


**ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ**

СОГЛАСОВАНО

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


Ю. А. Васильев
«11» июля 2024 г.

РЕКОМЕНДОВАНО

Экспертным советом по науке
Департамента здравоохранения
города Москвы № 11



«11» июля 2024 г.

**ПОДГОТОВКА НАБОРОВ ДАННЫХ,
ОБОГАЩЕННЫХ КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ**

Методические рекомендации № 41

УДК 004.89+614.2

ББК 32.813

П 44

Серия «Лучшие практики лучевой и инструментальной диагностики»
Основана в 2017 году

Организация-разработчик:

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

Авторы-составители:

Васильев Ю. А. – канд. мед. наук, главный внештатный специалист по лучевой и инструментальной диагностике ДЗМ, директор ГБУЗ «НПКЦ ДиТ ДЗМ»

Казаринова В. Е. – техник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Бобровская Т. М. - младший научный сотрудник отдела инновационных технологий ГБУЗ «НПКЦ ДиТ ДЗМ»

Никитин Н. Ю. – канд. физ.-мат. наук, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Памова А. П. – канд. мед. наук, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Арзамасов К. М. – канд. мед. наук, руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

П 44 Подготовка наборов данных, обогащенных клинической информацией: методические рекомендации / сост. Ю. А. Васильев, В. Е. Казаринова, Т. М. Бобровская [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. ___. – М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2024. – 39 с.

Рецензенты:

Синицын Валентин Евгеньевич – д-р мед. наук, профессор, заведующий кафедрой лучевой диагностики и терапии факультета фундаментальной медицины МГУ им. М.В. Ломоносова, заведующий отделом лучевой диагностики МНОЦ МГУ им. М.В. Ломоносова

Буренчев Дмитрий Владимирович – д-р мед. наук, заведующий отделением рентгенодиагностических и радиоизотопных методов исследования ГБУЗ «ГКБ им. Е.К. Ерамишанцева»

Методические рекомендации предназначены для организаторов здравоохранения, медицинских работников и инженерного персонала, задействованного в подготовке наборов медицинских данных для проведения научных исследований, а также для обучения и тестировании программного обеспечения на основе технологий искусственного интеллекта в составе систем поддержки принятия врачебных решений.

В издании излагаются практические рекомендации по всем этапам подготовки наборов данных, содержащих медицинские изображения, а также дополнительные клинические данные.

Данные методические рекомендации разработаны в ходе выполнения научно-исследовательской работы «Разработка платформы подготовки наборов данных лучевых диагностических исследований»

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

ISSN

© Департамент здравоохранения города Москвы, 2024

© ГБУЗ «НПКЦ ДиТ ДЗМ», 2024

© Коллектив авторов, 2024

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ	5
ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ	6
ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ	7
ВВЕДЕНИЕ	8
1. ЭТАП ИНИЦИИРОВАНИЯ	10
1.1. ПОСТАНОВКА ЦЕЛИ.....	10
2. ЭТАП ПЛАНИРОВАНИЯ	11
2.1. ТЕХНИЧЕСКОЕ ЗАДАНИЕ.....	11
2.2. ОЦЕНКА ОБЪЕМА ДАННЫХ, НЕОБХОДИМОГО ДЛЯ СОЗДАНИЯ НАБОРА ДАННЫХ, ОБОГАЩЕННЫХ КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ ВЫБОРКИ.....	12
2.2.1. ОЦЕНКА ОБЪЕМОВ ВЫБОРКИ В ЭКСПЕРИМЕНТАЛЬНЫХ И ОБСЕРВАЦИОННЫХ ИССЛЕДОВАНИЯХ	13
2.2.2. КОРРЕКТИРОВКА ЧИСЛА ИССЛЕДОВАНИЙ ПРИ УБЫТИИ ДАННЫХ.....	15
2.3. ОБЗОР ЛИТЕРАТУРЫ	16
2.3.1. ФОРМУЛИРОВКА ЦЕЛИ И НАЗНАЧЕНИЯ НД.....	16
2.3.2. ФОРМУЛИРОВКА КЛЮЧЕВЫХ СЛОВ ДЛЯ ПОИСКА	16
2.3.3. ОПРЕДЕЛЕНИЕ НАИБОЛЕЕ ПОДХОДЯЩИХ БАЗ ДАННЫХ ДЛЯ ПОИСКА ЛИТЕРАТУРЫ	21
2.3.4. ОБОСНОВАНИЕ ВЫБОРА КЛИНИЧЕСКИХ ПАРАМЕТРОВ НА ОСНОВЕ ОБЗОРА ЛИТЕРАТУРЫ.....	23
3. ЭТАП ФОРМИРОВАНИЯ	24
3.1. СБОР ДАННЫХ	24
3.2. ОБРАБОТКА ДАННЫХ	28
3.3. README-ФАЙЛ	32
4. ПРИМЕР ПОДГОТОВКИ НАБОРА ДАННЫХ, ОБОГАЩЕННОГО КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ.....	33
ЗАКЛЮЧЕНИЕ	35
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	36

НОРМАТИВНЫЕ ССЫЛКИ

В настоящем документе использованы ссылки на следующие нормативные документы (стандарты):

1. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации».
2. Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации».
3. ГОСТ 33707-2016. Информационные технологии. Словарь.
4. ГОСТ Р ИСО 21549-3-2017. Информатизация здоровья. Структура данных на пластиковой карте пациента. Часть 3. Основные клинические данные.
5. 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.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем документе применены следующие термины с соответствующими определениями:

Выборка – часть генеральной совокупности элементов, которая охватывается экспериментом.

Искусственный интеллект (ИИ) – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе то, в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

Клинические данные – данные, которые могут включать информацию о состоянии здоровья пациента и событиях медицинской помощи; описание и оценку работником здравоохранения характера событий медицинской помощи; сведения о планируемых, назначенных или выполненных действиях, связанных с оказанием медицинской помощи.

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

Проспективное исследование – исследование, в котором группа наблюдения, сформированная в настоящее время, прослеживается в будущем.

Ретроспективное исследование – исследование, которое опирается на информацию о событиях, имевших место в прошлом.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем документе применены следующие обозначения и сокращения:

БДТ	– базовые диагностические требования
БФТ	– базовые функциональные требования
ГБУЗ «НПКЦ ДиТ ДЗМ»	– Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»
ДЗМ	– Департамент здравоохранения города Москвы
ЕМИАС	– Единая медицинская информационно-аналитическая система
ЕРИС	– Единый радиологический информационный сервис
ЗНО	– злокачественное новообразование
ИИ	– искусственный интеллект
ИМТ	– индекс массы тела
МИС	– медицинские информационные системы
ММГ	– маммография
НД	– набор данных
НТБ	– Научно-техническая библиотека
НЭБ	– Научная электронная библиотека
ПО	– программное обеспечение
РГБ	– Российская государственная библиотека
РИНЦ	– Российский индекс научного цитирования
ТЗ	– техническое задание
ТИИ	– технологии искусственного интеллекта
Ф.И.О.	– фамилия, имя, отчество
ЭМК	– электронная медицинская карта
DICOM	– англ. Digital Imaging and Communications in Medicine (медицинский отраслевой стандарт создания, хранения, передачи и визуализации цифровых медицинских изображений и документов обследованных пациентов)
MeSH	– англ. Medical Subject Headings (всеобъемлющий контролируемый словарь, предназначенный для индексации журнальных статей и книг по наукам о жизни)
NaN	– Not a Number

ВВЕДЕНИЕ

В настоящее время в нашей стране происходит активная цифровизация здравоохранения согласно указу Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» [1]. Внедрение цифровых технологий в клиническую практику сопровождается высоким темпом роста объема медицинских данных, хранящихся в медицинских информационных системах (МИС). Клиническая информация в МИС представлена в неструктурированном виде, что делает ее непригодной для обучения и тестирования программного обеспечения (ПО) на основе технологий искусственного интеллекта (ТИИ). Поэтому одной из основных задач в области разработки, тестирования и дальнейшего внедрения ПО на основе ТИИ является подготовка качественных наборов данных (НД). Эталонные наборы данных должны иметь структуру и характеристики, необходимые для возможности применения методов машинного обучения, а также должны соответствовать поставленной задаче как с точки зрения компьютерных наук, так и с точки зрения медицины [2].

С теоретическими основами и практическими методами создания наборов данных в лучевой диагностике можно ознакомиться в учебном пособии Ю. А. Васильева, К. М. Арзамасова, А. В. Владзимирского и др. «Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта» [3].

Настоящее издание содержит рекомендации по обогащению НД клинической информацией. Клинические данные могут включать:

- информацию о состоянии здоровья пациента и событиях медицинской помощи;
- описание и оценку работником здравоохранения характера событий медицинской помощи;
- сведения о планируемых, назначенных или выполненных действиях, связанных с оказанием медицинской помощи [4].

Подготовка НД должна происходить в соответствии со ст. 13 ч. 3 Федерального закона от 21.11.2011 № 323-ФЗ [5].

Обогащение НД клинической информацией требуется как для научных целей, так и для обучения и тестирования ПО на основе ТИИ. В клинической практике при принятии решений относительно диагноза и планирования лечения врачи используют информацию из анамнеза жизни и анамнеза заболевания, жалоб пациента и результатов осмотра, данных лабораторной и инструментальной диагностики. Для установления точного диагноза с использованием ПО на основе ТИИ необходимо полное моделирование клинической практики врача. Это включает учет таких факторов, как пол, возраст пациента и другие медицинские параметры, которые могут повлиять на окончательный диагноз. Опыт зарубежных исследователей показывает, что использование клинической информации при обучении ПО на основе ТИИ

повышает значения метрик диагностической точности [6, 7, 8] и улучшает прогностические модели [9].

Конечная цель настоящих методических рекомендаций заключается в том, чтобы помочь научному и медицинскому сообществу в процессе обогащения существующих наборов данных дополнительной клинической информацией.

1. ЭТАП ИНИЦИИРОВАНИЯ

1.1. Постановка цели

Согласно учебному пособию «Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта» [3], **первым этапом при создании любого набора данных является постановка цели**, согласно которой будет вестись дальнейший процесс планирования и сборки набора данных. Как отмечалось выше, обогащение клинической информацией, как правило, требуется для задач обучения и тестирования ПО на основе ТИИ, для научных исследований и с целью обучения медицинского персонала. Однако в случае обогащения возможен выбор двух стратегий создания НД, от которых будут зависеть дальнейшие процессы и трудозатраты:

1. Обогащение уже имеющегося набора данных.
2. Создание нового набора данных, обогащенного клинической информацией.

У каждой стратегии имеются свои преимущества и недостатки. В первом случае процесс упрощается за счет уже выполненной части работ по сбору, фильтрации и структуризации данных, однако в этом случае важную роль играет качество исходного набора данных. Стоит отметить, что реализация такой стратегии в крупных исследовательских центрах, имеющих возможность создавать большое количество НД, оптимизируется в случае использования инструментов управления, таких как реестр НД [10]. Благодаря регламентированному учету всех процессов создания и использования НД, а также структурированной информации о них, контроль качества и принятие решений о выборе данных, подвергаемых обогащению, существенно упрощается. Тем не менее еще одним недостатком данной стратегии является ретроспективный характер сбора данных. Это может привести к наличию пропусков в данных (из-за их отсутствия), несоответствию критериям включения/исключения, невозможности балансировки НД по клиническим параметрам, а также отсутствию исследований с требуемыми клиническими параметрами.

В случае создания «с нуля», сбор данных происходит проспективно, что позволяет оптимально сбалансировать НД (если это требуется) и обеспечить полноту заполнения данных. Однако при этом по сравнению с предыдущим методом обогащения уже готового НД повышаются трудозатраты для осуществления всех основных этапов создания НД в соответствии с методологией [3]. Также возможна комбинация этих двух стратегий: обогащение уже готового НД и дополнение его новыми исследованиями в соответствии с техническим заданием (ТЗ).

2. ЭТАП ПЛАНИРОВАНИЯ

2.1. Техническое задание

Одним из важнейших этапов при создании наборов данных является планирование. Именно от того, насколько тщательно будут регламентированы все процессы создания, начиная от определения финансирования и сроков выполнения проекта и заканчивая обоснованием параметров и количества исследований итогового НД, будет зависеть качество выполненных работ. В случае разработки и тестирования ПО на основе ТИИ при планировании следует учитывать базовые диагностические и базовые функциональные требования (БДТ и БФТ), но основным документом при создании любого набора данных является техническое задание.

Подробно процесс подготовки ТЗ для создания НД рассмотрен в учебном пособии [3], здесь представлена краткая версия для стратегии обогащения уже имеющегося НД. Перечень рекомендуемых полей представлен в приложении А. Ниже описаны ключевые параметры, необходимые для описания процесса обогащения клинической информацией.

Безусловно, в ТЗ должна быть обозначена **цель** создания набора данных (см. пункт

1.1.цели), исходя из которой в дальнейшем будут определяться параметры, которыми будет обогащен набор данных (пункт 2.3.литературы). Также необходимо указать **наименование процедуры**, т. е. основное исследование, данные которого будут обогащаться клинической информацией. В случае, если исходный набор данных содержал признаки **целевой патологии(ий)**, это также необходимо указать. Исходя из этих данных, рекомендуется формировать и название самого набора данных, например: «Набор данных компьютерной томографии с признаками внутричерепных кровоизлияний, обогащенный клиническими данными». В указанном примере за основу брался набор данных, содержащий исследования компьютерной томографии с признаками внутричерепных кровоизлияний, и обогащался клиническими параметрами, например с целью повышения диагностической точности ПО на основе ТИИ (разработка ИИ).

Следующее, что необходимо указать в ТЗ, – это непосредственно **клинические параметры**, которыми НД будет обогащаться. Как уже отмечалось выше, в случае ретроспективного сбора данных невозможно настраивать **баланс классов** каждого **клинического параметра** (в случае классификации) в НД, а в случае проспективного – необходимо указать это условие, например, представленность курящих и некурящих пациентов в наборе данных должна составлять 50/50. В силу особенностей некоторых параметров они могут быть представлены не во всех исследованиях (например, при сборе анамнеза не уточнялось, курит пациент или нет), поэтому существует риск того, что в итоговый набор данных будет включено много пропущенных значений, что будет негативно сказываться на решении задачи, для которой НД создается. С целью предотвращения таких негативных

последствий в ТЗ необходимо **указать допустимое значение пропусков** для каждого параметра (например, 5%), при превышении которого этот параметр исключается из НД.

На этапе подбора клинических данных необходимо учитывать не только вид самих параметров, но и **временной интервал между основным исследованием и обогащаемыми данными**. Например, положительный результат ПЦР-теста на COVID-19, проведенный через 2 месяца после компьютерной томографии (основное исследование), никак не повлияет на интерпретацию найденных на КТ признаков.

Так как обогащение клинической информацией в большинстве случаев представляет из себя работу с разнородными данными медицинской карты, автоматизация процесса на сегодняшний день затруднена, поэтому раздел, посвященный выгрузке данных из МИС, преобразуется в поле **источник данных**, в котором указывается непосредственно основной источник, откуда получают данные для обогащения (например, электронная медицинская карта) и/или конкретное исследование (например, общий анализ крови или консультация врача-терапевта). Остальные разделы ТЗ остаются без изменений [3], однако необходимо пересмотреть **критерии включения и исключения** исследований и пациентов в набор данных, так как при обогащении эти параметры могут измениться. Также раздел ТЗ, посвященный разметке, относится к работе с клиническими параметрами, соответственно **разметчик** в данном случае – это специалист, который осуществляет извлечение клинической информации.

В ТЗ в обязательном порядке фиксируется целевой объем набора данных. Этот объем может быть известен заранее, например, в случае предварительной разработки дизайна обучения или тестирования ПО на основе ТИИ. В случае, если предварительно такая работа проведена не была, ее необходимо провести как один из этапов при подготовке ТЗ. Самый простой вариант – воспользоваться фиксированным значением объема выборки в зависимости от цели работы [11], номограммой Альтмана [12] или способом расчета, учитывающим ряд факторов, приведенным в следующем разделе.

2.2. Оценка объема данных, необходимого для создания набора данных, обогащенных клинической информацией выборки

При формировании набора данных, обогащенного клинической информацией, может возникнуть потребность в проведении оценки объема данных, необходимых для дальнейшего проведения научных исследований с использованием ПО на основе ТИИ.

Для проведения оценки необходимого количества данных следует руководствоваться следующими принципами:

1. Определить, о каком типе данных идет речь.
2. Оценить объем данных, содержащийся в исходном наборе данных.
3. Определить, сколько классов содержится в исходном наборе данных.

4. Оценить возможный уровень потерь в данных при сборе клинической информации.
5. Оценить допустимую ошибку в данных.
6. Оценить доверительную вероятность при анализе данных.

При проведении медицинских исследований наиболее часто доверительная вероятность принимается равной 95 %, а значение ошибки – 5 %. При бинарной классификации количество классов составляет два – «норма» и «патология». Уровень потерь данных при сборе клинической информации может варьироваться от 50 до 5 %, в случае редко встречаемых патологий возможны большие потери клинической информации. Допустимое количество пропусков в данных определяется требованиями ТЗ.

2.2.1. Оценка объемов выборки в экспериментальных и наблюдательных исследованиях

При проведении наблюдательных или экспериментальных исследований основной задачей является сравнение двух и более групп, а также определение объемов выборки для двух и более групп. Вычисление объема выборки для двух сравниваемых групп на основе пропорций осуществляется по уравнению (1):

$$n = \frac{(t_{\alpha} * \sqrt{2 * p * q} + t_{\beta} \sqrt{p_1 * q_1 + p_2 * q_2})^2}{(p_1 - p_2)} \quad (1)$$

где n – количество исследований, необходимое в каждой группе; p_1 – доля в группе № 1; p_2 – доля в группе № 2; q_1 и q_2 – обратные p_1 и p_2 величины; p и q – соответственно:

$$p = \frac{p_1 + p_2}{2} \quad (2)$$

$$q = 1 - p \quad (3)$$

Доля признака в группе вычисляется по уравнению (4):

$$p_1 = \frac{N_1}{N} \quad (4)$$

где N_1 – количество исследований в первой группе; N – общее количество исследований в классе. Для второй группы вычисление доли признака производится по тому же уравнению, с той разницей, что вместо N_1 подставляется N_2 .

t_{α} – коэффициент Стьюдента, соответствующий уровню статистической значимости (ошибке первого рода); t_{β} – коэффициент Стьюдента, соответствующий значению мощности (для простоты вычислений принять мощность равную 95%).

В таблице 1 представлены коэффициенты Стьюдента для различных доверительных вероятностей.

Таблица 1 – Доверительная вероятность и соответствующие ей коэффициенты Стьюдента

Значение коэффициента t	Доверительная вероятность $P(t)$
1	68,3%
1,96	95,0 %
2	95,5
2,58	99,0
3	99,7

Вычисление объемов выборки на основе дисперсий для двух групп проводится по уравнению (5):

$$n = \frac{2 * s^2 * (t_{\alpha} + t_{\beta})^2}{(\bar{x}_1 - \bar{x}_2)^2} \quad (5)$$

где \bar{x}_1 – среднее значение исследуемой величины в первой группе; \bar{x}_2 – среднее значение исследуемой величины во второй группе. Средние значения вычисляются по уравнению (6):

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (6)$$

где N – количество исследований в группе.

Величина s^2 – определяется по уравнению (7):

$$s^2 = \frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2} \quad (7)$$

где n_1 и n_2 – объем выборки в каждой из групп пилотного исследования; s_1 и s_2 – дисперсии в каждой из групп пилотного исследования.

Дисперсия в каждой группе вычисляется по уравнению (8):

$$S = \sqrt{\frac{1}{N - 1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (8)$$

где N – количество исследований в классе.

Вычисление объема выборки при исследованиях, содержащих много уровней одного фактора и много факторов, может быть проведено на основании средневзвешенных значений дисперсии или пропорций. Оценка общего числа необходимого для проведения исследования объема выборки для дисперсий осуществляется по уравнению (9):

$$n = \frac{\bar{s}^2 * t_{\alpha}^2 * N}{(N - 1) * e^2 + t_{\alpha}^2 * \bar{s}^2} \quad (9)$$

где N – размер генеральной совокупности, а \bar{s}^2 вычисляется по уравнению (10):

$$\bar{s}^2 = \frac{\sum_{i=1}^k s_i^2 * n_i}{\sum_{i=1}^k n_i} \quad (10)$$

где n_i – количество исследований в пилотном эксперименте или литературных источниках для i-го класса; s_i^2 – дисперсия i-й группы исследуемого класса.

Для случая бесконечного объема выборки:

$$n = \frac{t_{\alpha}^2 * \bar{s}^2}{e^2} \quad (11)$$

Оценка общего количества исследований, необходимых для подтверждения эффекта с заданной точностью по долям каждого класса, находится по уравнению (12):

$$n = \frac{t_{\alpha}^2 * (\bar{p} * \bar{q}) * N}{(N - 1) * e^2 + t_{\alpha}^2 * (\bar{p} * \bar{q})} \quad (12)$$

где N – размер генеральной совокупности;

$$(\bar{p} * \bar{q}) = \frac{\sum_{i=1}^k p_i * q_i * n_i}{\sum_{i=1}^k n_i} \quad (13)$$

где p_i – ожидаемый процент эффекта от i-го группе класса; $q_i=1-p_i$ – процент отсутствия отклика от i-го группе класса, n_i – количество исследований в i-й группе класса.

2.2.2. Корректировка числа исследований при убытии данных

При проведении клинических исследований объем исследуемой выборки может подвергаться изменениям по причинам естественной убыли или по причине отказа от участия субъектов в исследованиях. Такие изменения могут повлиять на конечный результат исследований,

соответственно, данные изменения должны быть учтены при планировании исследования.

Расчет объемов выборки с учетом возможного изменения выполняется по уравнению (14):

$$N = \frac{n}{(1 - q)} \quad (14)$$

где n – количество исследований, полученное без учета изменений в объемах данных; q – доля отсутствующих данных [13].

2.3. Обзор литературы

Этот раздел посвящен методам поиска надежных источников литературы на медицинскую тематику с целью определения перечня дополнительных клинических параметров для обогащения НД. Будут рассмотрены доступные инструменты и стратегии, используемые для поиска научных статей, книг, баз данных и других источников, содержащих ценную информацию для клинических исследований.

2.3.1. Формулировка цели и назначения НД

Первым этапом является формулировка цели подготовки набора данных и примерные ответы на вопросы:

- 1) для чего создается конкретный набор данных;
- 2) какая патология/заболевание, *возможно*, могут быть обнаружены в этом наборе данных;
- 3) с какой клинической информацией патология/заболевание, *возможно*, могут коррелировать;

Важно: четкое понимание необходимо только относительно первого вопроса, однако понимание (2) и (3) вопросов значительно ускорит поиск.

Пример

Предположим, что исследователя интересует обогащение НД, содержащего цифровые маммографические снимки пациентов с наличием или отсутствием признаков злокачественных новообразований (ЗНО) в молочной железе.

1. НД создается для обучения ПО на основе ТИИ распознаванию типа ЗНО.
2. Рак, гемангиоэпителиома, киста и т.д.
3. Время наступления менархе, количество родов, кормление грудью и т.д.

2.3.2. Формулировка ключевых слов для поиска

Вторым этапом является формулирование основных, «ключевых», базовых слов, на основе которых будет строиться весь дальнейший поиск литературных источников и составление из них списка {ключей} на русском и на английском языках.

Важно: если вы плохо владеете английским языком, то перевод терминов с русского на английский язык советуем осуществлять в онлайн-словаре «Мультитран» (<https://www.multitran.com/>).

Пример

Начнем с очевидного. Первым {ключевым} словом является: *маммография (mamмоgraphy)*. Также можно назвать ключевыми следующие слова: *скрининг (screening)*, *рак (onco, cancer)*, *новообразование (neoplasm)*. Естественно, набор слов можно расширять.

Подбор ключевых слов в медицинских предметных рубриках (Medical Subject Headings, MeSH)

Важным моментом является то, что для поиска медицинской литературы в базах данных PubMed, Medline, библиотеке NLM или реестре ClinicalTrials необходимо подобрать MeSH-термины, обратившись к рубрикатору базы данных PubMed – MeSH: (<https://pubmed.ncbi.nlm.nih.gov>).

Пример

Переходим на сайт по ссылке выше, нажимаем ссылку «MeSH Database» (рисунок 1).



Рисунок 1 – Ссылка на переход в словарь MeSH терминов на сайте базы данных PubMed

Далее вводим в строку поиска подобранные ключевые слова, список которых был сформирован ранее, например *маммография*, после чего ищем в списке терминов близкое по значению и определению к нашим задачам слово и выписываем его (рисунок 2).

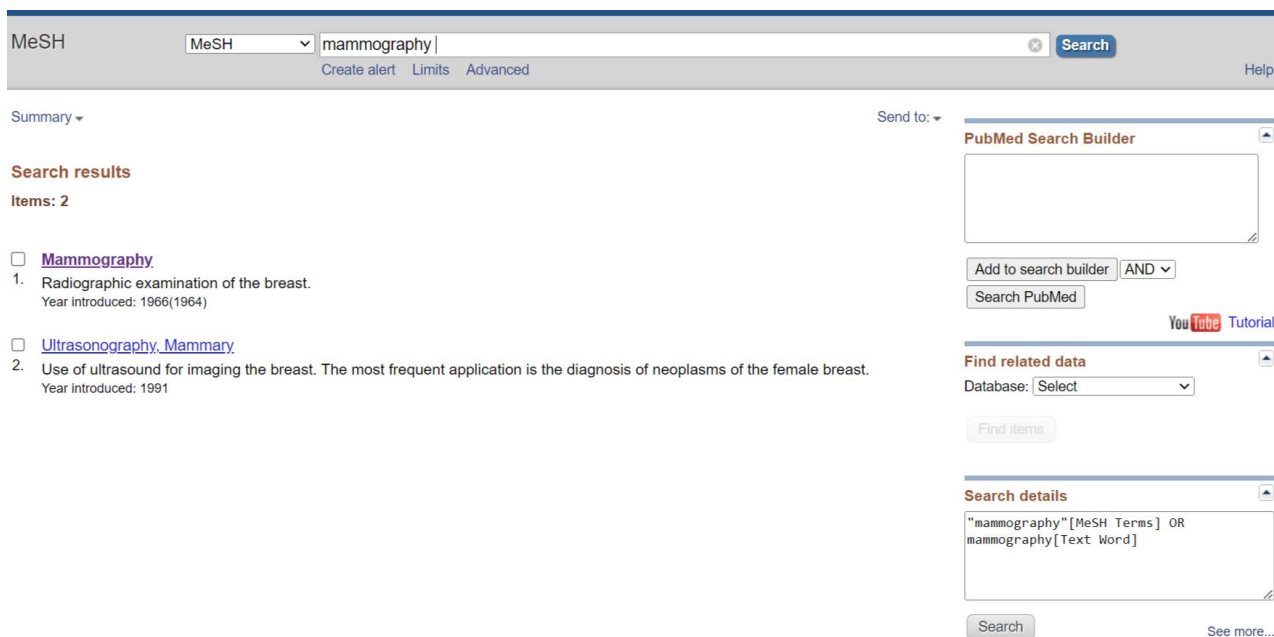


Рисунок 2 – Найденные в базе данных PubMed термины MeSH, удовлетворяющие нашим условиям

Использование MeSH-терминов

Медицинские предметные рубрики (Medical Subject Headings, MeSH) – это контролируемый словарь, используемый для каталогизации записей в базах данных PubMed, Medline, Cochrane и т.д. Эти рубрики помогают в индексации статей, обеспечивая эффективный поиск литературы. Рубрики позволяют пользователям расширять или сужать поиск литературы, сосредоточившись на конкретных вопросах рассматриваемой темы. Для более подробного знакомства с использованием MeSH-терминов необходимо воспользоваться ссылкой: <https://libguides.library.tmc.edu/PubMed/MeSH>, так как рассмотрение этого вопроса не является целью настоящих методических рекомендаций.

Пример

Приведем пример литературного поиска с использованием базы данных PubMed и MeSH-терминов. Для того чтобы сузить поиск относительно цифровой маммографии, добавим в поисковое дерево найденный ранее MeSH-термин (рисунок 3).

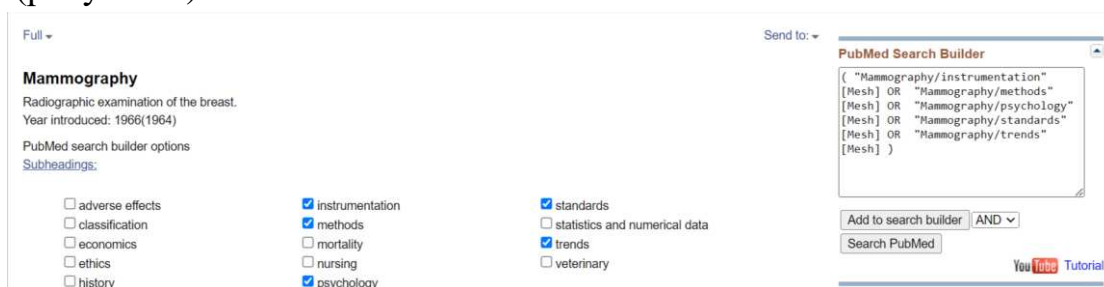


Рисунок 3 – Составление поискового запроса в базе данных PubMed с использованием MeSH-терминов

Таким образом, сузился поиск по MeSH-терминам, связанным с маммографией (отмечены галочкой на рисунке 3). Справа можно увидеть более узкий запрос на искомую тему, который сформируется автоматически после нажатия на кнопку «add to search builder». Далее можно переходить к поиску непосредственно в PubMed, нажав на кнопку «Search PubMed». На рисунке 4 ниже отображаются найденные по этому запросу публикации.

Рисунок 4 – Найденные на PubMed статьи с узким поисковым запросом по MeSH-терминам

Если есть необходимость, этот запрос можно сузить еще сильнее, воспользовавшись специальными символами (тэгами) и операторами, непосредственно встроенными в поиск PubMed, просто добавив их к уже существующему запросу через логические операторы (OR, AND, NOT) или воспользовавшись продвинутым поиском статей (Advanced).

Пример

("Mammography"[MeSH Terms] OR mammography[Title/Abstract]) AND ("Clinical Parameters"[Title/Abstract] OR "Clinical Importance"[Title/Abstract] OR "Clinical Significance"[Title/Abstract] OR "Clinical

Relevance"[Title/Abstract]) AND ("Breast Neoplasms"[MeSH Terms] OR "Breast Cancer"[Title/Abstract] OR "Breast Tumor"[Title/Abstract]) NOT ("animals"[MeSH Terms]).

Значение логических операторов и предназначение специальных символов (тэгов) PubMed представлено в таблице 2. Стоит отметить, что иногда те же самые операторы и символы можно использовать и в других базах данных.

Таблица 2 – Значение логических операторов при их использовании в PubMed у MeSH-терминов

Логические операторы и символы PubMed	Значение и использование
AND	<p>Оператор «И». При поиске статья должна соответствовать всем ключевым словам, перечисленным через AND. Т.е. она обязательно включает в себя все перечисленные слова.</p> <p>Пример: “diabet*[ti] AND obesit*[ti]”. В заголовке публикации должны быть как слова с разными окончаниями diabetes так и obesity.</p>
OR	<p>Оператор «ИЛИ». При поиске должно быть хотя бы одно из условий (ключевое слово), перечисленных через OR.</p> <p>Пример: “cancer[ti] OR oncology[ti]”. В заголовке статьи будет либо одно, либо другое, либо оба слова сразу.</p>
NOT	<p>Оператор «КРОМЕ». При поиске при добавлении ключевого слова через этот оператор – исключаются статьи с наличием данного ключевого слова.</p> <p>Пример: “diabetes[ti] NOT insulin[ti]”. В заголовке статьи будет слово диабет и не будет слова инсулин.</p>
()	<p>Скобки используются для группировки поисковых запросов. Помогают уточнить поиск, комбинируя различные логические операторы.</p> <p>Пример: “(diabetes OR obesity) AND (diet OR nutrition)”.</p>
« »	<p>Кавычки помогают для поиска точной фразы, заключенной в них.</p> <p>Пример: “diabetes”. Найдутся статьи на эту тематику. Не обязательно в заголовке.</p>

*	<p>Звездочка в конце корня слова используется для поиска вариантов слова. Результатом будет слово, которое содержит указанный корень.</p> <p>Пример: “child*[ti]”. В заголовке будут статьи с корнем слова child, такие как: child, children, childhood и т.д.</p>
[]	<p>В скобках указываются специальные символы (тэги), которые работают в PubMed. Эти скобки ставятся за искомым ключевым словом (не MeSH-термином).</p> <p>Пример: [Title] или [ti] – символ, указывающий, что это слово должно быть в заголовке статьи. [MeSH Terms] – указывает, что это ключевое слово должно быть в MeSH-терминах к этой статье. [Title/Abstract] – указывает, что ключевое слово должно быть в названии статьи или в ее описании.</p> <p>Более подробно со спецсимволами PubMed можно ознакомиться по ссылке: https://www.ncbi.nlm.nih.gov/pmc/about/userguide/</p>

Пользуясь данными инструментами, можно значительно сузить поиск литературы при использовании базы данных PubMed.

2.3.3. Определение наиболее подходящих баз данных для поиска литературы

После определения ключевых слов необходимо составить четкий план того, где именно будет происходить поиск литературы. На самом деле, выбор баз данных достаточно обширный, в каждой из них можно найти уникальную информацию, которой не будет в другой базе. Поэтому поиск литературы является очень трудоемким мероприятием и требует большого количества времени.

Обычно, для того чтобы обеспечить надлежащий и эффективный поиск, достаточно найти **хотя бы одну публикацию**, которая будет удовлетворять поставленным целям и отвечать на большую часть вопросов. Из нее возможно вычлениить надлежащую терминологию и искать публикации, содержащие аналогичную информацию. Часто именно **одна публикация становится фундаментальной для всего остального поиска**.

Ниже представлен список самых распространенных баз данных и ресурсов для поиска англо-/русскоязычных публикаций на биомедицинскую тематику (таблица 3).

Таблица 3 – Общедоступные ресурсы для поиска научной медицинской литературы

База данных или Поисковая система	Ссылка	Общая информация
Научная электронная библиотека (НЭБ)	https://www.elibrary.ru	Крупнейшая российская электронная библиотека, интегрированная с Российским индексом научного цитирования (РИНЦ). Есть доступ к полным текстам.
MedLine	https://www.nlm.nih.gov	Содержит в себе данные по биомедицинским и медицинским англоязычным журналам. Есть доступ к полным текстам.
Scopus	https://www.scopus.com	Содержит в себе журналы, книги, инструкции из разных областей, есть специализированный поиск. Есть доступ к полным текстам.
Google Scholar	https://scholar.google.com	Поисковая система для поиска академических статей, диссертаций, отчетов и т.д. Не предоставляет доступ к статьям напрямую. Хороший поиск по ключевым словам. Интегрируется с библиотеками РГБ* и НТБ** и др., если есть читательский билет – можно проводить онлайн-поиск с их использованием.
«КиберЛенинка»	https://cyberleninka.ru	Российская электронная библиотека, предоставляет доступ к полным текстам. Предоставляет услугу персональной подборки материалов.
PubMed	https://pubmed.ncbi.nlm.nih.gov	Содержит в себе данные по биомедицинским и медицинским журналам со всего мира, наиболее обновляемая база данных. Имеется специализированный поиск. Есть доступ к полным текстам.
<p>* РГБ – Российская государственная библиотека (Москва) ** НТБ – Научно-техническая библиотека (Москва)</p>		

Есть и другие способы поиска научной литературы, например через ResearchGate, однако эта платформа имеет иные задачи и цели, поэтому поиск в ней не является оптимальным. Необходимо понимать и то, что каждая база данных и поисковая система, научная библиотека – являются самостоятельными и уникальными ресурсами. Поиск в них проводится с помощью различающихся между собой инструментов. Однако для всех них одинаково правило: чем лучше подобраны ключевые слова и чем четче сформулирована цель и задачи поиска, тем быстрее, легче и эффективнее пройдет сам поиск.

Организация собранной литературы

После завершения поиска литературы и скачивания всех доступных публикаций необходимо организовать хранение найденной информации в одном месте для удобства последующего использования. Для этого можно применять бесплатное ПО, например, Mendeley или Zotero. С помощью тщательной организации библиотеки из найденных публикаций станет возможным быстрое переключение между ними, что ускоряет работу со списками литературы.

2.3.4. Обоснование выбора клинических параметров на основе обзора литературы

После завершения поиска литературы можно перейти к выбору клинических параметров, которыми будет обогащен набор данных. Рекомендуем придерживаться следующей последовательности действий для отбора клинических параметров из литературных источников:

1. Найдена публикация, которая наиболее полно отвечает на поставленные вопросы и соответствует цели литературного поиска.
2. Публикация изучена, из нее выделена нужная и общепринятая терминология, ключевые слова, возможно, найдены некоторые клинические параметры или корреляции нужных в исследовании параметров.
3. Изучен список литературы этой публикации. На него всегда необходимо обращать внимание, так как там часто можно найти близкие к искомой теме публикации.
4. Найдены публикации из списка литературы, они изучены.
5. Проведен дополнительный поиск литературы.
6. Сформирован и обоснован ряд клинических параметров, которыми нужно обогатить НД.

3. ЭТАП ФОРМИРОВАНИЯ

3.1. Сбор данных

Данный раздел посвящен сбору данных. С процессом выгрузки медицинских исследований в формате DICOM-файлов по уникальным идентификаторам можно ознакомиться в учебном пособии Ю. А. Васильева, К. М. Арзамасова, А. В. Владзимирского и др. «Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта» [3]. Приступим к сбору дополнительных клинических признаков. Источником клинических данных в ретроспективных исследованиях является электронная медицинская карта (ЭМК) пациентов. В ЭМК содержится клиническая информация, записанная врачом, например анамнез жизни и анамнез заболевания, жалобы и результаты осмотра на приеме, данные лабораторных и инструментальных методов диагностики, а также выписки из стационаров. Эту информацию разметчик должен уметь обработать и структурировать, чтобы включить ее в набор данных. Таким образом, поток информации от пациента до набора данных можно представить в виде схемы, представленной на рисунке 5.

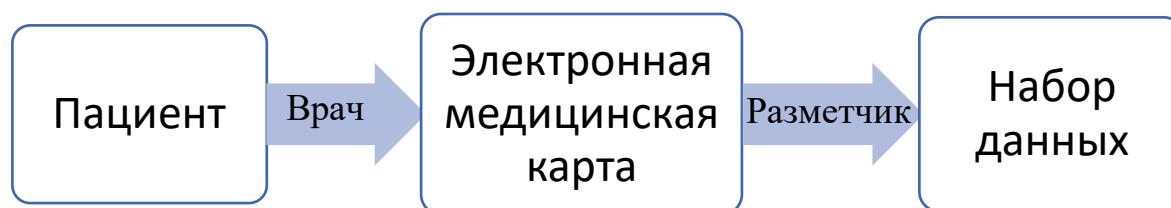


Рисунок 5 – Схема передачи информации от пациента до набора данных

Перед началом работы с ЭМК

По итогам анализа литературы таблицу разметки в Excel надо дополнить столбцами с названием медицинских параметров, которыми будет обогащен НД (таблица 4).

Таблица 4 – Пример таблицы разметки, дополненной столбцами с названиями медицинских параметров

Анонимизированный номер исследования	Наличие патологии	Пол	Возраст	ИМТ
1111111	1			
2222222	0			
3333333	1			

В случае возможности автоматизации процессов сбора данных (например, в случае лабораторных или инструментальных исследований, когда информация максимально структурирована) разметчик может использовать алгоритмы автоматизации сбора данных (например, обработки естественного языка). В остальных случаях сбор данных производится вручную.

Работа с ЭМК

Сбор клинической информации может быть осуществлен из ЭМК с помощью различных МИС. В настоящих методических рекомендациях будет рассмотрена работа с ЕМИАС, сбор данных – вручную.

Шаг 1. Поиск ЭМК пациента в МИС

В ЕМИАС поиск ЭМК пациента осуществляется по номеру полиса или Ф.И.О. и дате рождения (варианты поиска выделены красным прямоугольником на рисунке 6).

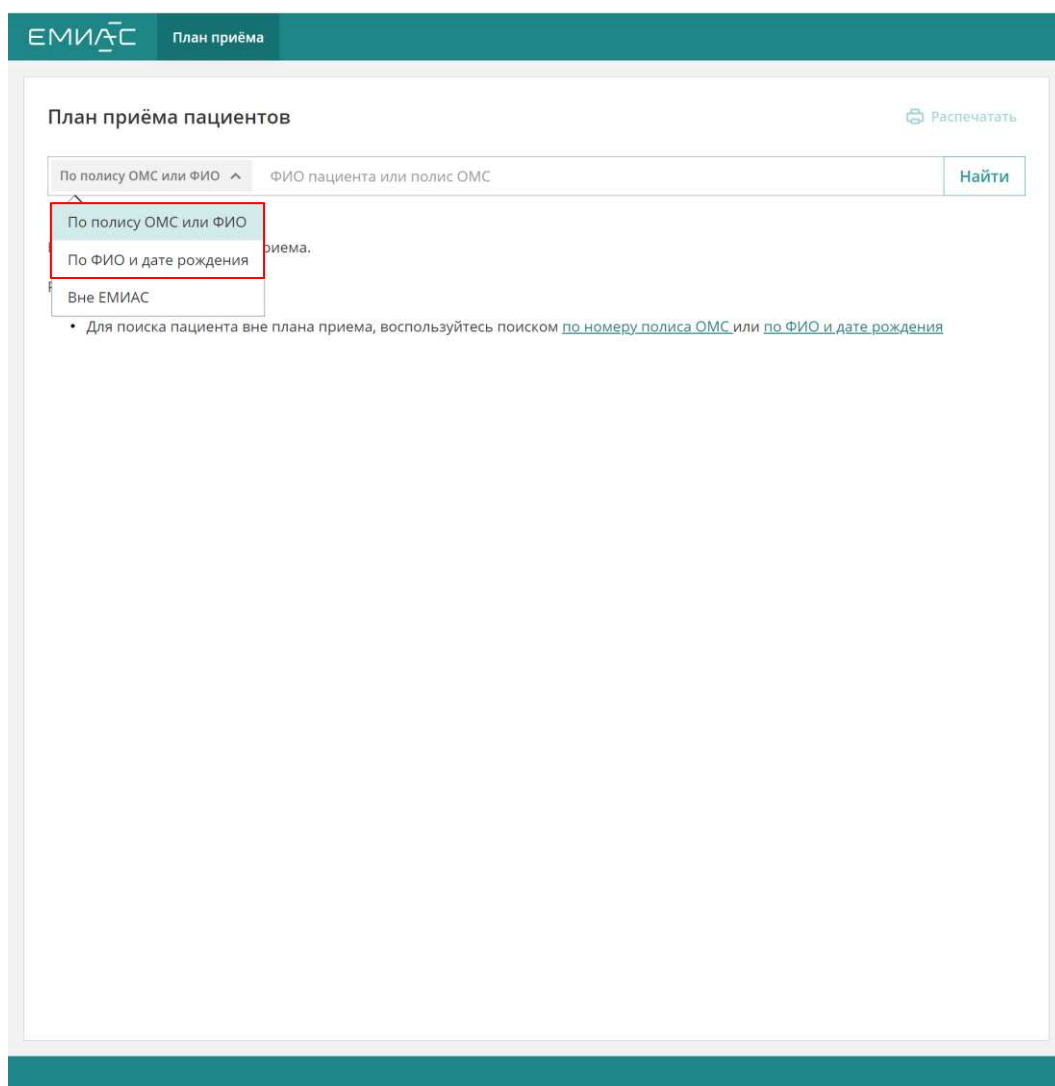


Рисунок 6 – Страница поиска ЭМК пациента

Шаг 2. Поиск клинической информации в ЭМК

В ЭМК пациента общие данные, такие как пол и возраст, можно найти в верхней части экрана (рисунок 7). В разделе «Осмотры» содержится история визитов пациента к врачам амбулаторно-поликлинического звена и выписки из стационаров, в разделе «Лабораторные исследования» – результаты проведенных лабораторных анализов. Вся информация упорядочена по датам.

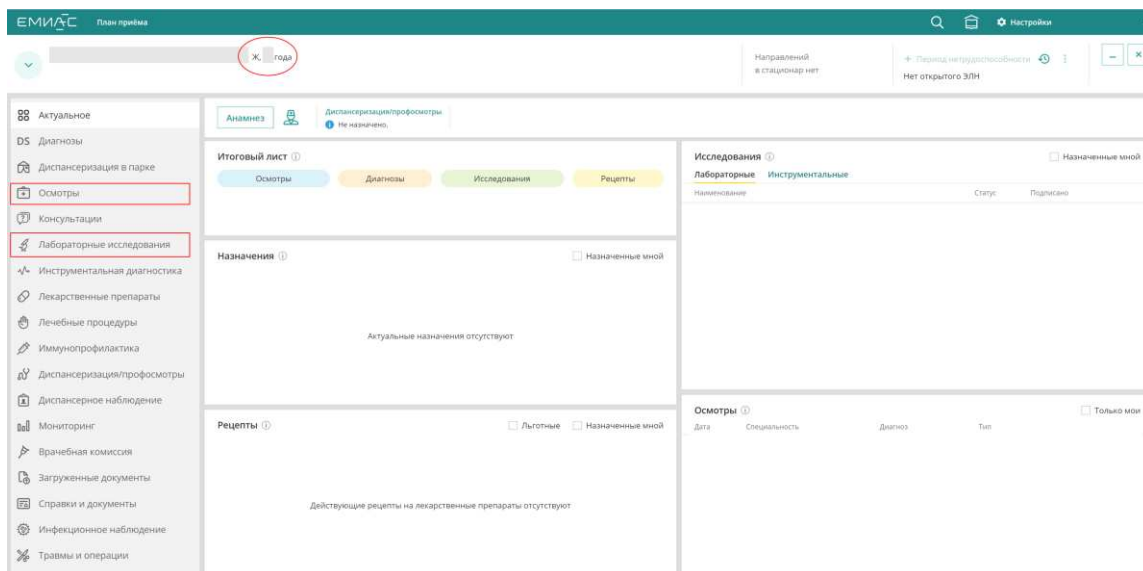


Рисунок 7 – Пример отображения ЭМК пациента: красным овалом обведены пол и возраст, красными прямоугольниками выделены разделы «Осмотры» и «Лабораторные исследования»

В зависимости от цели и задач исследования могут потребоваться данные, привязанные к моменту проведения лучевого исследования (например, симптомы, вес и т.д.). Для поиска такой клинической информации нужно перейти в соответствующий раздел: «Осмотры» или «Лабораторная диагностика» и найти ближайшую к дате проведения исследования запись врача. Например, нужно собрать информацию об индексе массы тела (ИМТ) пациента в момент проведения исследования. Для этого в разделе «Осмотры» нужно найти ближайшую (в рамках установленного в ТЗ допустимого временного интервала) к дате проведения исследования запись участкового терапевта, найти в ней ИМТ пациента и внести полученную информацию в таблицу разметки.

Таким образом, поочередно находя ЭМК пациентов в МИС и извлекая нужную клиническую информацию, разметчик заполняет таблицу. Результат этапа сбора данных – таблица разметки, дополненная клиническими признаками (таблица 5).

Таблица 5 – Пример таблицы разметки, дополненной клиническими признаками

Анонимизированный номер исследования	Наличие патологии	Пол	Возраст	ИМТ
1111111	1	ж	35	30
2222222	0	ж	41	19
3333333	1	м	67	24

В таблице 6 представлены основные проблемы, с которыми может столкнуться разметчик при сборе данных.

Таблица 6 – Возможные проблемы при сборе данных и способы их решения

Проблема	Возможная причина	Решение
Отсутствие / недостаточное количество необходимой клинической информации	Неполный анамнез. Например, врач детально указал анамнез жизни и заболевания при первичном обращении пациента и не дублировал эту информацию при повторной консультации.	Проверить протоколы осмотров в динамике и записи других врачей.
	Исследование параметра происходит в стационаре	Смотреть выписки из стационаров, в них может содержаться больше клинической информации о пациенте в связи с большим количеством обследований, проводимых при госпитализации.
	Редко встречаемый параметр	Оставить пропуск. Дальнейшая обработка пропусков в данных описана в пункте 3.2. Обработка данных.
Ошибки переноса данных из ЭМК в таблицу разметки	Неправильная интерпретация клинической информации	Сбор данных рекомендуется выполнять человеку с медицинским образованием для правильной интерпретации медицинских терминов.
	Ошибки и опечатки в протоколах осмотров	Рекомендуется сверять данные, полученные из протоколов осмотра на приеме, с другими протоколами осмотра врачей той же или другой специальности.
	Ошибки и опечатки разметчика при переносе информации	Рекомендуется проводить выборочную проверку собранных данных, сравнивая их с ЭМК.

3.2. Обработка данных

Обработка данных включает следующие основные этапы:

1. Загрузка собранных данных.
2. Предварительная обработка данных:
 - a) проверка данных на пропуски;
 - b) проверка данных на дубликаты (полные) – когда это необходимо;
 - c) проверка данных на аномалии или выбросы.
3. Формирование итогового набора данных.

Предварительная обработка данных

Не всегда дополнительно собранные клинические данные, которые были выбраны для обогащения НД, находятся в надлежащем виде. Поэтому прежде чем добавлять их в окончательный НД, на котором будет проводиться обучение ПО на основе ТИИ, необходимо провести их предварительную обработку.

Проверка данных на пропуски

В первую очередь необходимо понять, удалось ли выгрузить достаточно новых клинических данных из МИС (ЭМК). Для этого нужно подсчитать количество пропусков данных по каждому новому признаку, которым мы обогащаем НД. Количество пропусков по признаку лучше представлять в процентах, так легче воспринимать подобную информацию:

1. Если процент пропусков более 5 %, наиболее целесообразным будет этот признак удалить, если в ТЗ не предусмотрено иное. Если построить доверительный интервал для среднего значения признака, то с определенной вероятностью (95 %) можно сказать, что истинное среднее значение признака находится в этом диапазоне. Если в столбце с признаком процент пропусков в данных более 5 %, то это означает, что не имеется достаточного количества наблюдений для вычисления достоверного доверительного интервала. В этом случае нельзя быть уверенным, что среднее значение признака, вычисленное на основе имеющихся данных, является достоверным и точным.

2. Процент пропусков 5 % или менее. В этом случае необходимо точно понимать, с какой целевой патологией создается набор данных и для исследования какого заболевания он может быть использован. **Если это заболевание достаточно редкое** в исследуемой популяции больных, то любые данные (и признаки) очень важны. В этом случае можно попытаться восстановить пропуски. Осуществить это можно с помощью трех подходов:

- A. Клинический. Если есть возможность – дообследовать пациента и таким образом восстановить пропуски в данных (например, если это какой-то гормон или размер какого-то органа).
- B. Использовать метод KNN («метод k-ближайших соседей»). Объяснение принципа и методологии использования этого метода не

является целью настоящих методических рекомендаций. Подробно ознакомиться с ним можно в статье [14]. Пример практического использования данного метода в среде программирования Python представлен по ссылке [15].

С. Применить алгоритм Гиббса. С использованием данного метода можно ознакомиться здесь [16].

Если патология (заболевание) часто встречается в общей популяции, тогда можно пропуски заменить на NaN (Not a Number). NaN – понятный для электронной-вычислительной машины символ, означающий, что в ячейке нет числа, но есть пропуск. Это важный момент: нельзя оставлять ячейку пустой, также в нее нельзя вводить значение 0. В дальнейшем значение 0 может серьезно повлиять на обучение ПО на основе ТИИ в негативном ключе.

Для наглядности блок-схема последовательности действий представлена на рисунке 8.



Рисунок 8 – Блок-схема. Последовательность действий для проверки добавленных клинических данных

Проверка данных на дубликаты

Если в ходе обработки данных будут обнаружены полностью дублирующие друг друга строки – необходимо оставить только первое вхождение таких данных в НД. Остальную полностью повторяющуюся часть (дубликаты) необходимо удалить, если в ТЗ не оговорено иное.

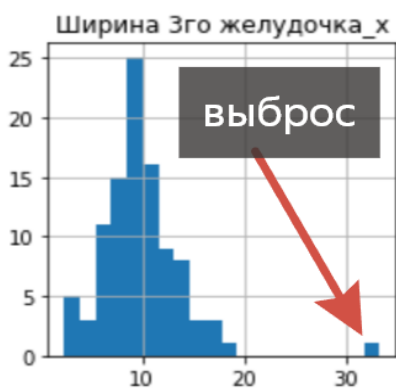
Проверка данных на выбросы

Помимо пропусков в данных, есть еще одна распространенная проблема, с которой можно столкнуться при обогащении НД – это «выбросы».

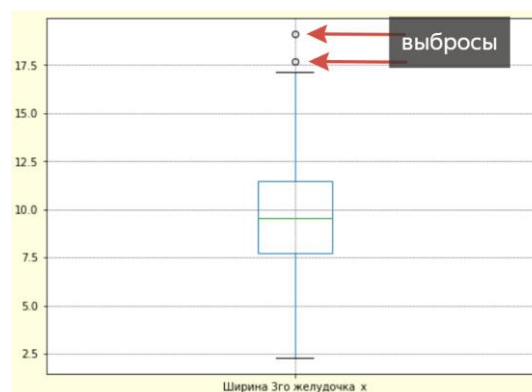
«Выброс» (или аномалия) – это значение (в разрезе одного признака), которое сильно отличается от остальных. Например, частота сердечных сокращений у пациентов с целевой патологией в спокойном состоянии в исследуемом наборе данных в среднем будет равняться 75 уд/мин, а среди них появится значение – 150 уд/мин. 150 уд/мин – и есть выброс (или аномалия). Это значение сердечных сокращений значительно отличается от остальных

значений признака в контексте данного исследования. Однако не всегда понятно, что делать с такими данными.

Для проверки НД на выбросы рекомендуем использовать следующие графические методы: построение графика «ящик с усами», построение гистограмм (накопления, с большим числом корзин). Дополнительно можно использовать статистические критерии: Граббса [17] или Диксона [18]. Оба критерия используются для обнаружения выбросов в данных, но критерий Граббса считается более консервативным, чем критерий Диксона, так как он учитывает только одно значение выброса, в то время как критерий Диксона учитывает несколько значений выброса. Выбор того или иного критерия зависит от конкретной задачи и характеристик данных (нормальное распределение). На рисунке 9 (А, Б) представлены примеры графических методов поиска выбросов. Стоит уточнить: это разные наборы данных. На рисунке под буквой А видно значение, значительно отстоящее от остальных. Вероятно, это выброс. Однако стоит иметь в виду то, что это может быть каким-либо казуистическим случаем (аномалия). Ширина 3-го желудочка головного мозга у этого больного действительно имела такой размер – может быть, у него была гидроцефалия? Для того чтобы убедиться в том, что это выброс, необходимо обратиться к МИС, из которой собираются дополнительные клинические признаки, и уточнить диагнозы больного. Под буквой Б показан пример поиска выбросов с помощью графика «ящик с усами». Все, что выше или ниже $1,5 \cdot IQR$ (усы), – считается выбросом.



А



Б

Рисунок 9 – Графические методы поиска выбросов:
А – гистограмма; Б – «ящик с усами»

В случае доказанного наличия выбросов рекомендуем придерживаться следующей схемы работы с ними:

1. Клиническое восстановление данных. Если это возможно – пересмотреть данные, откуда получили значение признака. Например, если

это электрокардиограмма – пересчитать частоту сердечных сокращений и заменить выброс на пересмотренное значение.

2. Если при наличии целевой патологии возможно подобное изменение в ваших данных, а клинически восстановить значение невозможно – оставить выброс как есть.

3. Если при наличии целевой патологии подобное изменение признака не типично и клинически пересмотреть значение признака также невозможно – целесообразно удалить такое значение. Необходимо уточнить, что настоящее издание носит рекомендательный характер, конечное решение об удалении выброса или о его сохранении принимает исследователь.

Общая блок-схема действий с выбросами представлена на рисунках 10 и 11.

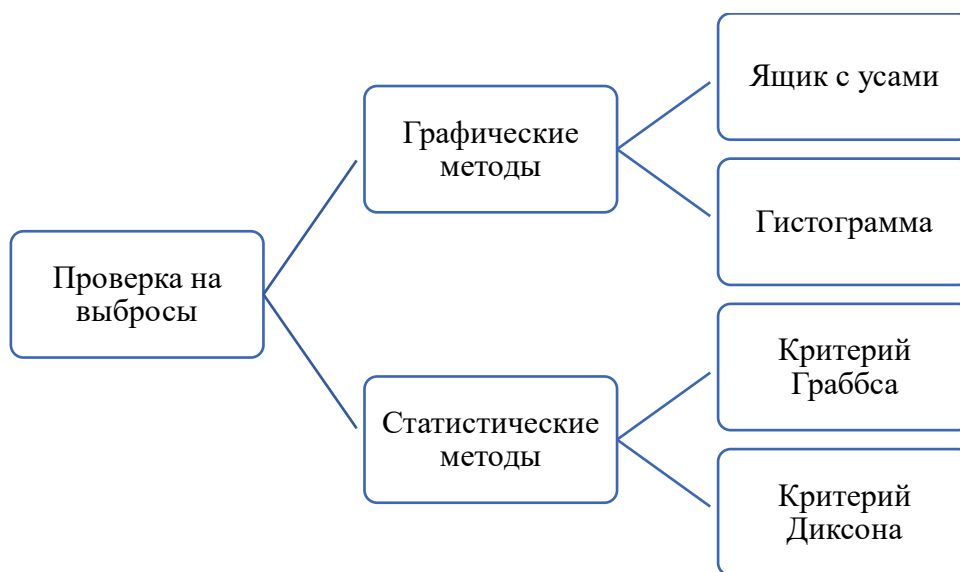


Рисунок 10 - Блок-схема проверки на выбросы

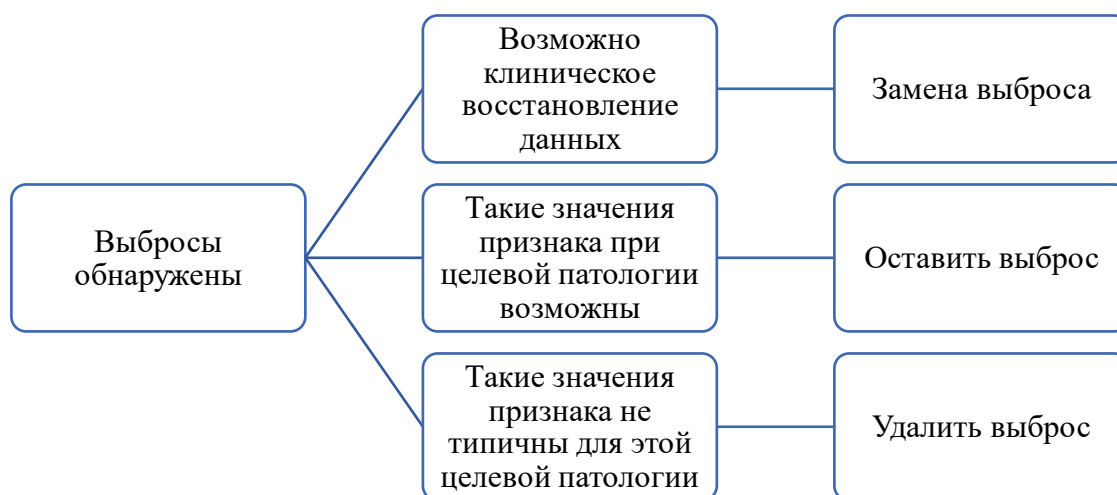


Рисунок 11 - Блок-схема действий с выбросами

Соблюдение этих простых рекомендаций позволит получать более точные и надежные результаты при работе с данными.

3.3. Readme-файл

Для удобства хранения, использования и публикации НД необходимо по завершении основных работ по подготовке НД сформировать сопроводительный текстовый файл (readme). Readme-файл содержит краткое структурированное описание НД, основные его параметры, данные об авторах, цитировании, организацию файлов. Он необходим для оперативного доступа к упорядоченной информации о НД. Особенно важно наличие таких файлов при длительном хранении файлов, большом количестве НД в хранилище, а также публикации в библиотеках НД. Readme-файл рекомендовано хранить в двух форматах (pdf и md) на двух языках (русский и английский). Такой формат охватывает максимальное количество конечных пользователей, как исследователей, так и разработчиков. Кроме информационной функции readme-файл может быть инструментом автоматизации при использовании НД (например, автоматическое извлечение информации при тестировании ПО на основе ТИИ на специализированных платформах).

На рисунке должно включать в себя следующие данные:

1. Общие положения:

Обзор данных

Параметр	Значение
Количество исследований, ед.	100
Количество пациентов, чел.	100
Распределение по полу, чел. (М/ Ж)	0/ 100
Распределение по возрасту, лет (мин./ медиана/ макс.)	48.0/ 63.0/ 71.0
Распределение по классам, ед. (С патологией/ Без патологии)	50/ 50
Распределение по классам, ед. (BIRADS 0 /1 / 2)	56/24/20
Распределение по классам, ед. (Код МКБ-10 C50.1, C50.2,C50.3,C50.4,C50.5,C50.8,C50.9,C50.1, N60.1, N60.3, N60.8, N60.9, N64.1, нет)	4/2/2/29/1/8/4/7/2/1/1/1/38
Распределение по классам, ед. (Количество родов 0, 1, 2, 3, 4, 6, n\а)	8/32/37/5/1/1/16

- наименование НД;
- версия;
- заказчик;
- источник финансирования;
- исполнитель;
- основание для создания НД;
- ключевые слова;

- авторы.
2. Назначение и цель создания:
 - назначение, область применения, целевая аудитория;
 - цель создания (тип НД).
 3. Параметры отбора данных:
 - наименование процедуры (основное исследование);
 - критерии включения;
 - критерии исключения;
 - период сбора данных;
 - количество исследований.
 4. Источник данных.
 5. Обезличивание:
 - дополнительная защита интеллектуальной собственности.
 6. Разметка данных:
 - необходимость разметки с привлечением медицинских специалистов по направлению;
 - количество и уровень медицинского персонала, привлекаемого к разметке;
 - разметчики;
 - требования, предъявляемые к разметчикам;
 - целевая патология;
 - клинические параметры;
 - допустимое значение пропусков клинических параметров;
 - временной интервал между основным исследованием и клиническими параметрами.
 7. Состав итогового НД.
 8. Требуемый объем памяти для хранения НД.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 141

Авторы-составители:

*Васильев Юрий Александрович
Казаринова Вероника Евгеньевна
Бобровская Татьяна Михайловна
Никитин Никита Юрьевич
Памова Анастасия Петровна
Арзамасов Кирилл Михайлович*

**ПОДГОТОВКА НАБОРОВ ДАННЫХ,
ОБОГАЩЕННЫХ КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ**

Методические рекомендации

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технический редактор В. П. Гамарина
Компьютерная верстка Е. Д. Бугаенко

ГБУЗ «НПКЦ ДиТ ДЗМ»

127051, г. Москва, ул. Петровка, д. 24, стр. 1 представлен фрагмент readme-файла (обзор данных) для обогащенного клинической информацией НД, описанного в главе ПРИМЕР

Рисунок 12 – Фрагмент readme-файла для НД, обогащенного клинической информацией

4. ПРИМЕР ПОДГОТОВКИ НАБОРА ДАННЫХ, ОБОГАЩЕННОГО КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ

Приведем пример подготовки набора данных, обогащенного клинической информацией, следуя настоящим методическим рекомендациям.

Этап 1 – инициирование

На этапе инициирования была определена цель – подготовить НД, обогащенный клинической информацией, для научных целей, а также для тестирования ПО на основе ГИИ, оценивающего не только маммографические изображения, но и клиническую информацию. Стратегия – обогатить существующий НД, содержащий маммографические изображения с наличием и отсутствием признаков ЗНО молочной железы [19].

Этап 2 – планирование

На этапе планирования было подготовлено ТЗ. Объем выборки был определен исходным НД, содержащим исследования 100 пациентов.

По результатам литературного анализа был сформирован список клинических параметров для обогащения. Он включал: возраст пациентов на момент проведения исследования, возраст наступления менархе, возраст наступления менопаузы, количество родов, возраст первых родов, наличие и продолжительность лактации, использование заместительной гормональной терапии, ИМТ, онкологический анамнез, мутации в генах BRCA1, BRCA2, курение, злоупотребление алкоголем и другое.

Этап 3 – формирование

Был произведен сбор клинической информации из МИС и обработка данных – параметры, заполненные менее чем на 30 %, были исключены (согласно ТЗ). Результат этапа – таблица разметки, дополненная клиническими признаками (рисунок 13).

Анонимизированный номер исследования	Патология	BIRADS	Диагноз (код по МКБ)	Возраст на момент исследования	Возраст наступления менопаузы	Количество родов
1.2.643.5.1.13.13.12.2.77.8252.05020007001	0	2	N60.1	51	37	1
1.2.643.5.1.13.13.12.2.77.8252.12110103050	1	0	C50.8	60	55	2
1.2.643.5.1.13.13.12.2.77.8252.10040909000	1	0	C50.4	66	56	2
1.2.643.5.1.13.13.12.2.77.8252.14010500060	1	0	C50.4	65	48	1
1.2.643.5.1.13.13.12.2.77.8252.02150106071	1	0	C50.4	59	55	2
1.2.643.5.1.13.13.12.2.77.8252.12061402090	1	0	C50.8	64	51	1
1.2.643.5.1.13.13.12.2.77.8252.14031004111	1	0	C50.4	64	56	0
1.2.643.5.1.13.13.12.2.77.8252.04130005110	1	0	C50.4	50	50	1
1.2.643.5.1.13.13.12.2.77.8252.08060813061	1	0	C50.4	67	52	2
1.2.643.5.1.13.13.12.2.77.8252.01100401031	1	0	C50.4	68	50	0
1.2.643.5.1.13.13.12.2.77.8252.11131404000	1	0	C50.4	63	50	2
1.2.643.5.1.13.13.12.2.77.8252.10140814010	1	0	C50.8	68	46	1
1.2.643.5.1.13.13.12.2.77.8252.11110302121	0	0	N60.1	58	55	1
1.2.643.5.1.13.13.12.2.77.8252.15100303140	1	0	C50.4	56	51	1

Рисунок 13 – Фрагмент полученной таблицы разметки, дополненной клиническими признаками

Подготовлен Readme-файл (см. рисунок) должно включать в себя следующие данные:

9. Общие положения:

- наименование НД;
- версия;
- заказчик;
- источник финансирования;
- исполнитель;
- основание для создания НД;
- ключевые слова;
- авторы.

10. Назначение и цель создания:

- назначение, область применения, целевая аудитория;
- цель создания (тип НД).

11. Параметры отбора данных:

- наименование процедуры (основное исследование);
- критерии включения;
- критерии исключения;
- период сбора данных;
- количество исследований.

12. Источник данных.

13. Обезличивание:

- дополнительная защита интеллектуальной собственности.

14. Разметка данных:

- необходимость разметки с привлечением медицинских специалистов по направлению;
- количество и уровень медицинского персонала, привлекаемого к разметке;
- разметчики;
- требования, предъявляемые к разметчикам;
- целевая патология;
- клинические параметры;
- допустимое значение пропусков клинических параметров;
- временной интервал между основным исследованием и клиническими параметрами.

15. Состав итогового НД.

16. Требуемый объем памяти для хранения НД.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 141

Авторы-составители:

*Васильев Юрий Александрович
Казаринова Вероника Евгеньевна
Бобровская Татьяна Михайловна
Никитин Никита Юрьевич
Памова Анастасия Петровна
Арзамасов Кирилл Михайлович*

**ПОДГОТОВКА НАБОРОВ ДАННЫХ,
ОБОГАЩЕННЫХ КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ**

Методические рекомендации

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технический редактор В. П. Гамарина
Компьютерная верстка Е. Д. Бугаенко

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

Далее НД был оформлен в реестре НД и зарегистрирован – получено свидетельство о государственной регистрации базы данных [20].

Таким образом, процесс подготовки НД, обогащенного клинической информацией, можно представить в виде схемы (рисунок 14).

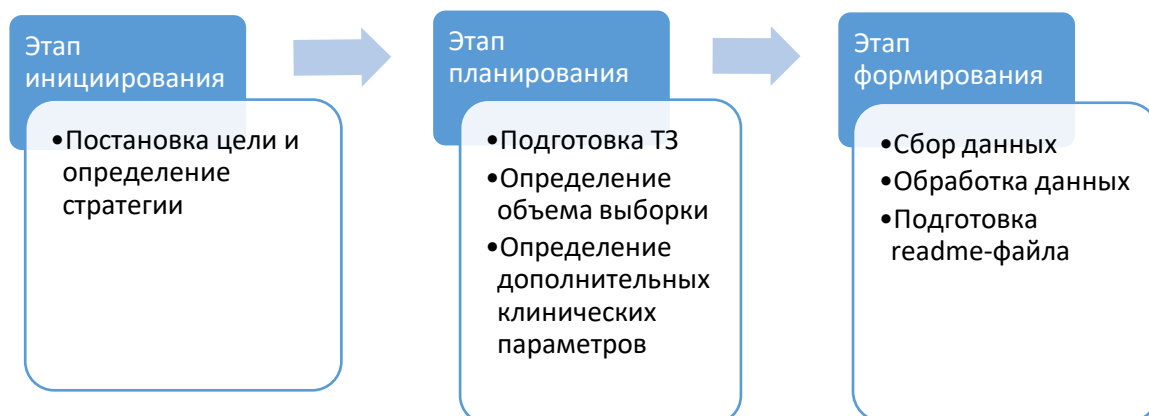


Рисунок 14 – Схема процесса подготовки НД, обогащенного клинической информацией

ЗАКЛЮЧЕНИЕ

Благодаря опыту подготовки обогащенных наборов данных, накопленному сотрудниками ГБУЗ «НПКЦ ДиТ ДЗМ», а также их последующей регистрации, удалось создать настоящие методические рекомендации. Основной целью являлось описание процесса поиска дополнительных клинических параметров, их сбора и дальнейшей обработки. Предложенный алгоритм позволяет оптимизировать процесс подготовки набора данных и избежать ошибок при его обогащении клиническими данными. Необходимо помнить, что сбор и обогащение набора данных должны происходить строго в соответствии с этикой работы с клинической информацией и со ст. 13 ч. 3 Федерального закона от 21.11.2011 № 323-ФЗ [5].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» : сайт. – URL: <http://www.kremlin.ru/acts/bank/44731> (дата обращения 23.05.2024). – Текст: электронный.
2. Регламент подготовки наборов данных с описанием подходов к формированию репрезентативной выборки данных. – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», 2022. – 40 с.
3. Васильев Ю. А., Арзамасов К. М., Владзимирский А. В. [и др.]. Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта : учебное пособие. – Ridero ; ГБУЗ «НПКЦ ДиТ ДЗМ», 2024. – 140 с.
4. ГОСТ Р ИСО 21549-3-2017. Информатизация здоровья. Структура данных на пластиковой карте пациента. Часть 3. Основные клинические данные : национальный стандарт Российской Федерации : дата введения 2019-07-01 / Федеральное агентство по техническому регулированию и метрологии. – Изд. официальное. – М. : Стандартинформ, 2017. – 19 с.
5. Российская Федерация. Законы. Об основах охраны здоровья граждан в Российской Федерации : Федеральный закон № 323-ФЗ : [принят Государственной Думой 1 ноября 2011 г. : одобрен Советом Федерации 9 ноября 2011 г.]. – Москва, Кремль 2011. – 120 с. – Текст : непосредственный.
6. Hsieh C., Nobre I. B., Sousa S. C., et al. MDF-Net for abnormality detection by fusing X-rays with clinical data // *Sci Rep.* – 2023. – Vol. 13(1). – P. 15873. – DOI: 10.1038/s41598-023-41463-0.
7. Fu Y., Xue P., Li N., et al. Fusion of 3D lung CT and serum biomarkers for diagnosis of multiple pathological types on pulmonary nodules // *Comput Methods Programs Biomed.* – 2021. – Vol. 210. – P. 106381. – DOI: 10.1016/j.cmpb.2021.106381.
8. Lin C. Y., Guo S. M., Lien J. J., et al. Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT // *Radiol Med.* – 2024. – Vol. 129(1). – P. 56-69. – DOI: 10.1007/s11547-023-01730-6.
9. Yala A., Mikhael P. G., Strand F., et al. Toward robust mammography-based models for breast cancer risk // *Sci Transl Med.* – 2021. – Vol. 13(578). – P. eaba4373. – DOI: 10.1126/scitranslmed.aba4373.
10. Васильев Ю. А., Бобровская Т. М., Арзамасов К. М. [и др.]. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской

- диагностике // Менеджер здравоохранения. – 2023. – № 4. – С. 28-41. – DOI: 10.21045/1811-0185-2023-4-28-41.
11. Отдельнова К. А. Определение необходимого числа наблюдений в социально-гигиенических исследованиях // Сб. трудов 2-го ММИ. – 1980. Т. 150 (6). – С. 18–22.
 12. Altman D. G. How large a sample? *Statistics in Practice*. London, UK: British Medical Association; 1982. // *Statistics in Practice*. – 1982.
 13. Whitley E., Ball J. Statistics review 4: sample size calculations // *Critical care*. – 2002. – Т. 6. – С. 1-7.
 14. Liu J., Lei J., Ou Y., et al. Mammography diagnosis of breast cancer screening through machine learning: a systematic review and meta-analysis // *Clin Exp Med*. – 2023. – Vol. 23(6). – P. 2341-2356. – DOI: 10.1007/s10238-022-00895-0.
 15. kNN Classifier Tutorial : сайт. – URL: <https://www.kaggle.com/code/prashant111/knn-classifier-tutorial> (дата обращения: 11.03.2024). – Текст: электронный.
 16. Liu M., Taylor J. M., Belin T. R. Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies // *Biometrics*. – 2000. – Vol. 56(4). – P. 1157-63. – DOI: 10.1111/j.0006-341x.2000.01157.x.
 17. Grubbs, F. E. Procedures for Detecting Outlying Observations in Samples // *Technometrics*. – 1969. – Vol. 11(1). – P 1. – doi:10.2307/1266761.
 18. Robert B. Dean, Wilfrid J. Dixon Simplified Statistics for Small Numbers of Observations // *Anal. Chem*. – 1951. – Vol. 23 (4). – P. 636–638. –doi: 10.1021/ac60052a025.
 19. Свидетельство о государственной регистрации базы данных № 2020621741 Российская Федерация. MosMedData: Результаты маммографических исследований для калибровки сервисов на основе искусственного интеллекта : № 2020621613 : заявл. 15.09.2020 : опубл. 24.09.2020 / С. П. Морозов, А. В. Владимирский, В. А. Гомболевский [и др.] ; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».
 20. Свидетельство о государственной регистрации базы данных № 2023624880 Российская Федерация. MosMedData: ММГ с наличием и отсутствием признаков злокачественных новообразований молочной железы, обогащенный клинической информацией : № 2023624807 : заявл. 14.12.2023 : опубл. 21.12.2023 / Ю. А. Васильев, А. В. Владимирский, О. В. Омелянская [и др.] ; заявитель ГБУЗ «НПКЦ ДиТ ДЗМ».

ТЗ должно включать в себя следующие данные:

17. Общие положения:

- наименование НД;
- версия;
- заказчик;
- источник финансирования;
- исполнитель;
- основание для создания НД;
- ключевые слова;
- авторы.

18. Назначение и цель создания:

- назначение, область применения, целевая аудитория;
- цель создания (тип НД).

19. Параметры отбора данных:

- наименование процедуры (основное исследование);
- критерии включения;
- критерии исключения;
- период сбора данных;
- количество исследований.

20. Источник данных.

21. Обезличивание:

- дополнительная защита интеллектуальной собственности.

22. Разметка данных:

- необходимость разметки с привлечением медицинских специалистов по направлению;
- количество и уровень медицинского персонала, привлекаемого к разметке;
- разметчики;
- требования, предъявляемые к разметчикам;
- целевая патология;
- клинические параметры;
- допустимое значение пропусков клинических параметров;
- временной интервал между основным исследованием и клиническими параметрами.

23. Состав итогового НД.

24. Требуемый объем памяти для хранения НД.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 141

Авторы-составители:

*Васильев Юрий Александрович
Казаринова Вероника Евгеньевна
Бобровская Татьяна Михайловна
Никитин Никита Юрьевич
Памова Анастасия Петровна
Арзамасов Кирилл Михайлович*

**ПОДГОТОВКА НАБОРОВ ДАННЫХ,
ОБОГАЩЕННЫХ КЛИНИЧЕСКОЙ ИНФОРМАЦИЕЙ**

Методические рекомендации

Отдел координации научной деятельности ГБУЗ «НПКЦ ДиТ ДЗМ»
Технический редактор В. П. Гамарина
Компьютерная верстка Е. Д. Бугаенко

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