Паспорт набора данных

1. Идентификатор набора данных.

1120 пациентов с установленным диагнозом колоректальный рак.

2. Версия набора данных.

1.0

3. Наименование набора данных.

Датасет по задаче определения прогностической значимости индекса массы тела в развитии колоректального рака и его локализации в толстой кишке.

4. Описание набора данных.

4.1. Аннотация

В датасете будут представлены данные 1120 пациентов с установленным диагнозом колоректальный рак, предназначенных для обучения моделей машинного обучения для оценки прогностической значимости индекса массы тела в развитии колоректального рака и определения его локализации в толстой кишке.

4.2. Клиническая задача

Колоректальный рак является одним из ведущих онкологических заболеваний. В последние годы отмечается неуклонное повышение уровня заболеваемости в развитых странах вследствие увеличения продолжительности жизни населения, возрастания влияния общепопуляционных факторов риска и улучшения выявляемости рака за счет внедрения скрининговых программ.

На сегодняшний день многочисленные обсервационные исследования сообщают о связи между индексом массы тела и повышенным риском развития колоректального рака, а также вовлеченностью в развитие онкологического процесса сразу нескольких отделов толстого кишечника.

В этих условиях, изучение влияния показателей индекса массы тела у лиц с колоректальным раком, могут помочь определить прогностическую значимость индекса массы тела в развитии колоректального рака и определения его локализации в толстом кишечнике, 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

- C18 Злокачественное новообразование ободочной кишки
- C18.0 Злокачественное новообразование: слепой кишки
- C18.1 Злокачественное новообразование: червеобразного отростка [аппендикса]
- C18.2 Злокачественное новообразование: восходящей ободочной кишки
- C18.3 Злокачественное новообразование: печеночного изгиба
- C18.4 Злокачественное новообразование: поперечной ободочной кишки
- C18.5 Злокачественное новообразование: селезеночного изгиба
- C18.6 Злокачественное новообразование: нисходящей ободочной кишки
- C18.7 Злокачественное новообразование: сигмовидной кишки

C18.8 Злокачественное новообразование: поражение ободочной кишки, выходящее за пределы одной и более вышеуказанных локализаций

- C18.9 Злокачественное новообразование: ободочной кишки неуточненной локализации

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, ЗЛОКАЧЕСТВЕННОЕ НОВООБРАЗОВАНИЕ ТОЛСТОЙ КИШКИ, КОЛОРЕКТАЛЬНЫЙ РАК, ЛОКАЛИЗАЦИЯ РАКА

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 колоректальный рак из архива ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет) с апреля 2020 года по январь 2024 года.

8.2. Объем набора данных

Планируется произвести ретроспективный набор данных 1120 пациентов с установленным диагнозом колоректальный рак.

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

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

Критерии включения:

Установленный диагноз колоректальный рак 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 528 пациентов с ранее установленным диагнозом колоректальный рак мужского пола и 592 пациентов женского пола.
- Минимальный возраст 18 лет, максимальный 99 лет.

8.6. Характеристики оборудования

Постановка диагноза колоректальный рак осуществляется на основании результатов морфологического анализа.

8.7. Дата публикации набора данных

30.01.2024

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 дирректорий 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
Индекс массы тела	число	Данные о пациенте
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 колоректального рака и его локализации в толстой кишке.

13. Правила использования и распространения

13.1. Копирайт

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

д.м.н., профессор, заведующий кафедрой
поликлинической терапии ИКМ
им. Н.В. Склифосовского
Осадчук Михаил Алексеевич