

Паспорт набора данных

1. Идентификатор набора данных.

Будет уточнен.

2. Версия набора данных.

1.0

3. Наименование набора данных.

Датасет по задаче прогнозирования развития доброкачественных новообразований толстой кишки.

4. Описание набора данных.

4.1. Аннотация

В датасете будут представлены данные 190 пациентов с установленным диагнозом доброкачественного новообразования толстой кишки, предназначенных для обучения моделей машинного обучения для прогнозирования развития доброкачественного новообразования толстой кишки.

4.2. Клиническая задача

Структура доброкачественных новообразований толстой кишки представлена тубулярными, ворсинчатыми, тубуловорсинчатыми и зубчатыми аденомами. Современные оценки распространенности аденом основаны на исследованиях колоноскопий, которые обычно сообщают об наличии доброкачественных новообразований толстой кишки у 20–53% лиц в возрасте 50 лет и старше, в том числе до 9,7% с запущенными формами, которые имеют высокий риск озлакоственения.

В этих условиях, изучение факторов риска и сопутствующих диагнозов у лиц с доброкачественными новообразованиями толстой кишки, могут помочь своевременно установить диагноз, что является стратегически важной задачей для современного здравоохранения и РФ.

4.3. Назначение набора данных

Датасет предназначен для разработки ML-решений автоматического анализа цифровых данных, что позволит упростить своевременную и точную постановку диагноза.

4.4. Нозологии

- D12 Доброкачественное новообразование ободочной кишки, прямой кишки, заднего прохода [ануса] и анального канала
- D12.0 Слепой кишки
- D12.1 Червеобразного отростка [аппендикса]
- D12.2 Восходящей ободочной кишки
- D12.3 Поперечной ободочной кишки
- D12.4 Нисходящей ободочной кишки
- D12.5 Сигмовидной кишки
- D12.6 Ободочной кишки неуточненной части
- D12.7 Ректосигмоидного соединения
- D12.8 Прямой кишки
- D12.9 Заднего прохода [ануса] и анального канала

5. Владелец набора данных

Федеральное государственное автономное образовательное учреждение высшего образования Первый Московский государственный медицинский университет имени И. М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет).

5.1. Контактные данные

119991, г. Москва, ул. Трубецкая, д. 8, стр. 2; Тел.: +7 (495) 609-14 00; Эл. почта.: rektorat@sechenov.ru;

5.2. Авторы

- Осадчук Михаил Алексеевич
- Бражников Константин Валерьевич
- Миронова Екатерина Дмитриевна

6. Порядок предоставления доступа к набору данных

Данные пациентов будут взяты из электронного архива данных ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет).

7. Тэги

ДАТАСЕТ, ЕДИНАЯ ГОСУДАРСТВЕННАЯ ИНФОРМАЦИОННАЯ СИСТЕМА В СФЕРЕ ЗДРАВООХРАНЕНИЯ, МЕДИЦИНСКАЯ ИНФОРМАЦИОННАЯ СИСТЕМА, РАЗМЕТКА ДАННЫХ, ГАСТРОЭНТЕРОЛОГИЯ, ГАСТРОЭНТЕРОЛОГИЕЧКЕ ЗАБОЛЕВАНИЯ, ДОБРОКАЧЕСТВЕННОЕ НОВООБРАЗОВАНИЕ ТОЛСТОЙ КИШКИ, АДЕНОМАТОЗНЫЙ ПОЛИП.

8. Источник данных

Федеральное государственное автономное образовательное учреждение высшего образования Первый Московский государственный медицинский университет имени И. М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет).

8.1. Глубина набора данных

В набор будут включены данные пациентов с установленным диагнозом доброкачественное новообразование толстой кишки из архива ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет) с апреля 2022 года по апрель 2023 год.

8.2. Объем набора данных

Планируется произвести ретроспективный набор данных 190 пациентов с установленным диагнозом доброкачественное новообразование толстой кишки.

8.3. Критерии отбора клинических параметров и пациентов им соответствующих, а также виды лабораторно-инструментальных исследований.

Для анализа будут отобраны пациенты с установленным диагнозом доброкачественное новообразование толстой кишки в ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет).

Критерии включения:

Установленный диагноз доброкачественное новообразование толстой кишки у лиц в возрасте от 18 до 99 лет

8.4. Особенности подготовки исследований

Данные пациентов обезличиваются в процессе выгрузки из архива ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет).

8.5. Поло-возрастные характеристики набора данных

- Планируется включить в базу данные 92 пациентов с ранее установленным диагнозом доброкачественное новообразование толстой кишки мужского пола и 98 пациентов женского пола.
- Минимальный возраст 18 лет, максимальный 99 лет.

8.6. Характеристики оборудования

Постановка диагноза доброкачественное новообразование толстой кишки осуществляется на основании результатов морфологического анализа.

8.7. Дата публикации набора данных

Будет уточнена

8.8. Дата обновления набора данных

Не обновлялся

9. Модели машинного обучения

Нейронные сети, градиентный бустинг, случайный лес и другие классифицирующие модели машинного обучения

10. Расположение набора данных

Будет уточнен

11. Структура набора данных

11.1. Схема структуры директорий и файлов

----- train_set

|----- Labels.xlsx

11.2. Описание файлов в корневой директории

Файл содержит данные в текстовом формате с разбиением на строки (по исследованиям) и столбцы. Набор столбцов соответствует параметрам, перечисленным в разделе «12. Обзор разметки данных».

12. Обзор разметки данных

12.1. Особенности формата разметки

Номера пациентов включенных в датасет соответствуют номерам исследований в архиве ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет).

12.2. Классы разметки

Результатом разметки будет является таблица, содержащая анамнестические данные, половозрастные характеристики, а также сопутствующие диагнозы.

Таблица 1. Размеченные параметры

Наименование	Тип данных	Раздел для поиска
№ пациента	Число	
Код-идентификатор пациента	цифра	Данные о пациенте
Возраст	Число	Данные о пациенте
Пол	М/Ж	Данные о пациенте
Основной диагноз	число+текст	Диагноз
Код МКБ-10 + диагноз	число+текст	Диагноз

12.3. Принципы разметки и верификации

Верификация датасета производится по следующему алгоритму:

1. При помощи использования регулярных выражений в исходных данных обнаруживаются записи, содержащие упоминание каждого из выявляемых критических событий в соответствии со словарем, содержащим все варианты медицинской идентификации критических событий;

2. Врачи-эксперты просматривают отобранные записи, исключая из них анамнестические данные и нерелевантные события;

3.Перекрестная проверка другим экспертом проводится для подтверждения наличия в электронной медицинской карте пациента записей о соответствующем состоянии. Установленный в медицинской карте диагноз считается достоверным;

4.В случае прохождения всех этапов верификации метаданные записей вносятся в датасет.

12.4. Статистика использования лейблов и классов

По итогу набора данных планируется произвести статистический анализ по прогнозированию развития доброкачественного новообразования толстой кишки.

13. Правила использования и распространения

13.1. Копирайт

ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России
(Сеченовский Университет)