ОРИГИНАЛЬНАЯ СТАТЬЯ
DOI: 10.21045/1811-0185-2023-4-28-41
УДК: 614.2

ОСНОВОПОЛАГАЮЩИЕ ПРИНЦИПЫ СТАНДАРТИЗАЦИИ И СИСТЕМАТИЗАЦИИ ИНФОРМАЦИИ О НАБОРАХ ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ В МЕДИЦИНСКОЙ ДИАГНОСТИКЕ

Ю.А. Васильев[1], Т.М. Бобровская[2] ✉, К.М. Арзамасов[3], С.Ф. Четвериков[4], А.В. Владзимирский[5], О.В. Омелянская[6], А.Е. Андрейченко[7], Н.А. Павлов[8], Л.Н. Анищенко[9]

[1, 2, 3, 4, 5, 6, 7, 8, 9] Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», г. Москва, Россия.

[1] https://orcid.org/0000-0002-0208-5218; [2] https://orcid.org/0000-0002-2746-7554;
[3] https://orcid.org/0000-0001-7786-0349; [4] https://orcid.org/0000-0002-3097-8881;
[5] https://orcid.org/0000-0002-2990-7736; [6] https://orcid.org/0000-0002-0245-4431;
[7] https://orcid.org/0000-0001-6359-0763; [8] https://orcid.org/0000-0002-4309-1868;
[9] https://orcid.org/0000-0002-2057-0452

✉ *Автор для корреспонденции: Бобровская Т.М.*

АННОТАЦИЯ

Обоснование: Активное внедрение технологий искусственного интеллекта в сферу здравоохранения, которое мы наблюдаем в последние годы, способствует резкому росту количества медицинских данных, собираемых для разработки моделей машинного обучения, в том числе данных лучевой и инструментальной диагностики. Для решения различных задач в области цифровых медицинских технологий посредством алгоритмов машинного обучения создаются все новые и новые наборы данных, поэтому становятся актуальными проблемы их систематизации и стандартизации, хранения, доступа, рационального и безопасного использования.

Цель: разработка подхода к систематизации и стандартизации информации о наборах данных для решения вопросов представления, хранения, применения и оптимизации использования наборов данных и обеспечения безопасности и прозрачности процессов разработки и испытаний медицинских изделий с использованием искусственного интеллекта.

Методы: анализ собственного и мирового опыта по созданию и использованию медицинских наборов данных, поиск и анализ медицинских справочников, разработка и обоснование структуры реестра, поиск научных публикаций с ключевыми словами «наборы данных», «реестр медицинских данных», размещенных в реферативных базах данных РИНЦ, Scopus, Web of Science.

Результаты. Разработана структура реестра наборов данных в медицинской инструментальной диагностике с разделами, отражающими информацию по всем этапам формирования и использования наборов данных для машинного обучения: 7 параметров на этапе инициирования, 8 – на этапе планирования, 70 – карточка набора данных, 1 – смена версии, 14 – на этапе использования, всего – 100 параметров. В работе предлагается классификация наборов данных по цели их создания, классификация методов верификации данных, а также принципы формирования названий для стандартизации и наглядности представления наборов данных. Кроме того, освещены основные особенности организации ведения данного реестра: управление, качество данных, конфиденциальность и безопасность.

Выводы. Впервые предлагается оригинальная технология структуризации и систематизации управления медицинскими наборами данных для инструментальной диагностики, в основу которой положены разработанная терминология и принципы классификации информации, что позволяет стандартизировать структуру информации о наборах данных для машинного обучения, обеспечивает централизацию хранения, удобный и быстрый доступ ко всей информации о наборе данных, а также прозрачность, надежность и воспроизводимость результатов в сфере искусственного интеллекта. Создание реестра дает возможность оперативно формировать наглядные библиотеки данных, позволяя обширному кругу исследователей, разработчиков и компаний выбирать наборы данных для своих задач, что обеспечивает их широкое использование, оптимизацию ресурсов и способствует быстрому развитию и внедрению искусственного интеллекта.

Ключевые слова: *набор данных, искусственный интеллект, машинное обучение, реестр, библиотеки наборов данных для машинного обучения.*

Для цитирования: *Васильев Ю.А., Бобровская Т.М., Арзамасов К.М., Четвериков С.Ф., Владзимирский А.В., Омелянская О.В., Андрейченко А.Е., Павлов Н.А., Анищенко Л.Н. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике // Менеджер здравоохранения. 2023; 4: 28–41. DOI: 10.21045/1811-0185-2023-4-28-41.*

Конфликт интересов. *Авторы заявляют об отсутствии конфликта интересов.*

© *Васильев Ю.А., Бобровская Т.М., Арзамасов К.М., Четвериков С.Ф., Владзимирский А.В., Омелянская О.В., Андрейченко А.Е., Павлов Н.А., Анищенко Л.Н., 2023 г.*

Введение

Различные технологии глубокого машинного обучения, искусственного интеллекта (ИИ), в частности технологии компьютерного зрения, активно внедряются в последние годы практически в каждую сферу нашей жизни. Одним из наиболее социально значимых направлений их применения являются медицина и организация здравоохранения [1]. Для диагностики, лечения и профилактики заболеваний, а также для стандартизации и повышения точности формирования медицинских документов ежегодно разрабатывается большое количество различных алгоритмов, основанных на технологиях ИИ [2]. Использование технологий ИИ способствует созданию условий для улучшения жизни населения, в т.ч. за счет повышения качества услуг в здравоохранении [3].

С развитием медицины, повышением ее доступности, а также повсеместного внедрения цифровых технологий в медицинскую практику [4–6] отмечается высокий рост количества медицинских данных: клинических, лабораторных и инструментальных [7].

Наиболее активно разработки ИИ используются в максимально стандартизированных и цифровизированных исследованиях, таких как лабораторные (патоморфологические исследования), лучевые (магнитно-резонансная, компьютерная томография, рентгенография, маммография, флюорография, ультразвуковые исследования) и сигнальные (ЭКГ, ЭЭГ, ЭНМГ, ФВД) методы диагностики [5, 8–11].

Пандемия COVID-19, начавшаяся в 2020, также показала, насколько важно и актуально оперативно реагировать на появление новых, не описанных ранее, заболеваний. Внедрение технологий ИИ [12, 13] в лучевую диагностику позволило снизить нагрузку на врачей и увеличить скорость обработки заключения в условиях острой нехватки медицинского персонала [14, 15].

Появление большого количества новых алгоритмов машинного обучения требует создания репрезентативных, релевантных и корректно размеченных наборов данных (НД) для разработки, дообучения и валидации этих алгоритмов, а также развития информационно-коммуникационной инфраструктуры для обеспечения доступа к таким данным [3, 16]. Производительность моделей ИИ зависит не только от количества данных, на которых он обучался, но и от их качества, обобщающей способности, структурированности и репрезентативности [17]. Цифровизация здравоохранения в Российской Федерации позволяет активно продвигать проекты по созданию эталонных НД, необходимых для успешного развития технологий ИИ и внедрения их в клиническую практику [3, 5, 16]. Количество новых НД ежегодно растет: только в ГБУЗ «НПКЦ ДиТ ДЗМ» в 2020 году было создано более 50 НД, а в 2021 уже более 120. Такое увеличение количества медицинских данных требует создания удобных инструментов для их хранения, администрирования и использования.

При подготовке НД для тестирования сервисов ИИ, а также решения ряда других научных задач, нами были сформулированы следующие проблемы:

1. Отсутствие единых стандартов представления информации о НД.

Успешное применение ИИ основывается на медицинских понятиях, требующих стандартизации и нормализации [18]. Современная процедурная терминология (Current Procedural Terminology) [19] предлагает стандартизированную номенклатуру и коды для медицинской визуализации, а онтологии обеспечивают семантическое отношение между терминами. На сегодняшний день существует множество справочников и словарей, разработанных с целью удобства представления данных и обеспечения электронного обмена медицинской информацией (например, SNOMED [20], LOINC [21], RadLex [22]), однако они имеют ряд ограничений [23], и многие пренебрегают их использованием, зачастую ограничиваясь лишь использованием Международной Классификацией Болезней (МКБ-10).

Кроме того, публикация каждого готового НД должна сопровождаться соответствующей документацией в виде текстового файла (readme-файл), в котором описаны основные параметры НД. На сегодняшний день единых стандартов такой документации не существует. Зачастую в readme-файлах упускается важная информация, которая могла бы позволить конечному пользователю принять решение, о применимости данного НД в его задачах. Или, наоборот, такой файл может содержать избыточную, несистематизированную информацию, что также затрудняет процесс поиска необходимых данных. В предыдущей работе была предложена базовая структура readme-файла[16].

2. Нерациональное использование данных и отсутствие централизованного хранения НД и информации о них.

Разметка результатов одного диагностического исследования – это дорогостоящая и трудозатратная процедура, поэтому необходимо обеспечить долгосрочное, надежное и централизованное

хранение данных с целью их возможного повторного использования для других задач, в том числе другими исследователями [24]. Кроме того, «разумная бережливость» – один из принципов развития и использования технологий ИИ [3], однако многие медицинские и научные учреждения, имея качественные и актуальные НД для решения задач в рамках машинного обучения, часто используют локальные специальные схемы кодирования, которые ограничивают повторное использование, в том числе и другими организациями [25].

3. Отсутствие НД для публичного использования.

Проблема отсутствия доступа к данным возникает не только в рамках одного учреждения, но и при сотрудничестве в области научных или коммерческих разработок, а это противоречит принципам поддержки конкуренции между организациями, осуществляющими деятельность в области ИИ [3]. Для продвижения научных исследований в этой области необходимо развитие исследовательской инфраструктуры и обеспечение доступа к НД посредством создания общедоступных платформ для их хранения, а также разработка унифицированных методологий описания, сбора и разметки данных и механизмов их контроля [3]. Для этих целей существуют библиотеки НД, которые позволяют собрать и предоставить краткую систематизированную информацию о НД (карточка НД) и сами НД для публичного использования. Однако из-за отсутствия единых стандартов представления данных, а также их разрозненного хранения часто данные не систематизированы и неудобны для изучения, например, НД на ресурсах https://github.com/, https://www.kaggle.com/. База данных Национального Института Рака (https://imaging.datacommons.cancer.gov/collections/) удобнее в использовании благодаря единому стилю представления данных, однако в карточках НД все же не хватает информации для принятия решения о возможности их применения в конкретных задачах машинного обучения, а названия НД неудобны в обращении, так как не отражают ключевую информацию для идентификации. Чтобы создавать удобные востребованные библиотеки необходим единый надежный источник информации о НД для синхронизации и предоставления в публичном поле.

Для решения этих проблем была поставлена следующая цель: разработать реестр данных инструментальной диагностики, предназначенных для разработки, дообучения и тестирования алгоритмов ИИ.

Материалы и методы

Данная работа является аналитическим исследованием, направленным на систематизацию и стандартизацию информации о НД для медицинской диагностики.

На первом этапе был проведен селективный анализ литературы: поиск и анализ научных публикаций с ключевыми словами «наборы данных», «реестр медицинских данных», «dataset», «register», размещенных в реферативных базах данных РИНЦ, Scopus, Web of Science с 2000 по 2022 год.

На втором этапе был проведен поиск и анализ медицинских справочников и приказов: федеральный справочник инструментальных диагностических исследований, федеральный справочник анатомических локализаций, тезаурус радиологических терминов RadLex, систематизированная машинно-обрабатываемая медицинская номенклатура SNOMED, база данных для идентификации медицинских врачебных и лабораторных наблюдений LOINC, МКБ-10, алфавитный указатель к международной статистической классификации болезней и проблем, связанных со здоровьем, справочник услуг ЕРИС, стандарт DICOM, номенклатура медицинских услуг [26], указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» [3].

На третьем этапе был проведен анализ библиотек медицинских НД, непосредственно самих НД, находящихся в открытом доступе (https://imaging.datacommons.cancer.gov/ и выложенных на ресурсах https://github.com/, https://www.kaggle.com/), а также их сопроводительной информации. Проведен анализ мирового опыта формирования репозиториев, таблиц НД и баз данных, а также изучена специфика формирования названий файлов в базах данных.

На четвертном этапе проведен анализ собственного опыта создания и использования НД в области лучевых и сигнальных методов диагностики [16, 27]. Выполнена аналитика сопроводительной информации (технических заданий, требований, readme-файлов, карточек НД и прочих документов), проанализирован наш опыт формирования номенклатуры наименований НД для инструментальной диагностики.

Структура и наполнение реестра разрабатывалась в соответствии с этапами жизненного цикла НД [27]. Жизненный цикл НД – это последовательность этапов, которую конкретная часть данных проходит от начального этапа создания

или получения до момента архивации или удаления [28].

1. Этап Инициации.

При возникновении необходимости создания НД по определенному направлению медицинской диагностики ответственными экспертами формируются или утверждаются к использованию ранее подготовленные базовые диагностические требования (БДТ). БДТ содержат информацию о целевой патологии, определяемой алгоритмами, а также о формате предоставления результата их работы. На их основе будет разработано техническое задание (ТЗ) на создаваемый НД (требования к составу, количеству исследований, типам и способам разметки и другая техническая информация).

2. Этап планирования работ по формированию НД.

На данном этапе планируются сроки проведения работ по созданию НД, разрабатывается ТЗ, распределяются ресурсы, назначаются ответственные лица, после чего происходит непосредственно сбор НД.

3. Этап регистрации готового НД (карточка НД).

Когда НД полностью сформирован происходит его размещение в хранилище НД, составление readme-файла и непосредственно регистрация, то есть внесение всей информации о нем в реестр.

4. Размещение в библиотеке НД.

На данном этапе происходит размещение ключевых параметров НД в карточке библиотеки НД и публикация ее на сайте.

5. Смена версии/утилизация.

В процессе использования НД с целью исправления ошибок или добавления новой информации, а также при создании новых НД на базе уже существующих с целью оптимизации выполнения работ и более рационального распределения происходит смена версии, которая регламентируется с помощью введения мажорных, минорных и патч-версий [16, 27].

6. Использование НД.

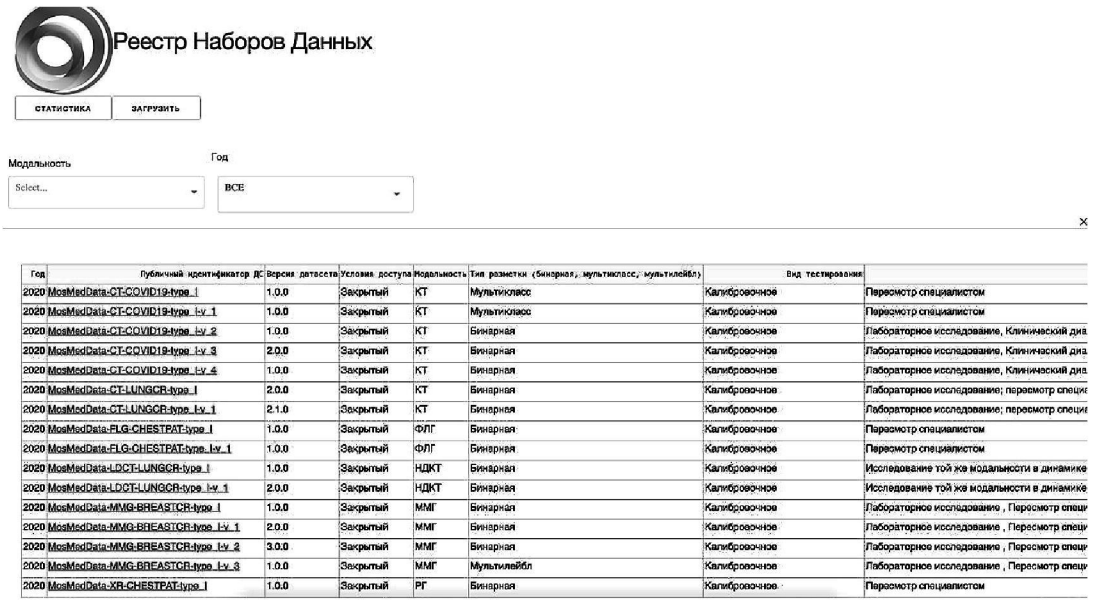
После прохождения всех этапов создания НД можно приступать к его использованию: разработке и тестированию алгоритмов машинного обучения, проведению испытаний, научных работ и т.д.

Результаты и обсуждение

По результатам анализа НД и их сопроводительной документации нами был создан «Реестр НД для медицинской диагностики» и определен, как систематизированный перечень сведений обо всех НД ГБУЗ «НПКЦ ДиТ ДЗМ», ведущийся уполномоченным сотрудником, с целью упорядочивания деятельности ГБУЗ «НПКЦ ДиТ ДЗМ» по формированию и использованию НД для машинного обучения.

На первых этапах функционирования реестра, он представлял собой таблицу в формате xls, однако такое представление базы данных неудобно в повседневной работе, поэтому был разработан графический интерфейс, более удобный и наглядный *(рис. 1)*.

В основу реестра легли этапы формирования и использования НД *(рис. 2)*. Мы предлагаем



Год	Публичный идентификатор ДС	Версия датасета	Условия доступа	Модальность	Тип разметки (бинарная, мультикласс, мультилейбл)	Вид тестирования	
2020	MosMedData-CT-COVID19-type_I	1.0.0	Закрытый	КТ	Мультикласс	Калибровочное	Пересмотр специалистом
2020	MosMedData-CT-COVID19-type_I-v_1	1.0.0	Закрытый	КТ	Мультикласс	Калибровочное	Пересмотр специалистом
2020	MosMedData-CT-COVID19-type_I-v_2	1.0.0	Закрытый	КТ	Бинарная	Калибровочное	Лабораторное исследование, Клинический диа
2020	MosMedData-CT-COVID19-type_I-v_3	2.0.0	Закрытый	КТ	Бинарная	Калибровочное	Лабораторное исследование, Клинический диа
2020	MosMedData-CT-COVID19-type_I-v_4	1.0.0	Закрытый	КТ	Бинарная	Калибровочное	Лабораторное исследование, Клинический диа
2020	MosMedData-CT-LUNGCR-type_I	2.0.0	Закрытый	КТ	Бинарная	Калибровочное	Лабораторное исследование; пересмотр специ
2020	MosMedData-CT-LUNGCR-type_I-v_1	2.1.0	Закрытый	КТ	Бинарная	Калибровочное	Лабораторное исследование; пересмотр специ
2020	MosMedData-FLG-CHESTPAT-type_I	1.0.0	Закрытый	ФЛГ	Бинарная	Калибровочное	Пересмотр специалистом
2020	MosMedData-FLG-CHESTPAT-type_I-v_1	1.0.0	Закрытый	ФЛГ	Бинарная	Калибровочное	Пересмотр специалистом
2020	MosMedData-LDCT-LUNGCR-type_I	1.0.0	Закрытый	НДКТ	Бинарная	Калибровочное	Исследование той же модальности в динамике
2020	MosMedData-LDCT-LUNGCR-type_I-v_1	2.0.0	Закрытый	НДКТ	Бинарная	Калибровочное	Исследование той же модальности в динамике
2020	MosMedData-MMG-BREASTCR-type_I	1.0.0	Закрытый	ММГ	Бинарная	Калибровочное	Лабораторное исследование , Пересмотр специ
2020	MosMedData-MMG-BREASTCR-type_I-v_1	2.0.0	Закрытый	ММГ	Бинарная	Калибровочное	Лабораторное исследование , Пересмотр специ
2020	MosMedData-MMG-BREASTCR-type_I-v_2	3.0.0	Закрытый	ММГ	Бинарная	Калибровочное	Лабораторное исследование , Пересмотр специ
2020	MosMedData-MMG-BREASTCR-type_I-v_3	1.0.0	Закрытый	ММГ	Мультилейбл	Калибровочное	Лабораторное исследование , Пересмотр специ
2020	MosMedData-XR-CHESTPAT-type_I	1.0.0	Закрытый	РГ	Бинарная	Калибровочное	Пересмотр специалистом

Рис. 1. **Фрагмент реестра наборов данных**

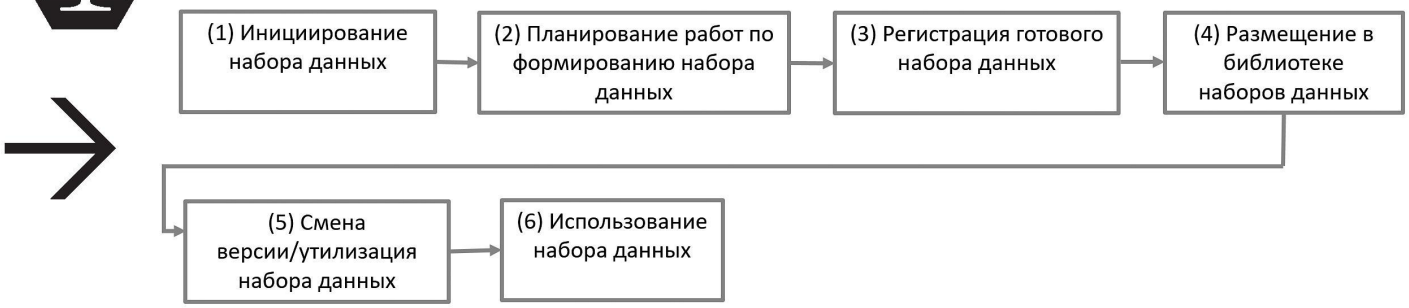


Рис. 2. **Этапы формирования и использования НД**

следующую структуру (подробный перечень полей представлен в приложении): 7 параметров на этапе инициирования, 8 – на этапе планирования, 70 – карточка НД, 1 – смена версии, 14 – на этапе использования, всего – 100 параметров.

1. Этап Инициации.

На данном этапе в реестр вносится следующая информация: рабочее название планируемого НД, ответственные лица, сроки выполнения работ, ссылки на базовые диагностические требования, а также тип НД, т.е. цель, с которой он создается. По результатам анализа основных направлений работ по применению НД мы выделили следующие типы назначений:

I – Проведение тестирований с целью оценки функционала (функциональное тестирование) и оценки метрик диагностической точности, настройки алгоритмов (калибровочное тестирование).

II – «Селф-тест технический» – проведение самостоятельной проверки разработчиками способности алгоритма машинного обучения обрабатывать исследования с диагностических устройств разных производителей и моделей [27].

III – «Селф-тест диагностический» – проведение самостоятельной проверки корректности клинической интерпретации исследований алгоритмом.

IV – Выполнение клинических испытаний – оценка безопасности и эффективности медицинского изделия [29].

V – Выполнение технических испытаний – оценка соответствия характеристик алгоритма требованиям нормативно-правовой, технической и эксплуатационной документации [29].

VI – Проведение разметки текстовых протоколов с помощью программ автоматизированного анализа текстов (например, MedLabel [30]).

VII – Проведение научных исследований.

VIII – Разработка ИИ: обучение и дообучение алгоритмов ИИ.

Также с целью удобства обращения с файлами порядковый номер, присваиваемый НД на этапе инициирования, используется в наименовании файлов технического задания.

2. Этап планирования работ по формированию НД.

На данном этапе фиксируются планируемые сроки проведения работ по созданию НД (начало и окончание подготовки НД), актуальный статус (например, подготовка технического задания или сбор НД) и дата его смены, ответственные за разметку данных и за НД, ссылка на техническое задание, а также поле для комментария в свободной форме.

Заполнение реестра на этапах инициирования и планирования позволяет не упустить информацию о готовящихся НД и отслеживать сроки проведения работ по их созданию, по необходимости обращаться к ответственным лицам для решения возникающих вопросов, контролировать процесс планирования и сбора НД, а также формировать справки о ходе работы для отчетности.

3. Этап регистрации готового НД (карточка НД).

Это самый большой раздел реестра, который содержит структурированное описание НД, и из него формируется readme-файл и карточка НД в библиотеке НД.

Идентификация НД

В реестре, помимо рабочего названия, присваиваемого НД на этапе инициирования, имеется еще 2 наименования: идентификатор НД и публичное наименование НД. Идентификатор уникален и позволяет однозначно установить, о каком НД идет речь в каждом конкретном случае. Публичное наименование формируется на русском языке и необходимо для публичного представления НД. Основываясь на примерах формирования наименований медицинских данных [31] и данных в других областях мы разработали следующие правила

формирования названия НД *(рис. 3, 4)*. Единый стандарт идентификации позволяет только исходя из названия понять то, какой организацией НД был подготовлен, исследования какой модальности (вид медицинского исследования) и нозологии в него включены, а также его назначение (закодировано в порядковом номере типа), что способствует упорядочиванию, наглядному представлению данных и удобному обращению с ними.

Также в разделе идентификации НД указывается порядковый номер НД, год его создания, информация о версиях и условия доступа.

Клинические параметры.

В этом разделе карточки указывается следующая информация:

- модальность;
- анатомическая область исследования на русском и английском языке согласно федеральным справочникам [32, 33];
- идентификаторы и коды справочников (идентификатор Федерального справочника инструментальных исследований [33], код услуги ЕРИС, код RadLex [22], код LOINC [21], код SNOMED [20]);
- название целевой патологии;
- внутренний код (формируется хранителем реестра на английском языке исходя из названия целевой патологии и необходим для формирования идентификатора НД);
- коды МКБ-10 направляющего диагноза и целевой патологии;
- уникальный идентификатор нозологии согласно алфавитному указателю к МКБ-10 [34];
- критерии включения/невключения пациентов в исследование.

Популяционные параметры

В этом разделе вносится информация о возрасте и поле пациентов, география и период сбора, эпидемиологическая обстановка, а также претестовая вероятность [35] и источник данных (фантомные, синтетические или пациенты).

Назначение (область применения)

Указываются задачи создания НД, сценарий применения моделей ИИ, созданных на основе

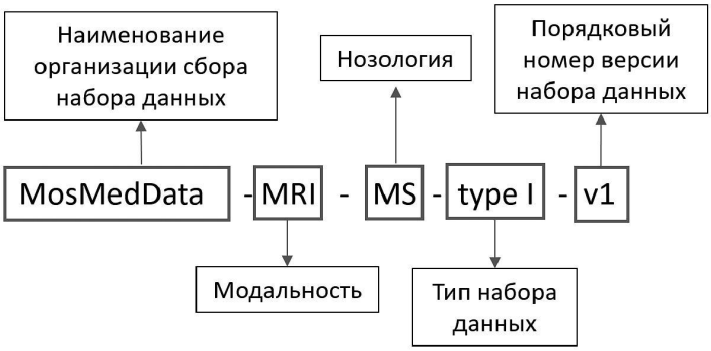


Рис. 3. **Правила формирования идентификатора набора данных**

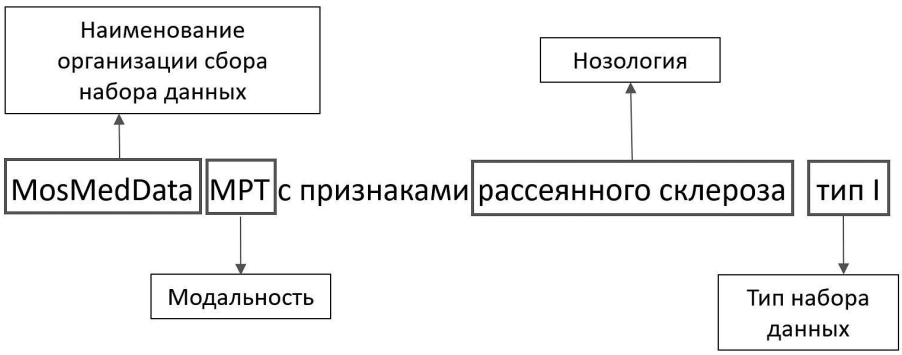


Рис. 4. **Правила формирования публичного наименования (полного) набора данных**

НД, виды и варианты тестирования, для которых создавался НД.

Параметры разметки [27]

- Способы предразметки.
- Уровень разметки (пациент, исследование, серия, изображение).
- Тип разметки (бинарная, мультикласс, мультилейбл).
- Количество лейблов.
- Характер разметки (бинарная, категориальная, регрессионная).
- Уровень детализации лейблов (исследование/серия/изображение; находка (локализация); Находка (сегментация)).
- Названия лейблов.
- Количество классов.
- Названия классов.
- Количество по классам.
- Класс разметки.
- Метод валидации/ верификации.
- Количество специалистов.
- Опыт (стаж работы) специалистов.
- Временной промежуток между входными данными и данными верификации.
- Используемая при верификации информация из МК пациента.
- Критерии отнесения к классам.

Для понимания структуры НД следует более подробно остановиться на следующих терминах в контексте НД для медицинской диагностики:

1. Предразметка (предварительная разметка) – это способ предварительного отбора информации в НД. Данные могут отбираться вручную или с использованием какого-либо алгоритма, например, с использованием анализатора текстовых протоколов Medlabel [30].

2. Класс – это множество всех объектов с заданным значением метки. В медицинских данных чаще всего встречаются классы «наличие патологии» / «отсутствие патологии» в случае бинарной классификации или одно из подразделений в классификации патологического состояния, например шкалы степени тяжести заболевания (KT-COVID, BIRADS, ASPECTS и т.д.).

3. Лейбл (от англ. label – ярлык, этикетка) – название патологического (или нормального) состояния, которое подвергается классификации, например, в НД компьютерной томографии грудной клетки может быть 2 лейбла: «признаки рака легких» и «признаки коронавирусной инфекции».

Исходя из количества лейблов и классов определяется тип разметки (бинарная, мультикласс, мультилейбл).

Для стандартизации и унификации методов валидации и верификации на основе собственного опыта [27] создания НД и рекомендаций управления по санитарному надзору за качеством пищевых продуктов и медикаментов (Food and Drug Administration, FDA [36]) нами был разработан справочник методов валидации и верификации *(таблица 1)*:

Технические параметры

- Критерии включения/ невключения в НД.
- Протоколы и условия сбора данных.
- Единичная запись НД: объект разметки и результат разметки.
- Форматы записи НД: объект разметки и результат разметки.
- Количество записей НД.
- Общий объем НД (Гб).
- Количество уникальных источников (диагностических устройств).
- Перечень моделей и производителей.
- Степень анонимизации.
- Комментарий к НД.

Карточка НД позволяет наглядно и структурированно продемонстрировать всю необходимую

Таблица 1

Методы валидации/верификации

Метод валидации/ верификации	*Пример*
Исследование другой модальности	Для верификации патологии на рентгенологическом исследовании: компьютерная томография той же области
Лабораторное исследование	Гистологическая верификация рака предстательной железы
Исследование той же модальности в динамике	Для верификации перелома позвонков на компьютерной томографии: признаки перелома позвонков в заключении компьютерной томографии в динамике
Клинический диагноз	Установленный диагноз U07.1 по данным медицинской карты
Пересмотр специалистом	Пересмотр разметчиком и экспертом
Согласно тексту описания исследования	Поиск ключевых слов в тексте описания исследования

информацию НД, что также дает возможность пользователям (публичным или внутренним) при необходимости обратиться к реестру с целью поиска подходящего НД. Это также позволяет в случае наличия такого НД избежать дополнительных трудозатрат для создания нового. В полях карточки заключена исчерпывающая информация для понимания возможности использования уже готового НД или формирования нового набора на базе уже существующего.

4. Смена версии/утилизация.

Если НД был сформирован на базе другого или направлен на утилизацию, эта информация вносится в реестр в поле «Смена версии». Это позволяет отслеживать ход работы над НД, внесение изменений, смену ответственных лиц, позволяет в будущем избежать возможных ошибок. Процесс изменения версионности также отражен в графе «Версия НД» карточки с помощью введения мажорных, минорных и патч-версий [16].

5. Использование НД.

Когда НД полностью сформирован необходимо отслеживать информацию о его использовании. Как правило, в медицинской организации ведутся отдельные журналы и документы для фиксирования информации о тестированиях на различных платформах, о научном сотрудничестве, публикациях, доступе для разработчиков и другое. Вся эта информация хранится разрозненно и при необходимости ее получения для отчетов или других целей требуется координация деятельности многих сотрудников. Во избежание этого ссылки на такие журналы и другая информация по использованию (публикации, сотрудничества) также фиксируются в реестре.

Также в разделе «Использование» указывается информация о регистрации НД в ФИПС (Федеральный Институт Промышленной Собственности): необходимость и статус регистрации. Для обеспечения централизованного хранения и оперативного доступа к подробным, максимально структурированным данным указывается ссылка на readme-файл, формат хранения файла и ссылки на место хранения НД с разметкой и без.

На сегодняшний день реестр успешно функционирует в рамках научно-практических задач ГБУЗ «НПКЦ ДиТ ДЗМ» (в т.ч. [12]): внесена информация о 334 НД. Разработка интерфейса продолжается, предполагается введение уровней доступа, автоматизация обработки данных, формирование справок и аналитических отчетов, содержание полей также пересматривается и актуализируется под текущие задачи.

Для широкого использования НД, а не только в пределах одной организации, существуют библиотеки НД (например, https://mosmed.ai/datasets/ [37]). Библиотеки представляют собой систематизированное собрание НД, доступных для использования. НД представлены в виде каталога карточек, в котором вся информация стандартизирована и отображена в наглядной форме, что позволяет исследователям, разработчикам и компаниям быстро оценить применимость конкретного НД для их задач. Использование реестра позволяет оперативно выгрузить информацию в карточки каталога, при этом избежать ошибок и не потерять данные. Библиотеки НД позволяют разработчикам решений на основе машинного обучения получать актуальную информацию, что позволяет повышать качество конечного продукта.

Особое внимание при создании реестра следует уделить вопросам его качества. Среди факторов, влияющих на качество отмечают следующие: управление, качество данных, конфиденциальность и безопасность [38].

Под управлением подразумевается организационная основа реестра, обеспечение ресурсов (финансовых, людских и технических) для его функционирования [38]. Применение надлежащих принципов управления должно обеспечивать четкий и легкий способ сбора данных на всех этапах. С этой целью нами были разработана сопутствующая документация: «Правила создания, изменения и использования наборов данных и их учета», регламентирующие порядок взаимодействия персонала при создании, изменении и использовании НД, и «Руководство по заполнению реестра наборов данных», подробная инструкция по заполнению всех полей реестра со всеми необходимыми ссылками и справочниками.

Качество реестра определяется не только результативностью использования данных и самого реестра (количество и качество публикаций, тестирований, разработок, запросов и т.д.)[38], но и «точностью» и «полнотой» [39]. Точность – степень, в которой зарегистрированные данные соответствуют истине, а полнота – степень, в которой все необходимые данные, которые могли бы быть зарегистрированы, действительно были зарегистрированы [39]. Надлежащее качество реестра обеспечивается процессами управления, описанными выше. Следует также отметить, что сам реестр позволяет проводить оценку выполненной работы в рамках задач по созданию и использованию НД, например, в виде статистических отчетов,

сформированных на запросах по необходимым параметрам.

Вопросы конфиденциальности и безопасности данных связаны не только с защитой интеллектуальной собственности (НД является интеллектуальной собственностью), но и с этическими вопросами и неприкосновенностью частной жизни. Непосредственно в самом реестре персональных данных пациентов нет, а вопросы анонимизации НД отражены в поле «степень анонимизации» реестра. Все меры информационной безопасности регламентируются действующим законодательством [41].

Следует отметить, что в мировой литературе в последнее время поднимается вопрос стандартизации и учета НД. Например, в работе Gebru T. авторы также в основу структуры «таблиц данных для НД» («datasheets for datasets») ставят жизненный цикл НД, однако эти разработки более общие, предназначены для НД в различных сферах и заполняются в более свободной форме [40]. Мы же создаем реестр непосредственно для данных медицинской инструментальной диагностики, предлагаем строгую унификацию и классификацию, минимальное количество полей, подразумевающих ответ в свободной форме, что способствует более четкой стандартизации и, как следствие, удобству работы с данными.

Описанные процессы создания и менеджмента реестра наборов медицинских данных имеют более широкое практическое применение и не ограничиваются только лишь инструментальной диагностикой. Рассмотренные в нашей работе принципы могут быть расширены на другие направления применения ИИ в здравоохранении: лабораторная диагностика, системы поддержки принятия врачебных решений для терапевтических, хирургических и прочих направлений медицинской деятельности.

Реестр НД может использоваться как в рамках одной медицинской организации, так и централизовано для групп медицинских организаций. Централизация процесса ведения реестра позволит стандартизировать и систематизировать подготовку медицинских НД в разных медицинских учреждениях. При этом, конечно же, требуется детальная проработка организационных и юридических аспектов данного процесса, а также перевод архитектуры реестра по аналогии с системами управления баз данных. В целевой модели должна быть единая база данных с распределенными по уровням доступа пользователями, представленная в наглядной форме, например, при помощи графического интерфейса, BI-системы (bisness intelligence). В соответствии с принципами эффективного менеджмента необходима организация процессов управления, технической поддержки, контроля доступа, корректности заполнения и актуализации имеющихся данных.

В перспективе, возможно включение реестра в регистрацию и аттестацию разработчиков, как составляющей системы менеджмента качества. Так, например, разработчикам необходимо уметь работать с библиотеками НД, а также участвовать в подготовке технического задания на разработку определенных НД.

Выводы

Широкое применение технологий ИИ в сфере здравоохранения требует формирования большого количества качественных эталонных НД для решения задач разработки, обучения, тестирования и оценки качества алгоритмов машинного обучения, что в свою очередь диктует необходимость разработки инструментов, позволяющих организовать удобную работу с этими НД. В данной работе впервые предлагается оригинальная технология структуризации и систематизации медицинских НД, в основу которой положены уникальная терминология и принципы классификации информации, разработанные ГБУЗ «НПКЦ ДиТ ДЗМ».

Создание реестра позволяет:

1. Стандартизировать информацию о НД для машинного обучения.
2. Обеспечить централизацию хранения, удобный и быстрый доступ ко всей информации о НД.
3. Формировать отчеты и справки для регуляции и повышения эффективности деятельности медицинской или научной организации по подготовке НД.
4. Обеспечить прозрачность, надежность и воспроизводимость разработок в сфере искусственного интеллекта.
5. Оперативно сформировать унифицированные и наглядные карточки НД в библиотеках, позволяя пользователю принимать решение применимости НД для его задач, минуя изучение сопроводительной документации.

Публикация структурированных, качественных данных обеспечивает их широкое использование и способствует развитию и внедрению ИИ.

СПИСОК ИСТОЧНИКОВ

1. *Ranschaert E.R., Morozov S., Algra P.R.* Artificial intelligence in medical imaging: Opportunities, applications and risks. Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks // Published online January 29, 2019:1–373. DOI: 10.1007/978-3-319-94878-2

2. Группа Центра компетенций Национальной технологической инициативы на базе МФТИ по направлению «Искусственный интеллект». Искусственный интеллект. Индекс 2021 года. Аналитический сборник № 10. 2022. Доступно по: https://aireport.ru/ai_index_russia-2021. Ссылка активна на 12.09.2022.

3. Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации». 2019. Доступно по: http://www.kremlin.ru/acts/bank/44731/page/1.Ссылка активна на 12.09.2022.

4. *Соболева С.Ю., Голиков В.В., Тажибов А.А.* Информационные технологии в здравоохранении: особенности отраслевого применения. E-Management. 2021; 4(2):37–43. doi.org/10.26425/2658-3445-2021-4-2-37-43.

5. *Морозов С.П., Кузьмина Е.С., Ветшева Н.Н. и др.* Московский скрининг: скрининг рака легкого с помощью низкодозовой компьютерной томографии // Проблемы социальной гигиены, здравоохранения и истории медицины. – 2019. – Т. 27. – С. 630–636. DOI: 10.32687/0869-866X-2019-27-si1-630-636

6. *Белолипецкая А.Е., Головина Т.А., Полянин А.В.* Цифровая трансформация сферы здравоохранения: компетентностный подход // Проблемы социальной гигиены, здравоохранения и истории медицины. – 2020. – Т. 28. – С. 694–700. DOI: 10.32687/0869-866X-2020-28-s1-694-700

7. *Dash S., Shakyawar S., Sharma M., Kaushik S.* Big data in healthcare: management, analysis and future prospects // Journal of Big Data. 2019;6(1):1–25. DOI: 10.1186/S40537-019-0217-0/FIGURES/6

8. *Griffith B., Kadom N., Straus C.* Radiology Education in the 21st Century: Threats and Opportunities // Journal of the American College of Radiology. 2019;16(10):1482–1487. DOI: 10.1016/J.JACR.2019.04.003

9. *Attia Z., Harmon D., Behr E., Friedman P.* Application of artificial intelligence to the electrocardiogram // Eur Heart J. 2021;42(46):4717–4730. DOI: 10.1093/EURHEARTJ/EHAB649.

10. *Fürbass F., Kural M., Gritsch G., Hartmann M., Kluge T., Beniczky S.* An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: Validation against the diagnostic gold standard // Clin Neurophysiol. 2020;131(6):1174–1179. DOI: 10.1016/J.CLINPH.2020.02.032

11. *Dey P.* Artificial neural network in diagnostic cytology // Cytojournal. 2022;19:27. DOI: 10.25259/CYTOJOURNAL_33_2021

12. *Морозов С.П., Владзимирский А.В., Ледихова Н.В., Андрейченко А.Е., Арзамасов К.М., Баланюк Э.А., Гомболевский В.А., Ермолаев С.О., Живоденко В.С., Идрисов И.М., Кирпичев Ю.С., Логунова Т.А., Нуждина В.А., Омелянская О.В., Раковчен В.Г., Слепушкина А.В.* Московский эксперимент по применению компьютерного зрения в лучевой диагностике: вовлеченность врачей-рентгенологов // Врач и информационные технологии. – 2020. – № 4. – С. 14–23. DOI: 10.37690/1811-0193-2020-4-14-23

13. *Jin C., Chen W., Cao Y. et al.* Development and evaluation of an artificial intelligence system for COVID-19 diagnosis // Nature Communications. 2020;11(1). DOI: 10.1038/S41467-020-18685-1

14. *Logunova T., Andreychenko A.E., Klyashtorny V., Arzamasov K.M., Vladzymyrskyy A., Morozov S.* Artificial intelligence services' impact on radiologist's performance in the context of the COVID-19 pandemic // Insights Imaging. 2021, 12 (Suppl 2): 216. DOI: 10.1186/s13244-021-01014-5

15. *Морозов С.П., Гаврилов А.В., Архипов И.В. и др.* Влияние технологий искусственного интеллекта на длительность описаний результатов компьютерной томографии пациентов с COVID-19 в стационарном звене здравоохранения // Профилактическая медицина. – 2022. – Т. 25. – № 1. – С. 14–20. DOI: 10.17116/profmed20222501114

16. *Павлов Н.А., Андрейченко А.Е., Владзимирский А.В., Ревазян А.А., Кирпичев Ю.С., Морозов С.П.* Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике // Digital Diagnostics. 2021;2(1):49–66. DOI: 10.17816/DD60635

17. *Willemink M.J., Koszek W.A., Hardell C. et al.* Preparing medical imaging data for machine learning // Radiology. 2020;295(1):4–15. DOI: 10.1148/RADIOL.2020192224

18. *Newman-Griffis D., Divita G., Desmet B., Zirikly A., Rosé C.P., Fosler-Lussier E.* Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets // J Am Med Inform Assoc. 2021;28(3):516. DOI: 10.1093/JAMIA/OCAA269

19. UMLS Metathesaurus – CPT (CPT – Current Procedural Terminology) – Metadata. Accessed: 19.09.2022: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT/metadata.html

20. SNOMED International [Electronic resource]. URL: https://www.snomed.org/. Accessed: 12.09.2022.

21. Logical Observation Identifiers Names and Codes [Electronic resource]. Accessed: 19.09.2022: https://loinc.org/.

22. RadLex Term Browser [Electronic resource]. Accessed: 19.10.2022: http://radlex.org/.

23. *Filice R.W., Kahn C.E.* Biomedical Ontologies to Guide AI Development in Radiology // Journal of Digital Imaging. 2021;34(6):1331–1341. DOI: 10.1007/S10278-021-00527-1/FIGURES/4.

24. *Wilkinson M.D., Dumontier M., Aalbersberg Ij.J. et al.* The FAIR Guiding Principles for scientific data management and stewardship // Scientific. Data 2016 3:1. 2016;3(1):1–9. DOI:10.1038/sdata.2016.18

25. *Wang J.W., Williams M.* Registries, Databases and Repositories for Developing Artificial Intelligence in Cancer Care // Clin Oncol (R Coll Radiol). 2022;34(2): e97-e103. DOI: 10.1016/J.CLON.2021.11.040

26. Приказ Министерства здравоохранения Российской Федерации от 13.10.2017 г. № 804н «Об утверждении номенклатуры медицинских услуг». http://publication.pravo.gov.ru/Document/View/0001201711080036. Ссылка активна на 12.09.2022.
27. *Морозов С.П., Владзимирский А.В., Андрейченко А.Е., Ахмад Е.С., Блохин И.А., Гомболевский В.А., Зинченко В.В., Кульберг Н.С., Новик В.П., Павлов Н.А.* Регламент подготовки наборов данных с описанием подходов к формированию репрезентативной выборки данных. Часть 1. 2021. Доступно по: https://tele-med.ai/media/documents/MP__Регламент_подготовки_наборов_данных_Ч.1_Препринт.pdf. Ссылка активна на 10.10.2022.
28. *Mayer-Schonberger V., Ramge T.* Reinventing Capitalism in the Age of Big Data. 2018.
29. *Морозов С.П., Владзимирский А.В., Кляшторный В.Г. [и др.].* Клинические испытания программного обеспечения на основе интеллектуальных технологий (лучевая диагностика) / Москва: Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы, 2019. Доступно по: https://tele-med.ai/media/documents/klinicheskie_ispytaniya_po_12022020.pdf. Ссылка активна на 12.09.2022.
30. *Морозов С.П., Андрейченко А.Е., Кирпичев Ю.С. [и др.]* MedLabel – автоматизированный анализ медицинских протоколов. Свидетельство о государственной регистрации программы для ЭВМ № 2020664321 Российская Федерация.11.11.2020 / ГБУЗ «НПКЦ ДиТ ДЗМ». Доступно по: https://elibrary.ru/item.asp?id=44244882 Ссылка активна на 12.09.2022
31. *Khaled R., Helal M., Alfarghaly O. et al.* Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Scientific Data 20229:1. 2022;9(1):1–10. DOI:10.1038/s41597-022-01238-0
32. Справочник Анатомические локализации. Доступно по: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477/version/4.5. Ссылка активна на 10.10.2022.
33. Федеральный справочник инструментальных диагностических исследований. Доступно по: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1471/version/2.15. Ссылка активна на 10.10.2022.
34. Алфавитный указатель к Международной статистической классификации болезней и проблем, связанных со здоровьем (10-й пересмотр, Том 3). Доступно по: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1489/version/2.22. Ссылка активна на 10.10.2022.
35. *Thomas G., Kenny L., Baker P., Tuytten R.* A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests. Diagnostic and Prognostic Research 20171:1. 2017;1(1):1–9. DOI:10.1186/S41512-017-0017-Y
36. Food and Drug Administration (FDA). Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in-Premarket Notification (510(k)) Submissions Guidance for Industry and FDA Staff Preface Public Comment // Published online 2012. Accessed: May 19, 2022. https://www.fda.gov/media/77642/download
37. https://mosmed.ai/datasets/ [интернет] Наборы данных. / ГБУЗ «НПКЦ ДиТ ДЗМ» Доступно по: https://mosmed.ai/datasets/. Ссылка активна на 09.11.2022.
38. *Zaletel M., Kralj M., Magajne M.* Methodological guidelines and recommendations for efficient and rational governance of patient registries. Institute PDLN, 2015. Published online 2015. Accessed: May 21, 2022. https://ec.europa.eu/health/system/files/2016–11/patient_registries_guidelines_en_0.pdf
39. *Danielle G.T. Arts, Nicolette F. de Keizer, Gert-Jan Scheffer,* Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study and Generic Framework // Journal of the American Medical Informatics Association, November 2002; 9(6): 600–611. https://doi.org/10.1197/jamia.M1087
40. *Gebru T., Morgenstern J., Vecchione B. et al.* Datasheets for Datasets. Documentation to facilitate communication between dataset creators and consumers // Communications of the ACM. 2021;64(12). DOI:10.1145/3458723
41. Федеральный закон от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации».

Источник финансирования: Данная статья подготовлена авторским коллективом в рамках научно-практического проекта в сфере медицины «Эталонные наборы данных для устойчивого развития технологий искусственного интеллекта в медицинской диагностике с целью минимизации долгосрочных последствий пандемии коронавирусной инфекции для здоровья населения города Москвы».

Конфликт интересов: авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Благодарности: авторский коллектив выражает благодарность младшему научному сотруднику сектора разработки систем внедрения медицинских интеллектуальных технологий ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ» Кирпичеву Ю.С. за реализацию графического интерфейса реестра.

Участие авторов

Васильев Ю.А. – руководство проектом.
Бобровская Т.М. – оформление рукописи, сбор литературных данных, сбор и систематизация данных, разработка сопроводительной документации, разработка справочников.
Четвериков С.Ф. – сбор и систематизация данных, разработка сопроводительной документации, разработка справочников и наименований.
Арзамасов К.М. – разработка концепции реестра, сбор литературных данных.
Владзимирский А.В. – разработка концепции реестра.
Омелянская О.В. – организация и управление бизнес-процессом подготовки наборов данных.
Андрейченко А.Е. – сбор литературных данных, разработка концепции реестра, сбор и систематизация данных, разработка справочников и наименований.
Павлов Н.А. – сбор и систематизация данных, разработка концепции реестра.
Анищенко Л.Н. – сбор литературных данных, разработка концепции реестра, разработка сопроводительной документации, разработка справочников и наименований.

Все авторы внесли значимый вклад в проведение исследования и подготовку статьи, прочли и одобрили финальную версию статьи перед публикацией.

ORIGINAL PAPER

MEDICAL DATASETS FOR MACHINE LEARNING: FUNDAMENTAL PRINCIPLES OF STANDARTIZATION AND SYSTEMATIZATION

Y.A. Vasilev[1], T.M. Bobrovskaya[2]✉, K.M. Arzamasov[3], S.F. Chetverikov[4], A.V. Vladzymyrskyy[5], O.V. Omelyanskaya[6], A.E. Andreychenko[7], N.A. Pavlov[8], L.N. Anishchenko[9]

[1, 2, 3, 4, 5, 6, 7, 8, 9] State Budget-Funded Health Care Institution of the City of Moscow «Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department», Moscow, Russia

✉ *Corresponding author: Bobrovskaya T.M.*

ABSTRACT

Background: Active implementation of artificial intelligence technologies in the healthcare in recent years promotes increasing amount of medical data for the development of machine learning models, including radiology and instrumental diagnostics data. To solve various problems of digital medical technologies, new datasets are being created through machine learning algorithms, therefore, the problems of their systematization and standardization, storage, access, rational and safe use become actual.

Aim: development of an approach to systematization and standardization of information about datasets to represent, store, apply and optimize the use of datasets and ensure the safety and transparency of the development and testing of medical devices using artificial intelligence.

Materials and methods: analysis of own and international experience in the creation and use of medical datasets, medical reference books searching and analysis, registry structure development and justification, scientific publications search with the keywords "datasets", "registry of medical data", placed in the databases of the RSCI, Scopus, Web of Science.

Results. The register of medical instrumental diagnostics datasets structure has been developed in accordance with stages of datasets lifecycle: 7 parameters at the initiation stage, 8 – at the planning stage, 70 – dataset card, 1 – version change, 14 – at the use stage, total – 100 parameters. We propose datasets classification according to the purpose of their creation, a classification of data verification methods, as well as the principles of forming names for standardization and datasets presentation clarity. In addition, the main features of the organization of maintaining this registry are highlighted: management, data quality, confidentiality and security.

Conclusions. For the first time, an original technology of medical datasets for instrumental diagnostics structuring and systematization is proposed. It is based on the developed terminology and principles of information classification. This makes it possible to standardize the structure of information about datasets for machine learning, and ensures the storage centralization. It also allows to get quick access to all information about the dataset, and ensure transparency, reliability and reproducibility of artificial intelligence developments. Creating a registry makes it possible to quickly form visual data libraries. This allows a wide range of researchers, developers and companies to choose data sets for their tasks. This approach ensures their widespread use, resource optimization and contributes to the rapid development and implementation of artificial intelligence.

Keywords: *dataset, artificial intelligence, machine learning, registries, libraries.*

For citation: *Vasilev Y.A., Bobrovskaya T.M., Arzamasov K.M., Chetverikov S.F., Vladzymyrskyy A.V., Omelyanskaya O.V., Andreychenko A.E., Pavlov N.A., Anishchenko L.N. Medical datasets for machine learning: fundamental principles of standartization and systematization // Manager Zdravoohranenia. 2023; 4: 28–41. DOI: 10.21045/1811-0185-2023-4-28-41.*

Conflict of interest: *The authors declare that there is no conflict of interest.*

REFERENCES

1. *Ranschaert E.R., Morozov S., Algra P.R.* Artificial intelligence in medical imaging: Opportunities, applications and risks. Artificial Intelligence in Medical Imaging: Opportunities, Applications and Risks // Published online January 29, 2019:1–373. DOI:10.1007/978-3-319-94878-2

2. Competence Center of the National Technology Initiative based on MIPT in the direction of "Artificial Intelligence". Artificial Intelligence. AI Index Russia 2021. Analytical collection № 10.2022 (In Russ). Accessed: 12.09.2022: https://aireport.ru/ai_index_russia-2021.

3. Ukaz Prezidenta Rossijskoj Federacii ot 10.10.2019 g. "O razvitii iskusstvennogo intellekta v Rossijskoj Federacii" № 490. (In Russ). Accessed: 12.07.2022: http://www.kremlin.ru/acts/bank/44731/page/1.

4. *Soboleva S.U., Golikov V.V., Tazhibov A.A.* Information technologies in healthcare: features of sectoral implementing. E-Management. 2021;4(2):37–43. (In Russ.). https://doi.org/10.26425/2658-3445-2021-4-2-37-43

5. *Morozov S.P., Kuzmina E.S., Vetsheva N.N. et al.* Moscow Screening: Lung Cancer Screening With Low-Dose Computed Tomography // Problems of Social Hygiene, Public Health and History of Medicine. – 2019. – Vol. 27. – P. 630–636. DOI: 10.32687/0869-866X-2019-27-si1-630-636

6. *Belolipetskaya A.E., Golovina T.A., Polyanin A.V.* Digital transformation of healthcare: a competency-based approach // Problems of Social Hygiene, Public Health and History of Medicine. – 2020. – Vol. 28. – P. 694–700. DOI: 10.32687/0869-866X-2020-28-s1-694-700

7. *Dash S., Shakyawar S., Sharma M., Kaushik S.* Big data in healthcare: management, analysis and future prospects // Journal of Big Data. 2019;6(1):1–25. DOI: 10.1186/S40537-019-0217-0/FIGURES/6

8. *Griffith B., Kadom N., Straus C.* Radiology Education in the 21st Century: Threats and Opportunities // Journal of the American College of Radiology. 2019;16(10):1482–1487. DOI:10.1016/J.JACR.2019.04.003

9. *Attia Z., Harmon D., Behr E., Friedman P.* Application of artificial intelligence to the electrocardiogram // Eur Heart J. 2021;42(46):4717–4730. DOI: 10.1093/EURHEARTJ/EHAB649.

10. *Fürbass F., Kural M., Gritsch G., Hartmann M., Kluge T., Beniczky S.* An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: Validation against the diagnostic gold standard // Clin Neurophysiol. 2020;131(6):1174–1179. DOI: 10.1016/J.CLINPH.2020.02.032

11. *Dey P.* Artificial neural network in diagnostic cytology // Cytojournal. 2022;19:27. DOI: 10.25259/CYTOJOURNAL_33_2021

12. *Morozov S., Vladzymyrskyy A., Ledikhova N. et al.* Moscow experiment on computer vision in radiology: involvement and participation of radiologists // Vrach i informacionnye tehnologii. 2020;(4):14–23. (In Russ.) DOI: 10.37690/1811-0193-2020-4-14-23

13. *Jin C., Chen W., Cao Y. et al.* Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nature Communications. 2020;11(1). DOI:10.1038/S41467-020-18685-1

14. *Logunova T., Andreychenko A.E., Klyashtorny V., Arzamasov K.M., Vladzymyrskyy A., Morozov S.* Artificial intelligence services' impact on radiologist's performance in the context of the COVID-19 pandemic // Insights Imaging. 2021, 12 (Suppl 2): 216. DOI: 10.1186/s13244-021-01014-5

15. *Morozov S.P., Gavrilov A.V., Arkhipov I.V. et al.* Effect of artificial intelligence technologies on the CT scan interpreting time in COVID-19 patients in inpatient setting // Profilakticheskaya Meditsina. 2022;25(1):14–20. (In Russ.) DOI: 10.17116/profmed20222501114

16. *Pavlov N.A., Andreychenko A.E., Vladzymyrskyy A.V., Revazyan A.A., Kirpichev Y.S., Morozov S.P.* Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics // Digital Diagnostics. 2021;2(1):49–66. (In Russ.) DOI: 10.17816/DD60635

17. *Willemink M.J., Koszek W.A., Hardell C. et al.* Preparing medical imaging data for machine learning // Radiology. 2020;295(1):4–15. DOI:10.1148/RADIOL.2020192224

18. *Newman-Griffis D., Divita G., Desmet B., Zirikly A., Rosé C.P., Fosler-Lussier E.* Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets // J Am Med Inform Assoc. 2021;28(3):516. DOI: 10.1093/JAMIA/OCAA269

19. UMLS Metathesaurus – CPT (CPT – Current Procedural Terminology) – Metadata. Accessed: 19.09.2022: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT/metadata.html

20. SNOMED International [Electronic resource]. URL: https://www.snomed.org/. Accessed: 12.09.2022.

21. Logical Observation Identifiers Names and Codes [Electronic resource]. Accessed: 19.09.2022: https://loinc.org/.

22. RadLex Term Browser [Electronic resource]. Accessed: 19.10.2022: http://radlex.org/.

23. *Filice R.W., Kahn C.E.* Biomedical Ontologies to Guide AI Development in Radiology // Journal of Digital Imaging. 2021;34(6):1331–1341. DOI: 10.1007/S10278-021-00527-1/FIGURES/4.

24. *Wilkinson M.D., Dumontier M., Aalbersberg I.J. et al.* The FAIR Guiding Principles for scientific data management and stewardship // Scientific. Data 2016 3:1. 2016;3(1):1–9. DOI:10.1038/sdata.2016.18

25. *Wang J.W., Williams M.* Registries, Databases and Repositories for Developing Artificial Intelligence in Cancer Care // Clin Oncol (R Coll Radiol). 2022;34(2): e97-e103. DOI: 10.1016/J.CLON.2021.11.040

26. Prikaz Ministerstva zdravoohranenija Rossijskoj Federacii ot 13.10.2017 № 804n «Ob utverzhdenii nomenklatury medicinskih uslug» (In Russ.). Accessed: 12.09.2022 http://publication.pravo.gov.ru/Document/View/0001201711080036

27. *Morozov S.P. et al.* Reglament podgotovki naborov dannyh s opisaniem podhodov k formirovaniju reprezentativnoj vyborki dannyh, Chast' 1, 2021, (In Russ.) Accessed: 12.09.2022: https://tele-med.ai/media/documents/MP__Регламент_подготовки_наборов_данных_Ч.1_Препринт.pdf

28. *Mayer-Schonberger V., Ramge T.* Reinventing Capitalism in the Age of Big Data. 2018.

29. *Morozov S.P., Vladzymyrskyy A.V., Klyashtornyy V.G. et al.* Clinical acceptance of software based on artificial intelligence technologies (radiology). 2019. (In Russ.) Accessed: 12.09.2022: https://tele-med.ai/media/documents/klinicheskie_ispytaniya_po_12022020.pdf

30. *Morozov S.P. et al.* MedLabel – avtomatizirovannyj analiz medicinskih protokolov. Svidetel'stvo o gosudarstvennoj registracii programmy dlja JeVM № 2020664321 Rossijskaja Federacija.11.11.2020. Accessed: 12.09.2022: https://elibrary.ru/item.asp?id=44244882

31. *Khaled R., Helal M., Alfarghaly O. et al.* Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Scientific Data 20229:1. 2022;9(1):1–10. DOI: 10.1038/s41597-022-01238-0
32. Spravochnik Anatomicheskie lokalizacii. (In Russ.). Accessed: 12.09.2022: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477/version/4.5
33. Federal'nyj spravochnik instrumental'nyh diagnosticheskih issledovanij. (In Russ.) Accessed: 12.09.2022: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1471/version/2.15
34. Alfavitnyj ukazatel' k Mezhdunarodnoj statisticheskoj klassifikacii boleznej i problem, svjazannyh so zdorov'em (10-j peresmotr, Tom 3) (In Russ.). Accessed: 12.09.2022: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1489/version/2.22
35. *Thomas G., Kenny L., Baker P., Tuytten R.* A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests. Diagnostic and Prognostic Research 20171:1. 2017;1(1):1–9. DOI: 10.1186/S41512-017-0017-Y
36. Food and Drug Administration (FDA). Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in-Premarket Notification (510(k)) Submissions Guidance for Industry and FDA Staff Preface Public Comment. Published online 2012. Accessed: May 19, 2022. https://www.fda.gov/media/77642/download
37. https://mosmed.ai/datasets. Accessed: 09.11.2022.
38. *Zaletel M., Kralj M., Magajne M.,* Methodological guidelines and recommendations for efficient and rational governance of patient registries. Institute PDLN, 2015. Published online 2015. Accessed: May 21, 2022. https://ec.europa.eu/health/system/files/2016–11/patient_registries_guidelines_en_0.pdf
39. *Danielle G.T.* Arts, *Nicolette F. de Keizer, Gert-Jan Scheffer.* Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study and Generic Framework, Journal of the American Medical Informatics Association, November 2002; 9(6): 600–611, https://doi.org/10.1197/jamia.M1087
40. *Gebru T., Morgenstern J., Vecchione B. et al.* Datasheets for Datasets. Documentation to facilitate communication between dataset creators and consumers // Communications of the ACM. 2021;64(12). DOI:10.1145/3458723
41. Federal'nyj zakon ot 27 iyulya 2006 g. № 149-FZ "Ob informacii, informacionnyh tekhnologiyah i o zashchite informacii" (In Russ).

ИНФОРМАЦИЯ ОБ АВТОРАХ / ABOUT THE AUTHORS

Васильев Юрий Александрович *– директор ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Yuriy A. Vasilev *– Director of the Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 4458–5608, ORCID: 0000-0002-0208-5218, e-mail: VasilevYA1@zdrav.mos.ru

Бобровская Татьяна Михайловна *– м.н.с. сектора разработки систем внедрения медицинских интеллектуальных технологий отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Tatiana M. Bobrovskaya *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia;*
eLibrary SPIN: 3400–8575; ORCID: 0000-0002-2746-7554; e-mail: BobrovskayaTM@zdrav.mos.ru

Арзамасов Кирилл Михайлович *– к.м.н., руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Kirill M. Arzamasov *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 3160–8062; ORCID: 0000-0001-7786-0349; e-mail: ArzamasovKM@zdrav.mos.ru

Четвериков Сергей Федорович *– к.т.н., начальник сектора разработки систем внедрения медицинских интеллектуальных технологий отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Sergey F. Chetverikov *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 3815–8870; ORCID: 0000-0002-3097-8881; e-mail: ChetverikovSF@zdrav.mos.ru

Владзимирский Антон Вячеславович *– д.м.н., заместитель директора по научной работе ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Anton V. Vladzymyrskyy *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 3602–7120; ORCID: 0000-0002-2990-7736; e-mail: VladzimirskijAV@zdrav.mos.ru

Омелянская Ольга Васильевна *– руководитель по управлению подразделениями Дирекции Наука ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Olga V. Omelyanskaya *– Head of Department Management of the Directorate of Science of GBUZ "NPCC diagnostics and Telemedicine DZM", Moscow, Russia.*
eLibrary SPIN: 8948–6152; ORCID: 0000-0002-0245-4431; e-mail: o.omelyanskaya@npcmr.ru

Андрейченко Анна Евгеньевна *– к.ф.-м.н., ведущий научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Anna E. Andreychenko *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 6625–4186; ORCID: 0000-0001-6359-0763; e-mail: a.andreychenko@npcmr.ru

Павлов Николай Александрович *– м.н.с. отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Nikolay A. Pavlov *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 9960–4160; ORCID: 0000-0002-4309-1868; e-mail: nickvolvap@gmail.com

Анищенко Леся Николаевна *– к.т.н, с.н.с. отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», г. Москва, Россия.*
Lesya N. Anishchenko *– Center for Diagnostics and Telemedicine of the Moscow Health Department, Moscow, Russia.*
eLibrary SPIN: 2991–2001; ORCID: 0000-0002-2057-0452; e-mail: anishchenko@rslab.ru