ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 59921.5— 2022

СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КЛИНИЧЕСКОЙ МЕДИЦИНЕ

Часть 5

Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов

Издание официальное

Москва
Российский институт стандартизации
2022

Предисловие

1 РАЗРАБОТАН Федеральным государственным бюджетным учреждением «Российский институт стандартизации» (ФГБУ «РСТ»), Государственным бюджетным учреждением здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ «НПКЦ ДиТ ДЗМ»)

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 31 марта 2022 г. № 180-ст

4 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КЛИНИЧЕСКОЙ МЕДИЦИНЕ

Часть 5

Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов

Artificial intelligence systems in clinical medicine. Part 5.
Requirements for the structure and application of dataset for training and testing algorithms

Дата введения — 2022—06—01

1 Область применения

Настоящий стандарт устанавливает общие требования к структуре и порядку применения наборов данных, которые используют на этапах разработки системы искусственного интеллекта (СИИ), включая обучение и внутреннее тестирование алгоритмов искусственного интеллекта, ее эксплуатации, а также внешнего тестирования (аналитическая и клиническая валидация).

Настоящий стандарт определяет методологическую основу для процесса подготовки и применения наборов данных, которые используют на этапах разработки, тестирования и эксплуатации систем искусственного интеллекта.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ 7.24 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению

ГОСТ 7.25 Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления

ГОСТ ISO 13485 Изделия медицинские. Системы менеджмента качества. Требования для целей регулирования

ГОСТ Р ИСО/МЭК 12207—2010 Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств

ГОСТ Р ИСО/МЭК 17826 Информационные технологии. Интерфейс управления облачными данными (CDMI)

ГОСТ Р ИСО 27799 Информатизация здоровья. Менеджмент защиты информации в здравоохранении по ИСО/МЭК 27002

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом ут-

Издание официальное

верждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями:

3.1 **аналитическая валидация** (analytical validation): Подтверждение способности системы искусственного интеллекта точно, воспроизводимо и надежно генерировать предполагаемые технические результаты вычислений из входных данных.

Примечания

1 См. [1].

2 Аналитическая валидация является частным случаем валидации в соответствии с ГОСТ Р ИСО/МЭК 12207—2010, пункт 4.54.

3.2 **верифицированный набор данных** (ground truth): Набор данных с верифицированной медицинской информацией.

3.3

верификация (verification): Подтверждение (на основе представления объективных свидетельств) того, что заданные требования полностью выполнены.

[ГОСТ Р ИСО/МЭК 12207—2010, пункт 4.55]

3.4 **воспроизводимость** (reproducibility): Свойство процесса получать одинаковые результаты испытаний в разных средах испытаний.

Примечание — Разные среды означают разные компьютеры, жесткие диски, операторы и т. д.

3.5

де-идентификация (de-identification): Общее название любого процесса удаления связи между совокупностью идентифицирующих данных и субъектом данных.

[ГОСТ Р 55036—2012, пункт 3.18]

3.6

искусственный интеллект (artificial intelligence): Комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение, поиск решений без заранее заданного алгоритма и достижение инсайта) и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.

Примечание — 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, анализу и синтезу решений.

[ГОСТ Р 59277—2020, пункт 3.18]

3.7

классификация (classification): Способ и результат упорядочения, структуризации некоторого множества объектов, разделения его на определенные подмножества путем артикуляции, выделения некоторого признака объектов исходного множества как основания их структуризации по данному признаку. Такого рода признак называется основанием классификации.

[ГОСТ Р 59277—2020, пункт 3.26]

3.8 **кластеризация** (claster analysis): Группировка экземпляров данных в кластеры со сходными характеристиками.

3.9 **клиническая валидация** (clinical validation): Подтверждение способности системы искусственного интеллекта выдавать клинически значимые выходные данные, связанные с целевым исполь-

зованием системы искусственного интеллекта в рамках установленного изготовителем функционального назначения.

Примечание — См. [1], пункт 7.0.

3.10 **контроль доступа** (access control): Средства, с помощью которых ресурсы системы обработки данных предоставляются только авторизованным субъектам в соответствии с установленными правилами.

3.11

метаданные (metadata): Информация о ресурсе.

Примечание — Метаданные бывают трех типов:
- описательные (служат для обнаружения, сбора или группирования данных по общим для них характеристикам);
- структурные (определяют состав или организацию набора данных);
- административные (используются для управления базой данных).

[ГОСТ Р 57668—2017, пункт 4.10]

3.12 **набор данных** (data set): Совокупность данных, прошедших предварительную подготовку (обработку) в соответствии с требованиями законодательства Российской Федерации об информации, информационных технологиях и о защите информации и необходимости для разработки программного обеспечения на основе искусственного интеллекта.

Примечание — См. [2].

3.13

обеспечение качества (quality assurance, QA): Совокупность систематических и планомерных действий, которые имеют целью обеспечить соответствие проведения исследования, сбора, регистрации и представления данных надлежащей клинической практике и нормативным требованиям.

[ГОСТ Р 52379—2005, пункт 1.34]

3.14 **обнаружение (детекция аномалий)** (detection): Идентификация редких экземпляров данных, существенно отличающихся от остальных.

3.15 **обучающая выборка** (training sample): Выборка, по которой производится настройка (оптимизация) параметров системы искусственного интеллекта.

3.16 **повторяемость** (repeatability): Свойство процесса, проводимого для получения одинаковых результатов тестирования в одной и той же среде тестирования.

Примечание — Одна и та же среда тестирования означает одинаковый компьютер, жесткий диск, режим работы и т. д.

3.17 **проверочная выборка** (validation sample): Выборка, на которой проводят проверку применимости параметров системы искусственного интеллекта для отличных от обучающей выборки наборов данных.

3.18 **размерность набора данных (арность)** (arity): Число атрибутов, которые имеют объекты в наборе данных (например, значение артериального давления, масса тела пациента, уровень холестерина и др.).

3.19 **разметка [аннотация] данных** (data labeling)**:** 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.

Примечание — См. [2].

3.20 **разреженность набора данных** (data sparsity): Доля атрибутов в наборе данных, содержащих недостающие, неизвестные либо пустые значения.

3.21 **регрессия** (regression): Аппроксимация и предсказание значения непрерывных параметров какого-либо объекта.

3.22 **ретроспективная разметка** (retrospective annotation): Сбор данных в соответствии с указанными метаданными, перечень которых выбирают в соответствии с поставленной целью формирования набора данных.

Примечание — Ретроспективная разметка не предполагает дополнительных манипуляций с элементами данных (например, постановка метки начала и окончания события, меток обнаружения признаков, обозначений патологий и т. п.)

3.23 **проспективная разметка** (prospective annotation): Сбор данных в соответствии с поставленной целью формирования набора данных, а также проведение дополнительных манипуляций с элементами.

Примечание — Проспективную разметку выполняют путем постановки метки начала и окончания события, меток обнаружения признаков, обозначений патологий и т. п.

3.24

сбор данных (data collection): Процесс объединения данных, поступающих из одного или более источников, в целях их использования при обучении и тестировании систем искусственного интеллекта.

[Адаптировано из ГОСТ 33707—2016, пункт 4.1218]

3.25

система искусственного интеллекта (artificial intelligence system): Программное обеспечение, в котором используются технологические решения искусственного интеллекта.

[Адаптировано из ГОСТ Р 59276—2020, пункт 3.16]

3.26 **система менеджмента качества систем искусственного интеллекта** (quality management system for artificial intelligence systems): Организационная структура, функции, процедуры, процессы и ресурсы, необходимые для скоординированной деятельности по руководству и управлению производителем системы искусственного интеллекта применительно к качеству.

3.27 **тестовая [контрольная] выборка** (test sample): Уникальная (отличная от обучающей и валидационной) выборка, на которой проводят объективную оценку качества параметров обученной системы искусственного интеллекта.

4 Общие положения

4.1 Введение

С целью повышения доступности и качества данных, необходимых для развития технологий искусственного интеллекта в сфере здравоохранения, в данном стандарте представлена унифицированная методология подготовки и использования набора данных, общая схема которого отображена на рисунке 1.

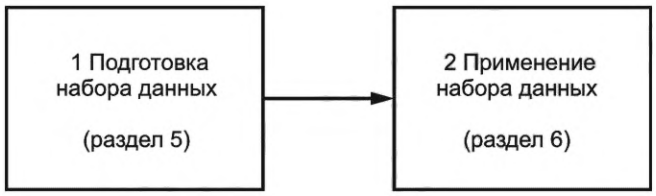

Рисунок 1 — Основные этапы жизненного цикла набора данных

4.2 Классификация наборов данных по виду разметки

Выделяют три вида набора данных, определяемых процессом выполнения разметки (рисунок 2):
- набор данных по ретроспективной разметке;
- набор данных по проспективной разметке;
- верифицированный набор данных.

Рисунок 2 — Классификация видов разметки по степени ценности

Примечание — См. [3].

4.2.1 Набор данных по ретроспективной разметке

Ретроспективная разметка представляет собой сбор элементов в соответствии с указанными метаданными, перечень которых выбирают в соответствии с поставленной целью. Такую разметку проводят путем выгрузки данных из медицинской информационной системы. Ретроспективная разметка не предполагает выполнение манипуляций или какой-либо обработки элементов. Для каждого элемента набора данных устанавливают соответствие с медицинской информацией (диагноз, результаты лабораторного тестирования и т. п.). Такая разметка не требует участия врача, а может быть выполнена техническим специалистом, который имеет опыт работы с наборами данных.

Пример — Ретроспективная разметка пациентов с подтвержденной коронавирусной болезнью. Перечень метаданных: идентификационный номер, дата рождения, дата выполнения лучевого исследования, результаты теста на полимеразную цепную реакцию и т. п.

4.2.2 Набор данных по проспективной разметке

Проспективная разметка представляет собой сбор элементов в соответствии с поставленной целью, а также проведение дополнительных манипуляций с элементами (например, постановка метки начала и окончания события, меток обнаружения признаков, обозначений патологий и т. п.). Такую разметку проводят с участием обученного медицинского персонала путем ручного аннотирования содержания данных или их частей, которое может быть выполнено в графической или текстовой форме, либо в их комбинации.

4.2.3 Верифицированный набор данных

Верифицированный набор данных получают при дополнении набора данных, подготовленных при проспективной разметке, данными из медицинских записей, в том числе об окончательном и/или патологоанатомическом диагнозе. В качестве метода для верификации набора данных можно применять метод «золотого стандарта» (ground truth) для целевой патологии (см. [3], [4], [5]), повторное исследование пациента через определенное время, результаты патогистологических, иммунологических исследований и др. (см. [6]).

Верификация набора данных может быть также обеспечена путем слепого анализа набора данных экспертами с достижением заданного уровня согласованности их решений.

Выделяют следующие критерии отнесения набора данных к верифицированному набору данных:

- данные получены из реальной практики (не допускается получение синтезированных данных, например ЭКГ от генератора физиологических сигналов);
- данные получены в «сыром виде» — без применения фильтров и математических средств постобработки;

- структура набора данных соответствует поставленной цели его формирования (обучение, аналитическая, клиническая валидация (см. [6] и др.);
- количество наблюдений (исследований) достаточно для достижения статистической значимости результата;
- разметка и/или аннотирование проведены экспертной группой, соответствующей критериям 5.8.3;
- разметка и/или аннотирование проведены с использованием тезауруса (кодированной библиотеки типовых формулировок, соответствующих рекомендации ассоциации специалистов в данной области по ГОСТ 7.24, ГОСТ 7.25).

5 Рекомендуемые этапы подготовки набора данных

5.1 Введение

Подготовка набора данных должна состоять из набора процедур, выполнение которых позволяет достигнуть цели обучения и тестирования системы искусственного интеллекта (СИИ) с обеспечением качества набора данных (см. [3], [4], [6]).

В настоящем стандарте рассматривается процесс подготовки набора данных, который может быть изменен в условиях конкретных задач (рисунок 3).

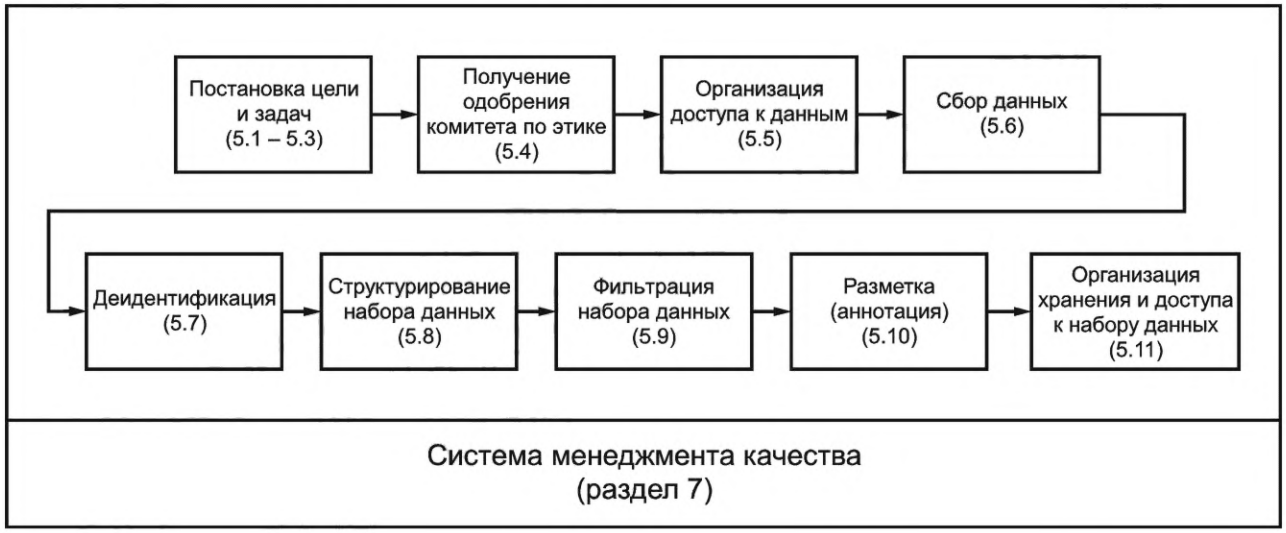

Рисунок 3 — Процесс подготовки набора данных для обучения и тестирования СИИ

5.2 Определение целей

Должна быть определена цель формирования набора данных, только тогда можно оценить, является ли доступ к данным или другая деятельность по их обработке допустимыми:
- какие данные допустимо собирать;
- как их следует использовать (применительно к каким задачам);
- кому их следует раскрывать (доступ третьими лицами);
- в течение какого времени они должны быть доступны.

Цели формирования набора данных могут быть разными, включая следующие:
- разработка СИИ, включающая этап обучения алгоритма искусственного интеллекта и выполнение внутреннего тестирования;
- выполнение аналитической или клинической валидации СИИ.

5.3 Постановка задачи

Постановка задачи подготовки набора данных должна включать определение предметной области и выбор методов обработки. Задача подготовки набора данных должна быть определена проблемой,

на решение которой направлено создание СИИ, классом СИИ или целью проведения тестирования, классификацией СИИ, а также задачей для СИИ (кластеризация, регрессия, рейтинг и др.).

Примечание — В клинической медицине используют два основных подхода машинного обучения, каждый из которых ассоциирован с определенным типом задач:

- при контролируемом машинном обучении (обучение с учителем или supervised machine learning) алгоритм наблюдает набор размеченных данных и обучается функции, позволяющей предсказывать аннотацию для новых входных данных. Возможными типами задач контролируемого машинного обучения являются классификация (см. 3.7) и регрессия. При регрессии аннотация может принимать любое действительное значение, а не ограничиваться конечным набором категорий, как при классификации;
- при неконтролируемом машинном обучении (обучение без учителя или unsupervised machine learning) алгоритм распознает паттерны (структуру) в неразмеченных данных. Возможными типами задач неконтролируемого машинного обучения являются кластеризация (группировка экземпляров данных в кластеры со сходными характеристиками) и детекция аномалий (идентификация редких экземпляров данных, существенно отличающихся от остальных).

5.4 Получение одобрения комитета по этике

Рекомендуется получить одобрение комитета по этике медицинской организации (МО) (при наличии такого органа в структуре МО) для сбора данных или использования де-идентифицированных данных с целью подготовки набора данных для разработки СИИ (включая этапы обучения алгоритма искусственного интеллекта и внутреннее тестирование на этапе разработки СИИ) и аналитической или клинической валидации СИИ.

5.5 Организация доступа к набору данных

МО, выполняющая подготовку набора данных, должна обеспечить доступ к требуемым данным, находящимся в медицинской информационной системе с возможностью выполнений операций поиска, чтения и сбора данных, а также выполнения де-идентификации (см. 5.7). Процесс организации доступа должен быть задокументирован, должны быть обеспечены процессы защиты данных, в том числе персональной информации согласно действующим нормативным правовым актам (НПА). Организация доступа должна обеспечивать скорость передачи данных, соответствующую целям и задачам такого доступа.

5.6 Сбор данных

Возможны два подхода к формированию наборов данных, в зависимости от поставленной цели:

а) представление медицинских данных (феноменов, синдромов, заболеваний, исходов), отражающее максимальную вариативность (то есть и частые, и редкие случаи представлены в одинаковом объеме). Данный подход должен быть применен в ходе формирования набора данных для аналитической валидации СИИ;

б) представление медицинских данных (феноменов, синдромов, заболеваний, исходов) согласно их частоте встречаемости, предтестовой вероятности, заболеваемости, распространенности в популяции. Данный подход должен быть применен в ходе формирования набора данных для клинической валидации СИИ.

5.6.1 Принцип сбора данных для аналитической валидации

- Набор данных для аналитической валидации должен быть подготовлен для определения следующих характеристик: производительность (например, время, затрачиваемое на обработку СИИ медицинского исследования при наличии функции автоматического расчета времени и т. д.), точность интерпретации исследований с учетом функциональных возможностей СИИ, повторяемость, воспроизводимость;
- Возможно включать в набор данных для аналитической валидации исследований с нарушением технологии (внешние помехи, артефакты, неверное наложение электродов/датчиков, нарушение последовательности регистрации, укладки пациента и т. п.). При этом такие исследования должны быть помечены должным образом. Метки возможно размещать в метаданных;
- При формировании набора данных следует использовать данные из разных МО и разных моделей/производителей оборудования, обработку данных с которых изготовитель СИИ включает в функциональное назначение. Это необходимо для снижения систематической ошибки, так как невключение в

набор данных элементов, получаемых, например на какой-либо модели оборудования, может привести к ограничениям в процессе использования набора данных.

5.6.2 Принцип сбора данных для клинической валидации

При проведении клинической валидации необходимо использовать верифицированный набор данных. В случае невозможности проведения данного вида валидации на верифицированном наборе данных необходимо представить соответствующее разъяснение причин и выбирать данные с иными методами разметки с учетом их ценности.

Принцип сбора данных для клинической валидации заключается в следующем:

а) соотношение «норма»/«патология» или разные заболевания в наборе данных определяют областью применения СИИ;

б) при формировании набора данных используют данные из разных МО и разных моделей/производителя оборудования. Это необходимо для снижения систематической ошибки, так как невключение в набор данных элементов, получаемых, например на какой-либо модели оборудования, может привести к ограничениям в процессе использования набора данных. Допустимо использовать данные из разных МО, обладающие одинаковой структурой, полученные в результате применения оборудования с одинаковым процессом работы (одинаковая модель/производитель), в случае указания этого в разделе области применения документации СИИ;

в) демографические, социально-экономические характеристики и основные показатели здоровья пациентов (репрезентативная выборка), чьи данные включают в набор данных, должны соответствовать усредненным характеристикам популяции территории, на которой планируется использование СИИ;

г) планируемый размер набора данных должен быть обоснован в документации испытаний, исходя из статистических соображений и желаемой точности оценки основных метрик:

- для алгоритмов бинарной классификации, включая данные с асимметричным распределением, а также в случае многомерных данных рекомендуемой метрикой является площадь под ROC-кривой (либо, в зависимости от типа данных, объем под поверхностью ROC-кривой или площадь ложно-положительных результатов под ROC-кривой, false discovery rate-controlled AUROC) со значением не менее 0,8;

- для алгоритмов бинарной классификации данных медицинской визуализации рекомендуется использовать метод ROC со свободным откликом (free-response ROC, FROC). СИИ с допустимой точностью соответствуют выпуклые FROC-кривые с углом наклона более 45°; графики FROC, представляющие собой прямую с углом наклона 45°, свидетельствуют о недостаточном качестве СИИ;

- для наборов данных с плохим балансом между классами (например, незначительное количество примеров патологии при существенно превосходящем числе экземпляров данных без отклонений) рекомендуется строить зависимость положительного предсказательного значения (positive predictive value) от чувствительности. Предпочтительным значением площади под кривой является 0,5;

- для алгоритмов, выполняющих функцию регрессии, рекомендуется использовать метрики средней абсолютной ошибки (mean absolute error), среднеквадратичной ошибки (mean squared error) и стандартного отклонения всех ошибок (root mean squared error). Во всех случаях значения должны стремиться к нулю для СИИ с высокой точностью. Другой допустимой метрикой является коэффициент детерминации R^2: если он равен 0, то предсказание аннотации невозможно, R^2 равно 1 означает полностью точное предсказание без изменчивости. Рекомендуемое значение R^2 для СИИ должно быть не менее 0,6;

д) МО должна иметь возможность самостоятельного формирования наборов данных для независимой клинической валидации СИИ без применения наборов данных, которые были использованы на этапе разработки СИИ, включая обучение алгоритма искусственного интеллекта и его внутреннее тестирование на этапе разработки СИИ.

5.7 Де-идентификация (обезличивание)

С целью применения набора данных для обучения и тестирования СИИ элементы набора данных не должны содержать какую-либо персональную информацию согласно действующим нормативно-правовым актам (НПА). Любая персональная информация должна быть удалена как из метаданных, так и из исходных данных. Согласно действующим НПА субъект персональных данных должен предоставить согласие на их использование для целей разработки СИИ.

Необходимо проводить также удаление из данных номера полиса обязательного медицинского страхования застрахованного лица, наименования МО, фамилии, имени, отчества пациента, места проживания, сведений о дате исследования и дате рождения. Дату рождения рекомендуется заменить на возраст (годы, месяцы) на момент исследования, чтобы была возможность сбора данных по возрасту пациента. Также должны быть удалены любые иные идентификаторы, с помощью которых потенциально возможно установить личность пациента (см. [3]).

Де-идентификация данных должна быть произведена в МО, в которой было проведено медицинское исследование, при условии наличия согласия пациента на обработку его персональных данных, включая де-идентификацию (обезличивание).

Пример — Де-идентификация метаданных изображений в формате DICOM проводится в соответствии с ГОСТ Р ИСО 17432—2009.

5.8 Структурирование набора данных

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.

5.9 Фильтрация набора данных

Качество СИИ зависит от качества данных, используемых для разработки СИИ. Этап фильтрации набора данных позволяет снизить затраты на разметку данных за счет исключения данных, не соответствующих заданным параметрам.

Процедура контроля качества (см. 7.4.1) включает нахождение, предотвращение и устранение проблем, связанных с качеством наборов данных.

Примеры проблем, связанных с наборами данных в клинической медицине

- Проблемы с метаинформацией о выполненном исследовании и пациенте (пропущенные значения, некорректные идентификаторы, некорректные значения тегов DICOM и т. д.);

- Проблемы с качеством исходных данных (смазанные изображения, неверные настройки оборудования, нерелевантные данные).

Фильтрацию и контроль качества наборов данных возможно осуществлять с помощью визуального контроля, специальных инструментов (например, DICOM-валидаторов), а также с использованием СИИ (например, для автоматической оценки качества изображения).

5.10 Разметка [аннотация] данных

5.10.1 Общие требования

Существующая классификация выполняемых разметок (аннотации) данных приведена в 5.1, решение о выборе вида разметки выполняется на этапе постановки цели и задачи формирования набора данных.

Существует ряд подходов к аннотации медицинских данных:

а) полуструктурированное текстовое описание визуальных наблюдений с указанием содержащих их анатомических объектов и типов нарушений.

Пример — Легочная паренхима: увеличивающееся образование размером 2,3 × 2,7 см, прилегающее к малой трещине в правой средней доле.

По причине возможных вариаций в используемой терминологии и структуре описаний, а также ориентировочной локализации наблюдений, автоматический поиск по таким аннотациям, а также использование их СИИ осложнены и малоэффективны;

б) структурированная аннотация, которая должна использовать согласованный набор терминов, для снижения вариабельности интерпретаций визуальных наблюдений.

Структурированная аннотация может быть сопровождена конкретизированной информацией о локализации наблюдений, которую могут выполнять с разным уровнем точности и детализации:

- с грубой локализацией — приблизительное обозначение координат объектов интереса, посредством задания ограничивающего параллелепипеда или эллипсоида;

- с полной сегментацией на основе маски минимальных элементов, обозначающей положение объекта интереса на фоне остальной части данных.

Стандартизированные методы аннотации приведены в приложении В.

5.10.2 Первичная разметка

В рамках проведения первичной разметки необходимо отметить и охарактеризовать все целевые структуры в подготовленном наборе данных.

Первичная разметка должна включать структурированные аннотации и быть выполнена в соответствии с установленными регламентами, характеризующими доступ к данным, используемые программные средства и методы разметки (см. [7]), а также шаблон выполнения аннотации элементов набора данных.

5.10.3 Экспертная валидация

Экспертную валидацию следует выполнять с привлечением экспертной группы в целях проверки и корректировки результатов первичной разметки.

Существуют две группы экспертных оценок:

- индивидуальные оценки основаны на использовании мнения отдельных экспертов, независимых друг от друга;
- коллективные оценки основаны на использовании коллективного мнения экспертов.

Основные этапы обработки экспертных оценок:

- определение компетенции экспертов;
- определение обобщенной оценки;
- построение обобщенной ранжировки объектов в случае нескольких оцениваемых объектов или альтернатив;
- определение зависимостей между ранжировками;
- оценка согласованности мнений экспертов. При отсутствии значимой согласованности экспертов необходимо выявить причины несогласованности (наличие групп) и признать отсутствие согласованного мнения (ничтожные результаты);
- оценка ошибки исследования;
- построение модели свойств объекта (объектов) на основе ответов экспертов (для аналитической экспертизы);
- подготовка отчета (с указанием цели исследования, состава экспертов, полученной оценки и анализа результатов).

5.11 Организация хранения и доступа к верифицированному набору данных

5.11.1 Метаданные

Метаданные применяют для классификации, упорядочения и описания данных. Метаданные должны быть составлены согласно принципам, содержащим базовые принципы улучшения возможностей поиска, обеспечения доступа к данным, их совместимости и повторного использования данных (см. [8]).

При аннотировании медицинских данных необходимо использовать библиотеки типовых формулировок (тезаурусы).

Список рекомендуемых метаданных для хранения медицинских изображений приведен в приложениях А, Б.

5.11.2 Организация хранения набора данных

Данные необходимо передавать в локальное хранилище (одноцентровое исследование) либо во внешнее хранилище данных (многоцентровое исследование). Хранение данных может быть организовано на локальном сервере или с использованием облачного хранения (ГОСТ Р ИСО/МЭК 17826). При этом доступность и безопасность обеспечивают на лучшем уровне при использовании локального сервера; совместное использование данных и резервное копирование возможно при использовании облачного хранения.

5.11.3 Доступ к верифицированному набору данных

Согласно ГОСТ Р ИСО 27799 статистические и научные данные, включая де-идентифицированные (обезличенные) данные, полученные посредством удаления идентифицирующих данных из персональной медицинской информации, должны быть защищены.

Должны быть установлены стандартные процедуры доступа к набору данных для третьих лиц, закрепленные в документе о политике по защите информации. При организации доступа к набору данных необходимо подписывать соглашение с МО, формирующей наборы данных.

6 Требования по использованию верифицированных наборов данных для обучения и тестирования систем искусственного интеллекта

6.1 Общие требования к описанию наборов данных

Набор медицинских данных должен содержать следующие сведения (описательного характера):

1) номер свидетельства о государственной регистрации базы данных в качестве результата интеллектуальной деятельности (рекомендательно);

2) характеристика популяции (возрастно-половые показатели, этнический состав, регионы проживания и т. д.); сведения о де-идентификации; сведения о МО, послуживших источниками для формирования базы данных; сведения о факторах риска;

3) характеристика исследования: анатомическая область(и), модальность, проекции, типы медицинских изделий — диагностических приборов, виды и характеристики протоколов исследования;

4) целевая патология согласно Международной классификации болезней (либо наименование феноменов в соответствии с рекомендациями профильной ассоциации специалистов), если применимо в соответствии с поставленной целью (см. 5.1);

5) общее количество клинических случаев, исследований, изображений, документов и их распределение по диагностическим группам;

6) соотношение случаев «норма»/«патология» (случаи «патология» могут быть разделены на несколько подклассов), если применимо в соответствии с поставленной целью (см. 5.1);

7) сведения о верификации (патогистологическом или ином окончательном диагнозе);

8) методология разметки.

Примеры рекомендованных параметров для описания наборов данных для медицинских изображений приведены в приложении А, для области клинической физиологии — в приложении Б.

6.2 Разделение набора данных на обучающую и тестовую выборки на этапе разработки системы искусственного интеллекта

В процессе разработки СИИ возможно использование обучающей, тестовой и в некоторых случаях проверочной выборок, которые выделены из одного или нескольких наборов данных.

Внутреннее тестирование СИИ должно быть проведено на наборе данных, который не был использован для обучения. Это необходимо для исключения явления переобучения, при котором в результате тестирования СИИ получается смещенная оценка.

Обучающая и тестовая выборки должны быть независимы, что обеспечит получение несмещенной оценки при тестировании СИИ.

Также в некоторых случаях используют проверочный набор данных для выбора оптимальной модели в процессе разработки СИИ.

Примечание — Возможно выполнение разделения набора данных на обучающую, проверочную и тестовую выборки в соотношениях 80/10/10 или 70/15/15, что определяется поставленной целью разработки СИИ.

6.3 Требования по размеру набора данных для обучения и тестирования на этапе разработки систем искусственного интеллекта

Размер набора данных должен быть определен поставленной целью его применения и зависит от таких факторов, как требуемое качество предсказаний СИИ, тип и архитектура алгоритма СИИ, количество параметров алгоритма СИИ, качество данных, включая качество аннотаций, распределение метрик и уровень шума в наборе данных. Например, изображения с более высоким разрешением (количеством пикселей) также усложняют архитектуру алгоритма, что обуславливает необходимость использования обучающих выборок большего размера. Теоретические обоснования оценки необходимого и достаточного размера набора данных в зависимости от указанных факторов находятся на стадии разработки. Возможно использование автоматизированных средств расчета размера набора данных,

которые основаны, например на использовании ширины 95 % доверительного интервала и допустимой ширины определения метрик.

Возможно формирование набора данных для обучения СИИ эмпирическим методом, используя правило, согласно которому размер набора данных должен в несколько раз превышать количество параметров алгоритма СИИ либо соответствовать другим обоснованным критериям (см. [4], [9]). Большее количество данных обеспечивает лучшее представленние комбинаций оцениваемых метрик и их вариаций, а также их принадлежности целевой(ым) патологии(ям) или прочим визуальным наблюдениям.

Набор данных для тестирования на этапе разработки СИИ должен быть обоснованного размера, определенного поставленной целью.

6.4 Требования к наборам данных для внешнего тестирования системы искусственного интеллекта

Набор данных для внешних тестирований должен быть обоснованного размера, определенного поставленной целью и дизайном испытаний.

Примечание — Внешние тестирования могут быть в том числе техническими испытаниями, предварительными клинико-техническими испытаниями, клиническими испытаниями и мониторингом.

Для оценки эксплуатационных характеристик СИИ набор данных для внешнего тестирования должен содержать данные эксплуатационных категорий. Допустимо добавлять в наборы данных дополнительные тест-случаи (контрольные тесты), соответствующие ситуациям, сложным для классификации экспертами: данные с высоким уровнем шума либо с ухудшенными характеристиками (например, в результате сбоя оборудования), изображения с недостаточной видимостью целевых объектов. Включение таких данных позволит проверить робастность СИИ, в дополнение к заявленным эксплуатационным характеристикам.

Наборы данных для внешнего тестирования СИИ не должны включать элементы из наборов данных для обучения данной СИИ.

До завершения этапа внешнего тестирования должен быть ограничен доступ к применяемому на данном этапе набору данных.

Примечание — Для выполнения этих требований должна быть обеспечена коммуникация между коллективами, выполняющими разработку и внешнее тестирование СИИ. Коллектив, выполняющий внешнее тестирование СИИ, самостоятельно определяет объем необходимой информации, сообщаемой коллективу, выполняющему обучение, руководствуясь исключением риска внедрения систематической ошибки (bias) на этапе тестирования.

6.5 Требования по характеристикам наборов данных

6.5.1 Размерность, разреженность, разрешение

Под размерностью набора данных понимают количество атрибутов, которые имеют объекты в наборе данных (например, значение артериального давления, масса тела пациента, уровень холестерина и др.). Наборы данных с высокой размерностью (с большим количеством атрибутов) выдвигают повышенные требования к алгоритмам СИИ, допустимому размеру таких наборов, а также к вычислительным ресурсам для их обработки. В зависимости от поставленной цели и дизайна исследования допустимо обоснованное снижение размерности набора данных, в частности за счет кластеризации данных либо группировки взаимосвязанных по какому-либо признаку атрибутов в объединенные категории.

Отсутствующие данные способны существенно осложнить для СИИ задачу поиска и категоризации объектов интереса. Характеристики размерности и разреженности должны быть подобраны под задачу. Для разработки СИИ всегда должны быть использованы данные в одном формате/разрешении/качестве.

6.5.2 Баланс данных, распределение классов

Сбалансированный набор данных должен содержать одинаковое количество примеров различных категорий (классов) объектов интереса, включая примеры неизмененных тканей (см. [10], [11]). В случае бинарной классификации это может соответствовать распределению 50/50 для случаев «патология»/«норма».

Для контроля дисбаланса классов необходимо использовать взятие подвыборок (понижающих выборок) — равных количеств случайно отобранных примеров каждого исследуемого класса. Допускается отхождение от сохранения баланса данных в некоторых сценариях исследования: например, при определении площади под кривой (AUC). Если набор данных нельзя сбалансировать, используются

техники обучения алгоритмов искусственного интеллекта (ИИ) по несбалансированным данным (under-sampling, over-sampling, модификации процесса обучения и др.).

Требования к балансу набора данных необходимо определять в соответствии с поставленной задачей (см. 5.1, 5.2). Требования по соотношению категорий (классов) для аналитической и клинической валидации в соответствии с 5.6.1, 5.6.2.

7 Система менеджмента качества при разработке и применении набора данных

7.1 Общие положения

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 организацией применительно к качеству.

7.2 Требования к персоналу

7.2.1 Требования к персоналу, выполняющему разметку

Персонал, осуществляющий деятельность, влияющую на качество подготовки набора данных, должен быть компетентным в соответствии с полученным образованием, подготовкой, навыками и опытом согласно ГОСТ ISO 13485. 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 (ЧСС, 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.

7.2.2 Требования к экспертам

В экспертную группу должны входить специалисты с большим опытом 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 экспертов следует учитывать наличие конфликтов интересов, которые могут стать существенным препятствием для получения объективного суждения.

Требования к экспертам, которых привлекают к подготовке набора данных, должны быть документированы в рамках СМК.

7.3 Требования к аппаратному обеспечению

Организация, проводящая подготовку и применение наборов данных, должна документировать требования к аппаратному обеспечению, необходимому для достижения соответствия требованиям к набору данных, организации хранения и управления наборами данных (включая ввод/вывод и обработку).

Аппаратное обеспечение включает рабочее пространство и связанные с ним системы инженерного обеспечения; оборудование, включая технические и программные средства; вспомогательные услуги (например, связь или информационные системы).

7.4 Контроль качества

7.4.1 Контроль качества при подготовке 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 участие в разметке и/или аннотировании, но будет регулировать срочность, очередность и объем рабо-

ты 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:

- проверка отсутствия пропусков элементов в наборе данных;
- проверка отсутствия некорректных элементов для решения поставленных задач;
- проверка качества элементов набора данных рекомендованным критериям профессионального медицинского сообщества.

7.4.2 Контроль качества при применении наборов данных

Должны быть подготовлены и внедрены стандартные процедуры применения наборов данных в рамках СМК. Должны быть указаны требования по организации доступа к наборам данных, в том числе реестр лиц, которые получили к нему доступ.

7.5 Управление изменениями наборов данных

После создания и регистрации набора данных может возникнуть необходимость внести изменения — например, в результате обнаружения ошибок или добавления новых данных (см. [12]). При внесении любых изменений необходимо документировать изменение версии набора данных.

Примечание — Это позволит избежать множества проблем, например связанных с невоспроизводимостью или некорректным сравнением результатов, полученных на разных версиях наборов данных.

Изменения в наборах данных должны быть задокументированы, эта документация должна быть приложена к набору данных.

Приложение А
(справочное)

Рекомендованный список метаданных для хранения верифицированного набора медицинских изображений (см. [4], [6])

1 Тип изображения:
- вид исследования (компьютерная томография, рентгеновское исследование и т. п.);
- разрешение;
- общее число изображений и по сериям.

2 Число исследований.
3 Источники исследований:
- оборудование;
- типы оборудования;
- МО.

4 Параметры сканирования изображений
5 Параметры хранения изображений:
- формат данных;
- уровень и тип сжатия данных.

6 Аннотация (разметка):
- тип;
- что и как описано;
- привлеченная экспертная группа.

7 Контекст.
8 Как определена истинная разметка и промаркирована.
9 Связанные данные:
- демографические;
- клинические;
- лабораторные;
- геномные;
- временные;
- принимаемые препараты (лекарства);
- другие.

10 Временной диапазон сбора изображений (дата и время исследования).
11 Использование данных:
- какое программное обеспечение использовать для просмотра данных.

12 Кому принадлежат данные.
13 Кто ответственен за данные.
14 Допустимое использование.
15 Назначение набора данных.
16 Информация об одобрении комитета по этике.
17 Информация о де-идентификации набора данных.
18 Информация о проведенном контроле качества набора данных.
19 Параметры доступа:
- доступность;
- цена и лицензионные соглашения.

20 Распределение случаев (если применимо):
- процент «норма/патология» (код МКБ);
- данные патологии: число исследований с каждой патологией.

Приложение Б
(справочное)

Рекомендованный список метаданных для хранения верифицированного набора физиологических данных

1 Параметры регистрации данных:
- модальность;
- длительность (продолжительность) регистрации;
- разрешение;
- частота дискретизации;
- частотный диапазон регистрации (диапазон пропускания сигнала);
- динамический диапазон;
- разрядность аналого-цифрового преобразователя;
- наличие калибровочных сигналов;
- количество и маркировка каналов (отведений), если в разных отведениях (каналах) регистрируются разные модальности (указать);
- схемы монтажей отведений (указать);
- в случае выполнения функциональных проб (указать действующий агент, протокол применения).

2 Число исследований.

3 Источники исследований:
- оборудование;
- типы оборудования;
- медицинские организации.

4 Параметры хранения данных:
- формат данных;
- уровень и тип сжатия данных;

5 Аннотация (разметка):
- тип;
- использованный словарь или тезаурус.

6 Контекст.

7 Связанные данные:
- демографические;
- клинические;
- лабораторные;
- других инструментальных методов исследования;
- геномные;
- принимаемые препараты (лекарства);
- результаты хирургического лечения;
- другие.

8 Временной диапазон:
- дата и время исследования;
- длительность исследования.

9 Использование данных:
- какое программное обеспечение использовать для просмотра данных.

10 Кому принадлежат данные.

11 Кто ответственен за данные.

12 Допустимое использование.

13 Назначение набора данных.

14 Информация об одобрении комитета по этике.

15 Информация о де-идентификации набора данных.

16 Информация о проведенном контроле качества набора данных.

17 Параметры доступа:
- доступность;
- цена и лицензионные соглашения.

18 Распределение случаев (если применимо):
- процент «норма/патология» (код МКБ);
- данные патологии: число исследований с каждой патологией.

Приложение В
(справочное)

Стандартизированные методы аннотации

Существует два основных международных стандартизированных метода аннотации: «аннотация и разметка изображений» (annotation and image markup, AIM) и «состояние представления DICOM» (DICOM Presentation State, PS).

AIM

AIM использует три базовых концепта:

- визуальные наблюдения: «масса», «поражение», «очаг»;
- анатомические объекты: «затылочная доля», «теменная доля», «медиальный сегмент средней доли правого легкого»;
- интерференция (нарушение): «поражение речевого центра», «плевральный выпот», «пневмония».

Визуальным наблюдениям и анатомическим объектам задают характеристики. «Предполагаемый», «кистозный» — характеристики наблюдений. «Расширенный», «разорванный» — примеры характеристик объектов. После задания характеристик наблюдений и объектов проводят их количественную оценку. Ее допустимо выражать в терминах «присутствует», «отсутствует», «не применимо», либо квартиль/процентиль, либо в произвольной шкале и др.

В стандарте AIM проводят совмещение этой описательной информации с графическими символами, располагаемыми экспертами на самом изображении, в единый тип данных.

DICOM PS

Presentation State (PS) — это независимый экземпляр класса типовой инструкции DICOM, который содержит информацию о том, как должно отображаться конкретное изображение с использованием всех возможных параметров и визуальных элементов, определенных в стандарте DICOM. Позволяет без потерь вернуться к оригинальному изображению, поскольку никак не модифицирует пиксельные данные.

Библиография

[1] IMDRF/SaMD WG/N41 — Software as a Medical Device (SaMD): Clinical Evaluation, 2017

[2] Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации»

[3] Willemink M.J. Preparing Medical Imaging Data for Machine Learning/M.J. Willemink, W.A. Koszek, C. Hardell. Radiology. 2020; 295(1):4-15

[4] Ranschaert E.R. Artificial Intelligence in Medical Imaging. Opportunities, Applications and Risks/E.R. Ranschaert, S.P. Morozov, R. Paul. Springer Nature Switzerland AG — 2019 — p. 705

[5] 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

[6] Kohli M.D. Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session/M.D. Kohli, R.M. Summers, J. Geis // Journal of Digital Imaging. 2017; 30:392—399

[7] Morozov S.P., Gombolevskiy V.A., Elizarov A.B., et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans. Computer Methods and Programs in Biomedicine. 2021. Vol. 206, P.106111

[8] Wilkinson M.D. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016; 3:160018

[9] Smith S.M., Nichols T.E. Statistical Challenges in “Big Data” Human Neuroimaging. NEUROVIEW. — Vol. 97. — No 2. — P. 263—268

[10] Medical Big Data and Internet of Medical Things. Advances, Challenges and Applications/E.H. Aboul [et al.]. Taylor & Francis Group., 2019. 357 c.

[11] Johnson, J.M. Survey on deep learning with class imbalance/J.M. Johnson, T.M. Khoshgoftaar // Journal of Big Data. 2019; 6 (27):1-54

[12] Klump J., Wyborn L., Wu M., Martin J., Downs R.R., Asmi A. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. Data Science Journal. — 2021. — V. 20. — No. 12; 1—13 p.

УДК 615.841:006.354 ОКС 11.040.01

Ключевые слова: система искусственного интеллекта, набор данных, метаданные, контроль качества

Редактор *Н.А. Аргунова*
Технический редактор *И.Е. Черепкова*
Корректор *И.А. Королева*
Компьютерная верстка *М.В. Малеевой*

Сдано в набор 01.04.2022. Подписано в печать 07.04.2022. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 2,79. Уч.-изд. л. 2,51.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «РСТ»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru