранения города Москвы.
4. **ЕРИС ЕМИАС** – Единый радиологический информационный сервис Единой медицинской информационно-аналитической системы г. Москвы.
5. **ИИ** – искусственный интеллект.
6. **МР** – методические рекомендации.
7. **FDA** – англ. Food and Drug Administration (Управление по санитарному надзору за качеством пищевых продуктов и медикаментов).
8. **ТЗ** – техническое задание.
9. **GDPR** – англ. General Data Protection Regulation (общие правила защиты данных).
10. **ЭМК** – электронная медицинская карта.
11. **МИС** – медицинская информационная система.
12. **СМК** – система менеджмента качества.

ВВЕДЕНИЕ

Цель данных методических рекомендаций – описать основные этапы процесса формирования наборов медицинских данных для использования алгоритмами машинного обучения с указанием требуемых специалистов, инструментов и инфраструктуры.

Методические рекомендации (далее – МР) обобщают практический опыт ГБУЗ «НПКЦ ДиТ ДЗМ» по формированию наборов медицинских данных для использования при тестировании, валидации и обучении интеллектуальных систем в здравоохранении, в том числе при реализации эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы (Постановление Правительства Москвы от 21.11.2019 № 1543-ПП, приказ Департамента здравоохранения города Москвы от 26.01.2021 № 51 (в ред. от 30.04.2021 № 413, в ред. от 23.06.2021 № 588)). Также МР расширены сводной аналитикой лучших мировых практик планирования, создания и работы с наборами медицинскими данных для применения в областях искусственного интеллекта и машинного обучения. В методических рекомендациях отражены все аспекты работы с медицинскими данными, обеспечивающие соответствие создаваемых наборов данных их целевому назначению и требованиям заинтересованных лиц. Аспекты включают в себя технологические, медицинские и регуляторные (стандарты) особенности подготовки и применения наборов данных для программного обеспечения с применением технологий искусственного интеллекта (далее – ИИ-сервиса) в здравоохранении. МР описывают универсальные подходы и могут быть применены в любой области здравоохранения, актуальной для внедрения ИИ-сервисов.

Назначение и актуальность методических рекомендаций

Одними из перспективных направлений развития современной медицины являются диагностика, лечение и профилактика заболеваний, базирующихся на интеллектуальном машинном анализе большого количества разнообразных данных о пациентах. В подобных целях, а также для стандартизации и повышения точности интерпретации результатов исследований и постановки диагнозов разрабатываются алгоритмы и сервисы, основанные на технологиях искусственного интеллекта. Ключевым фактором успешного развития этих направлений являются качественные наборы данных, собранные и структурированные из больших массивов медицинских данных.

Ежегодно наблюдается увеличение объема первично цифровых (машиночитаемых) медицинских электронных записей (electronic health records),

включающих в себя как клинические данные (результаты осмотров, физикальных исследований), так и результаты лабораторных и инструментальных исследований, в частности, лучевых (магнитно-резонансная, компьютерная томография, рентгенография, маммография, флюорография и др.), сигнальных (электрокардиография, электроэнцефалография, электронейромиография и др.).

Важным аспектом, ограничивающим простое использование накопленных результатов для эффективной разработки искусственного интеллекта (далее – ИИ), является тот факт, что медицинские электронные записи возникают в ходе рутинной медицинской деятельности организаций, а не специально в целях сбора данных для их последующей машинной обработки, что ведет к неструктурированности данных, различающимся форматам представления данных и т. п. Несмотря на ускоренные темпы развития сферы искусственного интеллекта в медицине, процесс сбора медицинских данных в единую структуру, имеющую необходимый набор характеристик и подходящую для дальнейших манипуляций и вычислений, еще не организован.

Ниже перечислены некоторые примеры актуальных проблем в области искусственного интеллекта в медицине:

1. Увеличение объема данных, в частности данных результатов лучевой диагностики, вследствие чего появляется необходимость налаживания процесса сбора и объединения данных в наборы, пригодные для применения в области машинного обучения;

2. Возникновение новых медицинских случаев, не зарегистрированных ранее, на которых нужно реагировать оперативно, создавать новые наборы данных для изучения врачами и исследователями (примером служит выявление новой коронавирусной инфекции COVID-19);

3. Активно создаются и развиваются алгоритмы машинного обучения, для валидации которых требуются эталонные наборы данных;

4. Эталонные наборы данных должны иметь структуру и характеристики, необходимые для возможности применений методов машинного обучения, кроме того, должны соответствовать поставленной задаче как с точки зрения компьютерных наук, так и с точки зрения медицины. Для этого необходимы специалисты, достаточно квалифицированные на междисциплинарном уровне, способные осуществить сбор и подготовку запрашиваемых данных.

Целью настоящих методических рекомендаций является систематическое обобщение мирового практического опыта и собственного опыта ГБУЗ «НПКЦ ДиТ ДЗМ» по подготовке медицинских наборов данных, готовых для разработки и валидации ИИ-сервисов для медицины. Методические рекомендации будут состоят из двух частей: часть 1 (данный документ) посвящена методологическим аспектам подготовки медицинских наборов данных, часть 2 отражает технические аспекты создания наборов данных.

Понятие наборов данных и разметки

Под набором данных (датасетом) понимается структурированный набор информации, объединенный по определенным логическим принципам, пригодный для машинной обработки компьютерными методами анализа данных, который характеризуется четырьмя основными этапами:

1) наличие содержимого (наблюдения, значения, записи, файлы и др.);

2) наличие цели (например, база знаний, использование для определенной задачи);

3) наличие группировки (агрегация и организация содержимого в наборы, коллекции и др.);

4) наличие связанности (отношение к субъекту, интегрированность, логическая коллекция содержимого и т.д.).

Под разметкой (аннотацией данных) понимается 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 систем искусственного интеллекта.

Цели создания набора данных

Формируемые наборы данных для одного и того же направления или клинической/практической задачи могут отличаться по конечным целям их применения. Предлагается следующее разделение наборов данных на классы:

1) селф-тест для проверки на техническое соответствие ИИ-сервиса;

2) тестирование на локальных данных для валидации и калибровки ИИ-сервиса;

3) дообучение для дополнительного обучения готовой модели ИИ;

4) машинное обучение для обучения новых моделей и решения новых клинических задач.

Понятие жизненного цикла набора данных

Модель жизненного цикла набора данных схематично представлена на рисунке 1, она включает в себя основные фазы при планировании, создании, модификации и эксплуатации наборов данных. За основу была взята адаптированная модель CRISP-DM [1] по исследованию данных.

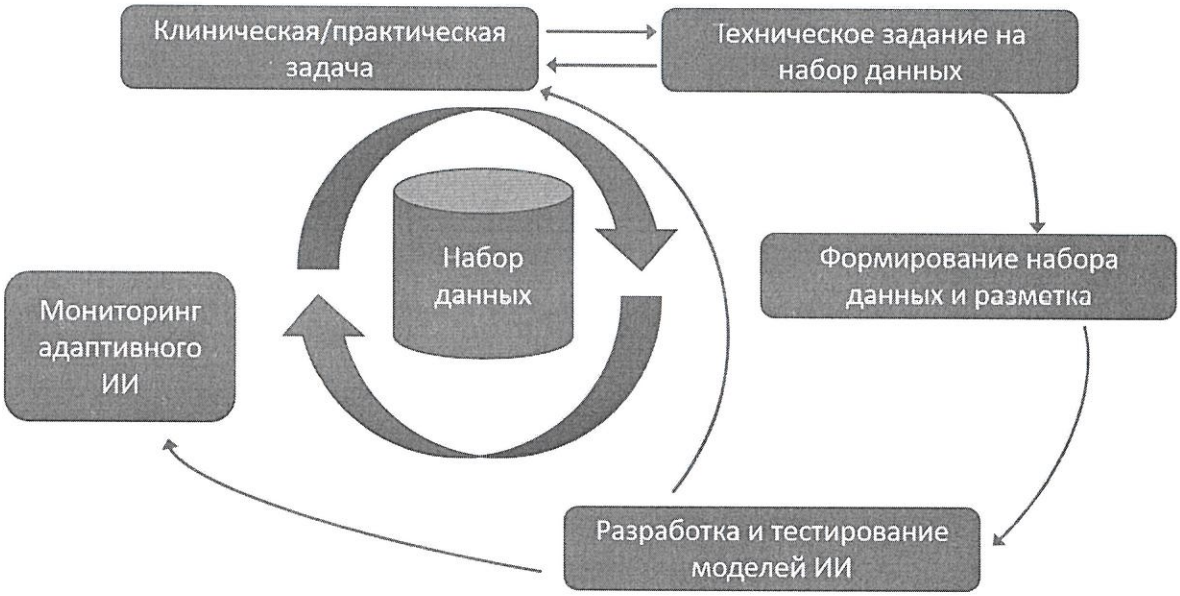

Рисунок 1 – Жизненный цикл набора данных для систем искусственного интеллекта в здравоохранении

В процессе подготовки набора данных следует определить следующие шаги по его развитию: изменению уровня доступа третьим лицам, частоте обновления, сроке поддержки и способе утилизации (уничтожения).

Некоторые категории наборов данных подлежат регулярным обновлениям. Обновления могут касаться как сопроводительной информации (например, при появлении исследований в динамике для верификации), так и самих единиц наборов данных (например, добавление новых случаев в особых эпидемиологических условиях). В этих случаях могут быть отдельно описаны принципы получения новых данных, внесения изменений, в том числе в номер версии.

Набор данных может быть запрограммирован на изменение уровня доступа с закрытого или ограниченного на открытый по прошествии определенного времени (например, 1 календарный год с момента публикации).

Наоборот, у набора данных может быть указан его «срок годности», после которого он должен быть либо скрыт из доступа (из открытого в ограниченный или закрытый, из ограниченного в закрытый), либо заархивирован в длительное хранение без возможности быстрого восстановления.

Бесследное удаление набора данных не приветствуется, так как в будущем может появиться необходимость восстановить источник «потерянных» исследований.

Понятие единицы и верифицированного набора данных

Под единицей набора данных понимается парная запись входных данных для систем ИИ и ожидаемые выходные данные от системы ИИ, после обработки и анализа входных данных (рисунок 2). Ожидаемые выходные данные формируются в процессе разметки, т.е. процесс разметки позволяет сформировать эталонные, «правильные» ответы, по которым затем будут оцениваться ответы от системы ИИ в целях их разработки, тестирования и/или пострегистрационного мониторинга.

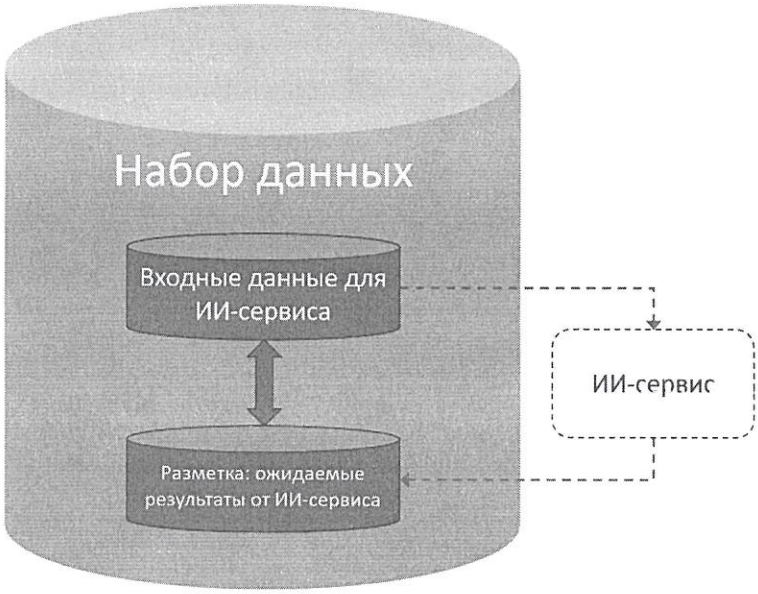

Рисунок 2 – Единица набора данных (в наборе данных присутствуют парные, однозначно соответствующие друг другу записи входных данных и разметки)

Под понятием верификации разметки в наборе данных понимается подтверждение результатов разметки единицы набора данных с помощью более точных диагностических методов и подходов, включающих: альтернативную диагностическую процедуру; клиническое подтверждение диагноза (например, биопсия, специфический лабораторный тест); результаты последующего клинического обследования; диагноз, поставленный клиницистом, и т.п. Метод верификации определяется при планировании набора данных и зависит от задачи, решаемой ИИ-сервисом, для которой такой набор данных применим.

1. ТИПЫ МЕДИЦИНСКИХ ДАННЫХ И ИХ ИСТОЧНИКИ

Медицинские данные пациентов в широком смысле – это любые данные, относящиеся к состоянию здоровья и качеству жизни человека или населения. Данные о здоровье включают клинические показатели, а также экологическую, социально-экономическую и поведенческую информацию, имеющую отношение к здоровью и благополучию.

1.1. Источники медицинских данных

Когда люди взаимодействуют с сервисами системы здравоохранения, собирается и используется множество данных о здоровье. Эти данные, собираемые поставщиками медицинских услуг, обычно включают записи о полученных услугах, условиях оказания этих услуг и клинических исходах.

1.2. Классификация типов медицинской информации

Типы медицинской информации представлены в таблице 1 [2].

Таблица 1 – Основные типы медицинских данных

Типы данных	Описание формата	Основные особенности (сложности использования)
Медицинские записи	Информация в печатных и рукописных документах	Неструктурированные записи на бумажных носителях
Электронная медицинская карта	Медицинская информационная система для сбора, хранения и отображения информации о пациентах	Неструктурированный текст
Лабораторные данные	Программное обеспечение и базы данных, используемые для управления и хранения результатов лабораторных тестов и данных о патологиях: в количественном, качественном и графическом представлениях	Отсутствие стандартизации в сборе, анализе и хранении, а также предоставлении доступа к данным
Медицинские изображения	Медицинские изображения получают для диагностики, определения состояния и планирования лечения. Наиболее распространенные модальности: ПЭТ, КТ, КЛКТ, МРТ и ультразвуковое исследование. Медицинские изображения регулируются общепринятым стандартом DICOM	Недостаточное следование регламентам стандартизации в сборе и анализе, дублирование данных в одном учреждении, доступность данных
Геномика	Отдельные наборы данных с крупномасштабными геномными данными	Неполные данные, доступность данных
Вспомогательные данные	Доход, социально-экономический статус, раса, этническая принадлежность, образование, жилье	Неструктурированные и неполные данные, доступность данных

Лабораторные данные: большая часть данных может быть представлена в виде цифровых значений или категориальной оценки. Отдельно следует выделить группу патоморфологических исследований, к которой относятся:

1) электронная микроскопия;
2) исследование «тотальной ткани»;
3) изучение постоянных препаратов;
4) исследование временных препаратов;
5) метод культуры тканей;
6) авторадиография.

Результаты этих методов исследований также могут быть представлены в виде изображений в различных форматах. При подготовке набора данных необходимо выбрать единый формат изображения, а также формат предоставления сопроводительной документации.

Медицинские изображения: изображения внутренних структур тела для клинического анализа и медицинского вмешательства, а также визуального представления функций некоторых органов или тканей, получаемые неинвазивно специальными устройствами и датчиками. Различают следующие основные модальности медицинских изображений, хранящихся в DICOM-формате:

EPS – Электрофизиология сердца (Cardiac Electrophysiology);
CR – Компьютерная рентгенография (Computed Radiography);
CT – Компьютерная томография (Computed Tomography);
DX – Цифровая рентгенография (Digital Radiography);
ECG – Электрокардиография (Electrocardiography);
ES – Эндоскопия (Endoscopy);
XC – Наружная фотография (External-camera Photography);
IVUS – Внутрисосудистый ультразвук (Intravascular Ultrasound);
MR – Магнитно-резонансная томография (Magnetic Resonance);
MG – Маммография (Mammography);
NM – Ядерная медицина (Nuclear Medicine);
OP – Офтальмологическая фотография (Ophthalmic Photography);
PX – Панорамная рентгенография (Panoramic X-Ray);
PT – Позитронно-эмиссионная томография (Positron emission tomography);
RF – Рентгенофлюороскопия (Radiofluoroscopy);
RG – Рентгенография (Radiographic imaging);
US – Ультразвуковая диагностика (Ultrasound);
XA – Рентгеновская ангиография (X-Ray Angiography);
BI – Биомагнитные изображения (Biomagnetic imaging);
CD – Цветовое доплеровское картирование (Color flow Doppler);
ST Однофотонная эмиссионная компьютерная томография (Single-photon emission computed tomography (SPECT));
TG – Термография (Thermography);
AU – Аудиозаписи (Audio);

SR – Документ структурированного отчета (SR Document);

SMR – Стереометрическое взаимодействие (Stereometric Relationship);

SC – Вторичный захват (Secondary Capture);

OT – Другое (Other).

Как правило, медицинские изображения хранятся в формате DICOM, который, однако, позволяет хранить не всегда полный объем диагностических данных (например, спектральные изображения).

После обработки исследования могут преобразовываться в другие форматы хранения, например, Neuroimaging Informatics Technology Initiative (NIFTI), наборы графических форматов (jpg, png) и т.д.

2. ПОДХОДЫ К ВЫБОРУ ВАРИАНТА ИСПОЛЬЗОВАНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЗДРАВООХРАНЕНИИ

Идеальный вариант использования (use case) алгоритма ИИ должен быть специфичным, измеримым и достижимым, с четко определенными пользователями и ценностью [3]. При формировании варианта использования можно ориентироваться на адаптированную шкалу SMART-GEM [4]. Варианты использования также помогают проиллюстрировать наличие стандартов или необходимость их разработки [5].

Ценность ИИ для врачей-клиницистов и пациентов заключается в том, что алгоритмы машинного обучения могут улучшить результаты лечения и снизить затраты на здравоохранение [6]. Однако коммерчески доступные алгоритмы анализа медицинских изображений требуют независимой проверки и потенциального одобрения FDA (Food and Drug Administration) [7]. Наборы данных для независимой валидации ИИ должны включать элементы данных, направленных на выявление пределов работы, погрешностей и даже потенциальных этических проблем, вызванных применением конкретного ИИ-сервиса [8].

Вариант использования алгоритма ИИ для медицинской организации должен отражать клиническую задачу, включая согласованные с врачами конечные точки, содержащие текущие клинические рекомендации и показатели, поддающиеся независимой оценке человеком [9]. В качестве ошибочного примера использования можно рассмотреть применение алгоритма ИИ, предназначенного для скрининговых целей [10] (например, для скрининга туберкулеза), у пациентов с ургентной патологией. Данное применение некорректно, т.к. алгоритм предназначен для выявления признаков инфекционных заболеваний в условиях скрининговых исследований у потенциально здоровых людей, и применение его в условиях больницы скорой помощи для выявления гемоторакса, пневмоторакса, гидроторакса и иных ургентных состояний может представлять опасность для пациента вследствие пропуска патологий.

Таким образом, алгоритм ИИ для медицинской организации должен обладать [11]:

1) клинической ценностью (clinical utility), оказывая положительное влияние на рабочий процесс в медицинской практике;

2) статистической достоверностью (statistical validity), посредством обучения модели на больших и разнообразных наборах данных для достижения высокой надежности на новой популяции;

3) экономической ценностью (economic utility), продемонстрированной проспективно или ретроспективно.

3. ПОДХОДЫ К ФОРМИРОВАНИЮ НАБОРОВ ДАННЫХ

3.1. Формирование технического задания

На рисунке 3 приведены основные этапы формирования набора данных, которые будут рассмотрены в данных методических рекомендациях.

Рисунок 3 – Этапы формирования набора данных

Важным этапом подготовки набора данных является этап планирования, или формирования требований к набору данных (техническое задание (ТЗ)). Этот этап необходим с целью определения технических требований к набору данных, которые будут строго соответствовать поставленным клинической и практической задачам его формирования.

3.1.1. Определение клинической и практической задач подготовки набора данных

Под клинической или практической задачей понимается сценарий применения (use-case) технологий искусственного интеллекта для диагностики, дифференциальной диагностики (осуществление независимого чтения), помощи в принятии решения (увеличение спектра входных данных для медицинского работника, принимающего решение), маршрутизации пациентов (обеспечение триажа), обеспечения технической поддержки (снижения бремени рутинной работы) и др.

Задача отвечает критериям актуальности, если:

1) она обусловлена потребностью со стороны врачебного сообщества, иными словами, врачи понимают, что именно получат в результате работы ИИ-сервисов, и заинтересованы в этом результате;

2) имеются ее существующие решения на рынке технологий искусственного интеллекта как в России, так и за рубежом;

3) решение этой задачи принесет значимый социоэкономический эффект.

Для того, чтобы принять решение об актуальности задачи, необязательно наличие всех трех пунктов; например, актуальной может быть задача автоматизации, для которой пока не существуют разработанные ИИ-сервисы, однако ее решение имеет стратегическое значение для здравоохранения.

Также при формировании клинической и практической задач необходимо оценить, является ли доступ к данным или другая деятельность по их обработке допустимыми, и определить:

1) какие данные допустимо собирать, в каком объеме и из каких источников;
2) как их следует использовать (применительно к каким задачам);
3) как защитить данные при применении сервисов ИИ;
4) кому их следует раскрывать (доступ третьим лицам);
5) в течение какого времени они должны быть доступны.

3.1.2. Определение параметров набора данных

Необходимо определить ожидаемые результаты работы сервисов ИИ, а также требования к набору данных (включая требования к разметке и верификации). В данном процессе участвуют научные сотрудники, врачи, биостатистики.

Пункты 3.1.2.1–3.1.2.7 настоящего раздела являются основой для заполнения ТЗ на набор данных. В результате выполнения этой работы будут сформированы техническое задание на набор данных, а также базовые требования к результатам работы ИИ-сервиса.

3.1.2.1. Характеристики клинической или практической задач

Клиническая задача может быть дихотомной (разделение на два класса, например, есть признаки целевой нозологии и нет признаков целевой нозологии, в этом случае решается задача «бинарной классификации») либо заключать в себе дифференциальный диагноз (например, разделение одной патологии от другой, т.е. «мультиклассовая классификация»). Клиническая задача может также заключаться в измерении какой-либо непрерывной величины (например, процента поражения паренхимы легкого), а в случае наличия находок на изображении – в их детекции, поиске и отображении, а также в выборе тактики лечения, предсказания исходов на основе аналитики и т.д.

В зависимости от клинической задачи меняются и требования к выходным форматам результатов работы ИИ-сервисов:

1) для бинарной классификации – выбор одного из классов или % вероятности одного из классов;

2) для мультиклассовой классификации – выбор наиболее вероятного класса или % вероятности каждого класса;

3) для непрерывной величины – предсказание этой величины с известной погрешностью;

4) для поиска находок на изображении – координаты находки, «тепловая карта» или контур на изображении.

3.1.2.2. Определение области применения набора данных

От области применения набора данных зависят такие параметры, как баланс классов (наличие равного количества элементов набора данных для разных классов, например, нормы и патологии в случае бинарной классификации, см. п. 3.1.2.6), целевое количество исследований и др.

На данном шаге необходимо определить, какая цель применения набора данных:

1) тестирование ИИ-сервисов с целью их валидации и верификации:

– путем выполнения функционального тестирования (проверка работоспособности алгоритма, визуальная оценка выходных данных и т.п.);

– в рамках клинической валидации (проверка метрик точности работы ИИ-сервиса в рамках его предназначения);

– «селф-тест» (самостоятельная проверка разработчиками способности ИИ-сервиса обрабатывать разнородные входные данные);

2) машинное обучение:

– дообучение алгоритмов (transfer learning);

– обучение алгоритмов.

В случае применения набора данных в машинном обучении необходимо учитывать модель применения ИИ-сервисов. Например, для разработки ИИ-сервиса, который будет работать «до врача», т.е. выполнять отсев исследований «без патологии», в наборе данных должны быть предусмотрены различные вариации нормы. Однако набор данных для разработки ИИ-сервиса «с врачом» должен быть нацелен на дифференциальную диагностику, сложные случаи и т.д.

3.1.2.3. Определение клинических параметров набора данных

Для каждого набора данных необходимо определить требуемый список для его описания, который должен в полной мере характеризовать клинические параметры набора данных применительно к конкретной области здравоохранения.

Предлагаемый вариант списка клинических параметров относится к инструментальной диагностике:

1) вид входных данных согласно общепринятым стандартам (например, для медицинских изображений – согласно стандарту DICOM [12]);
2) анатомическая локализация (согласно справочнику [13]);
3) целевая нозология (одна или несколько согласно справочнику [14]);
4) популяционные критерии:
- заданная верхняя граница возраста;
- заданная нижняя граница возраста;
- распределение по полу;
- география и даты сбора данных;
- характеристики медицинской организации (МО) для сбора данных:
  - наименование;
  - тип;
  - вид (детское, взрослое, смешанное);
- эпидемиологическая обстановка при сборе данных;
- иные критерии выбора пациента.

3.1.2.4. Определение технических параметров набора данных

Для набора данных необходимо определить следующие технические параметры:
1) особенности диагностических устройств, на которых данные были получены:
- перечень производителей (и их моделей при необходимости);
- технические характеристики;
- наличие особенных режимов исследований;
2) требования к техническим характеристикам данных (например, разрешение);
3) требования к обезличиванию исследования:
- обезличивание метаданных:
  - согласно общепринятым стандартам (например, для изображений это стандарт DICOM, секция E1, таблица E.1–1 [15]);
  - с сохранением особой персональной информации для последующего сопоставления с сопроводительными материалами;
- обезличивание пиксельных изображений:
  - детекция текста на изображениях;
  - удаление мягких тканей лица в исследованиях головы, шеи и головного мозга;

3.1.2.5. Определение критериев разметки

Критерии разметки, на которых строится ее дальнейшее планирование, включают:

1) входные данные набора данных:
- наименование единицы входных данных;
- формат единицы входных данных;

2) выходные данные набора данных:
- наименование единицы выходных данных;
- формат единицы выходных данных;

3) характер разметки:
- один лейбл;
- мультилейбл;

4) для каждого лейбла:
    - уровень разметки (выбор из списка):
        - пациент;
        - исследование;
        - серия;
        - изображение;
    - уровень детализации разметки:
        - исследование/серия/изображение;
        - находка (локализация);
        - находка (сегментация);
    - характер лейбла:
        - бинарная классификация;
        - мультикласс (более 2 зависимых классов);
        - непрерывная величина;
    - для каждого класса:
        - критерии включения в класс;
        - критерии невключения в класс;
        - источник данных о критериях (само исследование (изображения), метаданные, иные источники);
        - ссылка на литературный источник.

3.1.2.6. Определение баланса классов и целевого количества исследований

Для того, чтобы определить баланс классов и целевое количество исследований, необходимо обратиться к области применения набора данных, а также сложности разметки (количество лейблов и классов).

Классы одного лейбла могут быть:

1) сбалансированы (количество исследований совпадает в классах);
2) несбалансированы (количество исследований одного класса превалирует).

Одним из самых популярных из несбалансированных наборов данных являются наборы, основанные на предтестовой вероятности. В этом случае

количество исследований с признаками патологии будет сопоставимо с выявляемым количеством этих признаков в заданной популяции.

Для достижения целей тестирования, в котором используется анализ характеристической кривой (ROC-анализ), следует стремиться к сбалансированным наборам данных.

Целевое количество исследований каждого класса рассчитывается статистическими методами исходя из требуемой статистической мощности теста. Для целей расчетов объема набора данных необходимо обращаться к биостатистикам.

3.1.2.7. Определение источников исходных данных для набора

Для большинства случаев источником исходных данных будут медицинские информационные системы (например, в Москве – это система ЕРИС ЕМИАС), однако данные могут поступать и из других баз данных, материальных носителей; сопроводительные документы могут также быть дополнительно загружены из других источников (например, данные клинических диагнозов и др.).

3.2. Сбор исходных данных согласно техническому заданию

3.2.1. Введение в цифровые медицинские данные

Обоснование целесообразности разработки ИИ, подготовка данных для разметки и тестирования, а также проспективная работа ИИ связаны с доступом к исходным данным, которые могут быть как «сырыми данными», так и прошедшими предварительную обработку на медицинском оборудовании, и доступны для конечного пользователя (медицинского персонала). «Сырыми» являются необработанные данные, полученные с диагностических устройств, которые зачастую недоступны ни пользователю, ни медицинским информационным системам, и представляют собой сложный математический набор данных, не имеющий ценности для врача. «Сырые данные» представляют ценность для разработчиков ПО при оптимизации работы алгоритмов предварительной обработки сигналов, полученных с медицинского оборудования.

3.2.2. Юридическая основа сбора исходных данных

Доступ к исходным данным связан с ограничениями, которые накладываются рядом нормативно-правовых актов:

1. Федеральным законом Российской Федерации № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации» от 21.11.2011 (с изм. и доп., вступ. в силу с 13.07.2021): ст. 4; ст. 13, ч. 2–4; ст. 92.

2. Федеральным законом Российской Федерации № 152-ФЗ «О персональных данных» от 27.07.2006. Выдержки, относящиеся к вопросам сбора, обработки и передачи наборов данных: ст. 5,6.

Также, согласно Европейским нормам GDPR (General Data Protection Regulation) [16], к персональным данным относятся все сведения, которые прямо или косвенно могут идентифицировать лицо, что является более широким понятием, чем принято в Российской Федерации (согласно закону РФ № 152-ФЗ «О персональных данных» от 27.07.2006).

Обезличивание (деидентификация, деперсонификация, de-identification) – действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных их конкретному субъекту. Основная цель обезличивания – обеспечение конфиденциальности персональных данных.

В нормативно-методических документах, посвященных вопросу «удаления» связи между персональными данными и субъектом персональных данных, используются три основных термина:

1. Анонимизация (обезличивание) – действия, в результате которых удаляется связь между совокупностью идентифицирующих данных и субъектом данных. Необратимо: все атрибуты из записи удаляются либо изменяются таким образом, чтобы было невозможно выполнить идентификацию субъекта (необратимое обезличивание).

2. Псевдонимизация – особый случай обезличивания, при котором, помимо удаления прямой связи с субъектом данных, создается связь между конкретной совокупностью характеристик этого субъекта и одним или несколькими псевдонимами. Действие является обратимым. Как и при анонимизации, атрибуты изменяются или удаляются из записи, но при этом обезличенные данные о пациенте помечаются псевдонимом.

3. Обратный процесс – деобезличивание (обратная персонификация) – действия, в результате которых обезличенные данные принимают вид, позволяющий определить их принадлежность конкретному субъекту персональных данных – становятся персональными данными.

При формировании наборов данных термины «обезличивание», «анонимизация» и «деперсонализация» являются синонимами.

Согласно приказу Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор) от 5 сентября 2013 г. № 996 «Об утверждении требований и методов по обезличиванию персональных данных» к свойствам обезличенных данных относятся:

1) полнота (сохранение всей информации о конкретных субъектах или группах субъектов, которая имелась до обезличивания);

2) структурированность (сохранение структурных связей между обезличенными данными конкретного субъекта или группы субъектов, соответствующих связям, имеющимся до обезличивания);

3) релевантность (возможность обработки запросов по обработке персональных данных и получения ответов в одинаковой семантической форме);

4) семантическая целостность (сохранение семантики персональных данных при их обезличивании);

5) применимость (возможность решения задач обработки персональных данных, стоящих перед оператором, осуществляющим обезличивание персональных данных, обрабатываемых в информационных системах персональных данных, в том числе созданных и функционирующих в рамках реализации федеральных целевых программ (далее – оператор, операторы), без предварительного деобезличивания всего объема записей о субъектах);

6) анонимность (невозможность однозначной идентификации субъектов данных, полученных в результате обезличивания, без применения дополнительной информации).

Таким образом, существующая нормативно-правовая база ставит серьезные ограничения и предоставляет пути, по которым можно следовать при сборе исходных данных для работ, связанных с ИИ в медицине.

3.2.3. Подготовка инфраструктуры для получения исходных данных

В ходе первичного этапа разрабатываются программные решения, которые обеспечивают задачи автоматизированной подготовки исходных данных к разметке, самой разметки и формирования наборов данных. В работах этого этапа участвует команда разработки программного обеспечения (далее – ПО), включая архитектора информационных систем, архитектора баз данных, дизайнера интерфейсов, devops-инженера, программистов и тестировщиков. Получение исходных данных из медицинских информационных систем может быть как единичным, так как пакетным. Для подготовки наборов данных, предназначенных для работы с ИИ, приоритетным является пакетный путь сбора исходных данных, так как обычно требуется большое количество исследований от разных пациентов. Для этого необходимо программное обеспечение, соответствующее набору функциональных возможностей. Функциональные требования к ПО (на примере получения доступа к цифровым данным лучевой диагностики из организаций, подведомственных Департаменту здравоохранения г. Москвы) заключаются в:

1) поиске исходных данных в ЕРИС ЕМИАС;

2) выгрузке текстовых протоколов по предоставленным спискам уникальных идентификационных номеров исследований (study_uids);

3) отборе исследований по ключевым словам (предварительная сортировка);

4) отборе и фильтрации исследований по техническим параметрам;

5) обеспечении первого/второго чтения врачами (разметчиками и экспертами);

6) обеспечении верификации;

7) сохранении результатов разметки и верификации в машиночитаемом виде;

8) подготовке сопроводительной документации к набору данных.

3.2.4. Сбор исходных данных

Важным этапом при планировании сбора исходных данных является их доступность, которая зависит от источника данных (таблица 2).

Таблица 2 – Источники данных

Наименование типов источников данных	Наименование типов данных
Медицинские источники	– Изображения (КТ, МРТ, видеооперации и др.) – Текст (ЭМК, клинические протоколы и рекомендации и др.) – Звук (записи голоса пациентов, хрипов, кашля и др.) – Сигналы (ЭЭГ, ЭКГ, данных с прикроватных мониторов и носимых устройств и др.) – Генетические (NGS, ДНК-чип и др.)
Фармакологические источники	– Данные клинических исследований – Данные о препаратах – Данные о продажах – Данные фармаконадзора
Финансовые источники	– Данные о движении денежных средств ФОМС, МО – Данные о движении денежных средств страховых компаний
Административные источники	– Данные о работе МО (загрузка кабинетов, оборудования, расписание приемов и др.) – Реестры ФРМО, ФРМР, НСИ – Данные о жалобах и отзывах на МО – Данные о распространении заболеваний, эпидемиях и др.
Страховые источники	– Страховые отчеты – Отчеты по скорингу клиентов – Данные о претензиях
Внешние источники	– Демографические данные – Биомедицинская литература Данные, полученные с форумов о здоровье – База медицинских онкологических заболеваний (канцер-регистр) – База ЗАГС – Открытые базы данных

При включении дополнительных характеристик в планируемый набор данных уменьшается количество субъектов, которые будут обладать всем набором характеристик одновременно, поэтому этап планирования сбора

исходных данных должен включать в себя понимание цели для применения каждого набора данных.

С 2013 года Министерством здравоохранения Российской Федерации определена структура данных в электронной медицинской карте (далее – ЭМК) [17]. Ее необходимо использовать при создании и доработке медицинских информационных систем (далее – МИС) в лечебно-профилактических учреждениях.

ЭМК является долговременным накопителем информации о том, что произошло у пациента или было сделано для него. Карта должна содержать информацию, относящуюся ко всем видам медицинского обеспечения – результаты врачебных наблюдений, мнения и планы лечения.

Структура ЭМК включает 15 разделов: «Метрики пациента», «Результаты исследований», «Врачебные осмотры», «Заболевания и осложнения», «Рецепты на лекарственные средства» и другие.

Каждый из разделов состоит, в свою очередь, из десятков параметров (полей ЭМК). Например, раздел «Врачебные осмотры» должен содержать информацию о должности и Ф.И.О. врача, симптомах и жалобах, диагнозе и т.д. Такой структурированный подход создает условия для повышения точности формирования запроса для сбора тех или иных наборов данных.

3.2.5. Этапы формирования наборов данных

Большинство систем здравоохранения не имеют достаточного оборудования и возможностей для обмена большими объемами медицинских изображений [18]. Даже когда разработка/тестирование/ применение ИИ возможно с этической, нормативно-правовой и финансовой сторон, медицинские данные часто хранятся в разрозненных хранилищах, что не является оптимальным для работ по развитию медицинского ИИ, который может широко использоваться в клинической практике. Сбор исходных данных в большом объеме в одном месте не позволяет использовать их для разработки, тестирования и/или пострегистрационного мониторинга медицинского ИИ на должном уровне. Существует также проблема с методологией сбора и хранения исходных данных, разработка которой должна быть проведена до непосредственной работы с исходными данными. Один из вариантов решения лежит в делегировании ответственности и прав за формирование, подготовку и тестирование в общем цифровом пространстве от одной медицинской организации к другой. Такой путь принят в Москве, где ГБУЗ «НПКЦ ДиТ ДЗМ» приняло на себя ответственность за формирование методологии и внедрение ИИ в рамках Эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы [19].

Подготовка исходных данных должна проходить в несколько этапов:

1. Планирование целей применения набора данных. Этому важному пункту часто уделяют недостаточно внимания, однако тут должны быть определены количество пациентов/исследований, области применения; критерии включения/исключения; уровень достоверности маркировок набора данных (Ground truth); необходимость (или возможность) обновления (или расширения) набора данных для получения новых областей применения имеющихся данных, включая ограничения его применения.

2. Одобрение этического комитета на использование исходных данных для конкретной цели. Для использования исходных данных необходимо информированное согласие.

3. Получение доступа к соответствующим исходным данным по релевантному запросу.

4. Обезличивание данных и надежное хранение ключей для деанонимизации (деперсонализации).

5. Проверка качества данных. Качество и количество изображений варьируются в зависимости от целевой задачи и предметной области. Если данные предназначены для исследований с открытым исходным кодом, то дополнительная проверка каждого изображения человеком является стандартной практикой, поскольку некоторые изображения содержат аннотации произвольной формы, которые не могут быть гарантированно удалены автоматическими методами.

6. Структурирование данных в гомогенизированных и машиночитаемых форматы [20] (например, DICOM или NIFTI).

7. Присоединение к исследованиям достоверной информации (Ground truth), которая может быть одной или несколькими метками, сегментами или сторонним исследованием более высокой степени доказательности (например, биопсия или лабораторные результаты).

8. Регистрация полученного набора данных как самостоятельного объекта для интеллектуальной защиты.

3.2.6. Рекомендации по сбору данных

Что касается специалистов по работе с данными, которые могут использоваться для обучения, дообучения, тестирования, валидации или научного анализа, то наборы данных могут быть получены следующими путями:

1) ретроспективный/проспективный сбор данных (должен проходить путь указанный выше);

2) проведение научного исследования;

3) использование общедоступных баз данных с соответствующими разрешениями на их использование – самый простой, однако имеет свои ограничения:

– некоторые открытые наборы данных разрешено применять для научных целей, а не для разработки коммерческого ИИ;
– невозможность контроля качества набора данных;
– ограниченность объема набора данных.

Большой и важной проблемой качества подготовки наборов данных является поиск достоверной информации (Ground truth) для подтверждения целевой патологии на изображениях лучевой диагностики. Кроме формирования разметки на изображения, которые могут быть весьма трудоемкими, следует помнить, что для каждого исследования существует сформированный протокол описания – интерпретация врачом-рентгенологом этого исследования. Использование этих протоколов может либо быть самостоятельной разметкой, либо способствовать уменьшению количества исследований, которые потом потребуют разметки на изображениях. Существуют подходы по выполнению ретроспективной маркировки, от простой ручной маркировки [21] рентгенологами до автоматизированных подходов, которые могут извлекать структурированную информацию из рентгенологического заключения и / или электронной медицинской карты.

Существует тенденция к интерактивной отчетности, в которой отчет радиолога содержит гипертекст, напрямую связанный с аннотациями к изображениям [22]. Такие аннотации могут использоваться для маркировки наборов данных с открытым исходным кодом [23]. Однако это решение не может считаться высококачественным, так как около 2–20 % заключений рентгенологов содержат ошибки [24].

Когда речь идет о разметке изображений, то следует отдавать приоритет специалистам с достаточным опытом работы, не забывая, что среди них должен быть эксперт, который будет лишь валидировать эту разметку в роли аудитора. Данные задачи краудсорсинга должны ставиться после предварительного обсуждения с экспертами, так как количество разметчиков и наличие аудитора может различаться в зависимости от поставленной задачи. Так, для сложных задач количество разметчиков должно быть существенным, например, рекомендованное количество экспертов для разметки легочных узлов на каждое КТ-исследование – 4 разметчика и 1 эксперт-валидатор [25].

При планировании сбора данных рекомендуется планировать соотношение обучения, тестирования и валидации в наборе данных, например, 80:10:10 или 70:15:15. Чтобы обеспечить обобщаемость алгоритма ИИ, необходимо ограничить смещение набора обучающих данных. Если алгоритм ИИ обучен с изображениями из московского учреждения, а алгоритм используется для азиатского населения, то на производительность может влиять смещение населения или распространенности заболеваний. Аналогично, если все данные обучения визуализации были получены с использованием одного вида томографа, он может не работать так же хорошо на машинах других производителей.

Таким образом, рекомендуется использовать изображения из нескольких различных источников или, по крайней мере, изображения, представляющие целевую популяцию или систему здравоохранения, в которой будет применяться алгоритм. Для обеспечения обобщения часто необходимы большие наборы обучающих данных. Для конкретных целевых приложений или групп может быть достаточно относительно небольших наборов данных (сотни случаев). Большие размеры выборки особенно необходимы в популяциях со значительной гетерогенностью или когда различия между визуализируемыми фенотипами незначительны [26].

При расчете размера выборки для наборов тестовых данных следует использовать традиционные методы расчета мощности для оценки размера выборки. В целом, разработка обобщаемых алгоритмов ИИ в медицинской визуализации требует наборов статистически обоснованных данных порядка сотен тысяч, что является проблемой для многих исследователей и разработчиков. Частичным решением этой проблемы может быть полууправляемое обучение. Для обучения с учителем необходимы полностью аннотированные наборы данных, тогда как полууправляемое обучение [27] использует комбинацию аннотированных и неаннотированных изображений для обучения алгоритма.

Отдельного внимания заслуживает федеративное обучение. Ряд компаний [28, 29] представили возможность не передавать данные из медицинских организаций, а проводить обучение ИИ-моделей на вычислительных мощностях, развернутых в стенах медицинских организаций. За счет обмена обученными моделями между медицинскими учреждениями осуществляется дообучение и повышение метрик диагностической точности. Обучение, тестирование и проспективная работа осуществляются внутри медицинских организаций без необходимости передачи наборов данных (исследований) во вне. Несмотря на потенциальные преимущества этого метода, существуют важные проблемы, которые необходимо решить до внедрения федеративного обучения в широкую практику:

1. Необходимо обеспечить стандартизацию интерпретации и разметки медицинских данных в тех организациях, в которых будет развернут алгоритм медицинского ИИ.

2. В зависимости от сложности алгоритма ИИ могут потребоваться значительные вычислительные ресурсы, в т.ч. в каждом объекте федеративного обучения.

3. Ограниченности видимости данных для разработчика ИИ-алгоритмов, предварительная обработка и организация данных для приема алгоритмом являются сложными задачами.

4. Разные учреждения могут обладать неоднородными данными с точки зрения популяций пациентов, распределения патологии, объема данных, формата данных и так далее.

Одним из наиболее важных ограничений обучения алгоритмов ИИ на основе данных из одного учреждения или из нескольких учреждений в небольшой географической области является систематическая ошибка выборки. Если алгоритм ИИ, обученный таким образом, применяется к другой географической области, то результаты алгоритма могут быть ненадежными из-за различий между выборкой и целевой популяцией [30].

Следовательно, существуют различные рекомендации по сбору и маркировке данных для ИИ. Врачи, исследователи данных и лица, принимающие решения о предоставлении данных или внедрении новой модели ИИ, должны знать об источнике данных для обучения, потенциальных проблемах, которые могут повлиять на обобщаемость алгоритмов ИИ. Новые подходы, такие как федеративное обучение, интерактивная отчетность и синоптическая отчетность, могут помочь решить проблему доступности данных в будущем. Однако их редактирование и аннотирование, а также вычислительные требования являются существенными препятствиями на этом пути.

3.3. Классификация наборов данных по разметке

Тип разметки изображений варьируется в зависимости от задачи, выполняемой алгоритмом ИИ-сервиса [31].

Выделяют три вида набора данных, определяемых процессом выполнения разметки (рис. 4, 5) [18]:

1) набор данных по ретроспективной разметке;
2) набор данных по проспективной разметке;
3) набор данных с верификацией.

Рисунок 4 – Классификация видов разметки по степени ценности

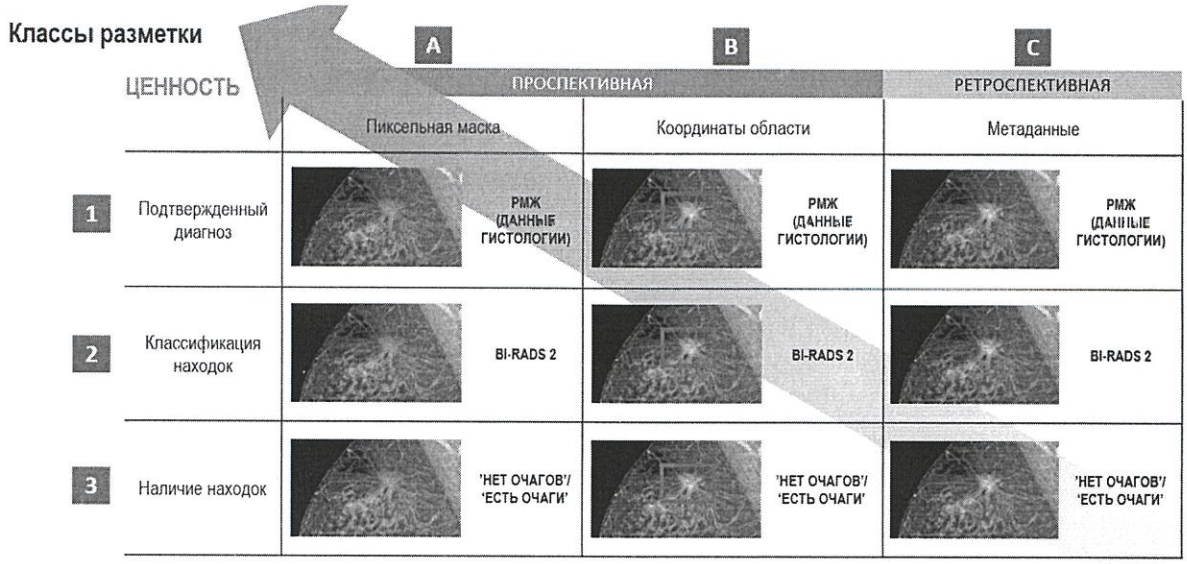

Рисунок 5 – Классификация разметки в лучевой диагностике

На рисунке 5 выше представлена классификация типов разметки [32]. Для ретроспективной разметки (1C, 2C, 3C) могут быть использованы данные сопроводительных документов (такими, например, являются тексты заключений для результатов инструментальных исследований), МИС, электронных медицинских карт и др. Примером могут служить и метаданные, генерируемые автоматически прибором при проведении исследования и хранящиеся в исходных данных. Очевидное преимущество ретроспективной разметки заключается в том, что на нее требуется значительно меньше времени со стороны медицинских специалистов, так как бóльшую часть подготовительной работы выполняет специалист по работе с данными.

В случае, когда имеется подтвержденный диагноз для элементов в наборе данных, например, результаты патогистологии, лабораторных данных, повторных исследований (если применимо), то данный набор данных является наиболее достоверным (строка 1).

Проспективная разметка (1A, 1B) предполагает активное вовлечение врачей в процесс включения в набор данных информации, например, позволяющей эффективно выполнить классификацию элементов набора данных. В лучевой диагностике под разметкой чаще всего понимают классификацию исследований по классам (наличие или отсутствие рентгенологических признаков выбранного заболевания), а также выделение в виде графического обозначения области интереса [18], соответствующей искомым признакам (например, очаги демиелинизации при рассеянном склерозе на МР-изображениях головного мозга). Степень вовлечения может быть разделена на более и менее затратную: в первом случае экспертам предлагается обвести контур области интереса, т.е. создать пиксельную маску на уровне области интереса (столбец А), во втором – обозначить ее координаты простой геометрической фигурой (столбец В).

3.3.1. Ретроспективная разметка

Ретроспективная разметка данных представляет собой сбор элементов в соответствии с метаданными, отбирающимися согласно поставленной цели. Такую разметку проводят путем минимальных трудозатрат: выгрузка данных 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.).

3.3.2. Проспективная разметка

Проспективная разметка аналогично ретроспективной разметке 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.). Данный вид разметки проводят с участием обученного медицинского персонала (зачастую квалифицированного врача в субспециализации размечаемого набора данных) путем ручного аннотирования содержания данных или их частей.

3.3.3. Верифицированный набор данных

Верифицированный набор данных получают при дополнении набора данных, подготовленного при проспективной разметке врачами, данными из медицинских записей, в том числе об окончательном (поставленном врачом-клиницистом) диагнозе и/или при постановке патологоанатомического диагноза. В качестве метода для верификации набора данных часто применяется метод «золотого стандарта» (ground truth) для целевой патологии, который представляет собой повторное исследование пациента через определенное время, результаты патогистологических, иммунологических и других исследований, ответ на терапию и т.д. [18, 20, 31, 33].

Существует также верификация набора данных путем его слепого анализа врачами-экспертами с достижением заданного уровня согласованности их решений. Когда задействовано несколько врачей-экспертов, следует описать процесс, с помощью которого происходило объединение их интерпретации для определения общего эталонного стандарта, и то, как ваш процесс учитывает несоответствия между экспертами, участвующими в ходе установления истины (вариативность

истинности) [34]. Отличие от проспективной разметки заключается в пересмотре данных группой врачей-экспертов и формировании единого экспертного мнения согласно методам принятия решений (например, «методу шара»).

Выделяют следующие критерии отнесения набора данных к верифицированному набору данных:

1) данные получены из реальной практики (не допускается получение синтезированных данных, например, от генератора физиологических сигналов ЭКГ);

2) структура набора данных соответствует поставленной цели его формирования (обучение, аналитическая, клиническая валидация и др.);

3) количество наблюдений (исследований) достаточно для достижения статистической значимости результата;

4) разметка проведена экспертной группой;

5) разметка проведена с использованием тезауруса (кодированной библиотеки типовых формулировок, соответствующих рекомендации ассоциации специалистов в данной области).

3.3.4. Требования к экспертам, принимающим участие в разметке набора данных

Специалисты (врачи, инженеры), формирующие набор данных, должны быть компетентными в соответствии с полученным образованием, подготовкой, навыками и опытом согласно ГОСТ ISO 13485. Сведения о необходимой квалификации, опыте и подготовке персонала необходимо указать в должностных инструкциях специалистов.

Разметчики должны подбираться по следующим критериям:

1. Компетентность в области конкретных типов данных: изображения, текстовые данные или сигнальные (ЭКГ, ЭЭГ и т.д.), количественные данные (ЧСС, артериальное давление, спирометрия и др.), бинарные данные (например, да/нет).

2. 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).

Например, признаки часто встречающихся нозологий (таких как пневмония или перелом трубчатых костей) могут быть размечены врачами более низкой квалификации, а сложные в диагностике и дифференциальной диагностике признаки (например, признаки демиелинизирующих болезней головного мозга) должны размечаться врачами высокой квалификации.

3.4. Контроль качества набора данных

Формирование набора данных должно быть спланировано и подвержено мониторингу и управлению для обеспечения соответствия качества.

Работой группы может руководить сотрудник, назначенный ответственным, который не принимает участия в разметке и/или аннотировании, но будет регулировать срочность, очередность и объем работы между экспертами. Обязанностью данного ответственного также является формирование рабочей группы для обеспечения объективности и достоверности результата.

Должны быть применены методы оценки качества набора данных, по которому будет производиться разметка:

1) проверка отсутствия пропусков элементов в наборе данных;

2) проверка отсутствия некорректных элементов для решения поставленных задач;

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

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

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 набора данных, что будет способствовать последующему обобщению шагов и проведению качественной разметки (аннотации) данных.

Фильтрация набора данных позволяет снизить 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 использованием системы искусственного интеллекта – СИИ (например, для автоматической оценки качества изображения).

3.5. Внесение изменений в наборы данных

После создания и регистрации набора данных может возникнуть необходимость внести изменения (в результате обнаружения ошибок или добавления новых данных) [35]. При внесении любых изменений следует документировать изменение версии набора данных (включая любую смену версии), позволяющей оценить вносимые изменения с течением времени. Эта документация должна быть приложена к набору данных. При смене версии используются трехзначные значения формата А.Б.В, где А – мажорная версия, Б – минорная версия, В – патч-версия [32]:

1. Мажорная версия увеличивается при изменении значимых параметров набора данных, связанных с клинической задачей, целью, принципами разметки и верификации данных.
2. 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 устанавливается в 0 при выпуске новой мажорной версии).
3. 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 единицы набора данных (патч-версия устанавливается в 0 при выпуске новой минорной или мажорной версии).

Следует учитывать, что при выпуске новой минорной или патч-версии ИИ-сервисы могут использовать набор данных без изменений кода, принимающего его элементы в качестве входных данных. При выпуске новой патч-версии набора данных количество и качество элементов наборов данных в качестве входных данных должны быть такими же, однако результат работы может быть иным (т.к. могут быть затронуты файлы разметки и верификации). При добавлении дополнительной серии единиц данных, что влечет за собой значимые изменения в клинической задаче набора данных и в целом в цели его создания, но не изменяет ее полностью – изменяется мажорная версия набора данных. Новый набор данных создается при условии полного изменения назначения набора данных, его цели создания и клинических задач.

ЗАКЛЮЧЕНИЕ

Настоящие методические рекомендации содержат описание практических подходов при планировании и создании наборов данных, необходимых для апробации и применения технологий искусственного интеллекта в здравоохранении. Внедрение подготовленных рекомендаций позволит повсеместно унифицировать разработку и обеспечить качество медицинских наборов данных, специально создаваемых для систем искусственного интеллекта. В силу того, что качество, надежность и безопасность работы систем искусственного интеллекта напрямую обусловлены наборами данных, которые легли в основу их создания и валидации, именно процесс их формирования в первую очередь требует прозрачной и понятной методологии для специалистов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining // Journal of Data Warehousing. 2000. № 5. P. 13–22.

2. Kazmierska J., Hope A., Spezi E. et al. From multisource data to clinical decision aids in radiation oncology: The need for a clinical data science community // Radiotherapy and Oncology. 2020. № 153. P. 43–54.

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

4. Bowman J., Mogensen L., Marsland E. et al. The development, content validity and inter-rater reliability of the SMART-Goal Evaluation Method: A standardised method for evaluating clinical goals // Australian occupational therapy journal. 2015. № 62.6. P. 420–427.

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

6. Lee D.H., Yoon S.N. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges // International Journal of Environmental Research and Public Health. 2021. Vol. 18, №1. P. 271.

7. Wu E., Wu K., Daneshjou R. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals // Nature medicine. 2021. Vol. 27, № 4. P. 582–584.

8. O'Reilly-Shah V., Gentry K.R., Walters A.M. et al. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment // British Journal of Anaesthesia. 2020. Vol. 125, № 6. P. 843–846.

9. Oren O. B., Gersh J., Bhatt D. L. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints / // The Lancet Digital Health. 2020. № 2.9. P. e486–e488.

10. Melendez J., Sánchez C.I., Philipsen R.H.H.M. et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information // Scientific Reports. 2016. № 6:25265. P. 1–8.

11. Sendak M.P., D'Arcy J., Kashyap S. et al. A path for translation of machine learning products into healthcare delivery // EMJ Innov. 2020. № 10. P. 1–14. DOI: 10.33590/emjinnov/19-00172.

12. DICOM Library – Anonymize, Share, View DICOM files ONLINE: [сайт]. Poland, 2021. URL: https://www.dicomlibrary.com/dicom/modality/ (дата обращения: 03.07.2021).

13. НСИ: [сайт]. РФ, 2021. URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477 (дата обращения: 03.07.2021).

14. НСИ: [сайт]. РФ, 2021. URL: https://nsi.rosminzdrav.ru/#!/ (дата обращения: 03.07.2021).

15. DICOM: [сайт]. United States, 2021. URL: http://dicom.nema.org/dicom/2013/output/chtml/part15/chapter_E.html (дата обращения: 03.07.2021).

16. General Data Protection Regulation (GDPR) – Official Legal Text: [сайт]. Germany, 2021. URL: https://gdpr-info.eu/General Data Protection Regulation (дата обращения: 03.09.2021).

17. Министерство здравоохранения Российской Федерации. Основные разделы электронной медицинской карты: [сайт]. РФ, 2021. URL: https://nsi.rosminzdrav.ru/#!/ (дата обращения: 03.07.2021).

18. Willemink M.J., Koszek W.A., Hardell C. et al. Preparing Medical Imaging Data for Machine Learning. DOI: 10.1148/radiol.2020192224 // Radiology. 2020. Vol. 295, № 1. P. 4–15.

19. Pavlov N., Kirpichev Y.S., Revazyan A. et al. Value of technical stratification of medical datasets for AI services. DOI: 10.1186/s13244-021-01014-5 // Insights Imaging. 2021. № 12 (Suppl 2): 75. P. 216.

20. Harvey H., Glocker B. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. DOI: 10.1007/978-3-319-94878-2_6 // Artificial Intelligence in Medical Imaging. 2019. P. 61–72.

21. Wang X., Peng Y., Lu L. et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 2097–2106.

22. Folio L.R., Machado L.B., Dwyer A.J. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology // RadioGraphics. 2018. Vol. 38, №2. P. 462–482.

23. Yan K., Wang X., Lu L. et al. Deep Lesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning // Journal of Medical Imaging. 2018. Vol. 5, №3. P. 1–11.

24. Brady A., Laoide R.O., McCarthy P. et al. Discrepancy and error in radiology: concepts, causes and consequences // The Ulster medical journal. 2012. Vol. 81, № 1. P. 3–9.

25. Кульберг Н.С., Решетников Р.В., Новик В.П. [и др.] Вариабельность заключений при интерпретации КТ-снимков: один за всех и все за одного // Digital Diagnostics. 2021. Т.2. № 2. С. 105–118.

26. Chang K., Balachandar N., Lam C. et al. Distributed deep learning networks among institutions for medical imaging // Journal of the American Medical Informatics Association. 2018. Vol. 25, № 8. P. 945–954.

27. Kingma D.P., Rezende D.J., Mohamed S. et al. Semi-Supervised Learning with Deep Generative Models. 2014. P. 1–9. URL: https://arxiv.org/abs/1406.5298 (дата обращения: 11.08.2021).

28. Sheller M.J., Edwards B., Reina G.A. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data // Scientific Reports. 2020. Vol. 10, №1. P. 12. DOI 10.1038/s41598-020-69250-1.

29. McMahan B., Ramage D. Federated Learning: Collaborative Machine Learning without Centralized Training Data // Google AI Blog: [website]. 2017. Apr. 6. URL: https://www.https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (дата обращения: 09.08.2021).

30. Toll D.B., Janssen K.J., Vergouwe Y. et al. Validation, updating and impact of clinical prediction rules: a review// Journal of clinical epidemiology. 2008. Vol. 61, №11. P. 1085–1094. URL: DOI 10.1016/j.jclinepi.2008.04.008.

31. 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, №5. P. 25–37.

32. 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.

33. Ranschaert E.R., Morozov S.P., Paul R. Artificial Intelligence in Medical Imaging. Opportunities, Applications and Risks // Artificial Intelligence in Medical Imaging. 2019. P. 705.

34. U.S. Food and Drug Administration. 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: [сайт]. Netherlands, 2021. URL: https:www.fda.gov/media/77642/download&lr=213&mime=pdf&l10n=ru&sign=5bc08065d038d478209b122441e2ffc4&keyno=0 (дата обращения: 03.07.2021).

35. Klump J., Wyborn L., Wu M. et al. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles // Data Science Journal. 2021. Vol. 20, №12. P. 1–13.