Public Medical Text Datasets

I. Clinical Notes and Electronic Health Records

1、MIMIC Series (latest version: MIMIC-IV) — Access requires “CITI Data or Specimens Only Research” training via the PhysioNet website

File format: CSV

MIMIC-III includes the electronic health records of more than 40,000 critical care patients from 2001–2012, along with a large number of clinical notes such as nursing notes, laboratory reports, and imaging records.
MIMIC-III is organized as a relational database containing 26 tables, which are connected through identifiers (usually ending with “ID”).

https://physionet.org/content/mimiciii/1.4/

MIMIC-IV covers patient admissions to the ICU or emergency department between 2008 and 2022. It includes data from over 65,000 ICU patients and more than 200,000 emergency admissions.

MIMIC-IV is divided into two modules: **hosp** and **icu**, reflecting the source of each dataset. The *hosp* module contains data from the hospital-wide electronic health record system, while the *icu* module comes from the ICU’s MetaVision clinical information system.
In version 3.0 of MIMIC-IV, there are 364,627 unique individuals identified by $subject_{id}$, with a total of 546,028 hospital admissions and 94,458 distinct ICU stays.

https://physionet.org/content/mimiciv/3.1/

2、eICU Collaborative Research Database — Access requires “CITI Data or Specimens Only Research” training via the PhysioNet website

File format: CSV

This is a large-scale, multi-center ICU database from the United States, containing over 200,000 ICU admissions recorded between 2014 and 2015.
The dataset includes vital signs, laboratory test results, medications, APACHE components, care plan information, admission diagnoses, patient history, timestamped diagnoses in structured problem lists, and details on selected treatments.

https://physionet.org/content/eicu-crd/2.0/

3、UK Biobank — Application and payment required

This dataset includes information from **500,000 participants**, managed by UK Biobank Ltd. It provides longitudinal health data such as outpatient and inpatient records, diagnostic codes, and selected raw clinical records, as well as survey responses.
Researchers must submit a **paid application for scientific use**, and the data cannot be accessed publicly without approval.

The UK Biobank dataset contains a wide range of information:

- **Imaging data** (e.g., MRI scans, bone density scans, carotid ultrasounds)
- **Biomarker data** (proteins, metabolites, infectious disease markers)

- **Genetic data** (genotyping, exome, and whole-genome data)
- **Medical records** (hospitalizations, cancer diagnoses, and causes of death)
- **Questionnaire data** (self-reported health and lifestyle information)
- **Physical measurements** (vision and hearing tests, body composition, and activity monitoring)
- **Demographic and environmental data** (e.g., air and noise pollution levels in participants' local areas)

https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/

4、i2b2 (also known as n2c2: National NLP Clinical Challenges) — Access requires signing a data use agreement (DUA)

Original format: TXT; **processed format:** CSV

This dataset series is organized by the **Department of Biomedical Informatics (DBMI)** at Harvard Medical School and has been used in multiple clinical NLP shared tasks. It includes de-identified discharge summaries and other medical text from institutions such as **Beth Israel** and **Partners Healthcare**, with annotations for entities, relations (e.g., diseases, medications, lab tests), and de-identification tags.
Access requires registration on the Harvard DBMI data portal and signing the DUA.

https://n2c2.dbmi.hms.harvard.edu/data-sets

Year	Task Topic	Approximate Scale	Description
2006 i2b2	De-identification	~889 records	Radiology reports from Partners Healthcare used for identifying PHI such as names, locations, and dates.
2008 i2b2	Obesity Challenge	1,237 records	Each record (1–2 pages) is manually annotated with 19 diseases and their associations with obesity.
2009 i2b2	Medication Extraction	871 records	Focuses on extracting drug names, dosages, and frequencies.
2010 i2b2	Concept, Assertion, Relation Extraction	394 records	Includes annotations for three concept types—problem, test, and treatment—and their relations.
2012 i2b2	Temporal Relations	310 records	Focuses on time expression recognition and event sequencing.
2014 i2b2/UTHealth	PHI De-identification	1,304 records	Derived from MIMIC-II data (~2.3M words) and includes 25 PHI categories.

Year	Task Topic	Approximate Scale	Description
2018 n2c2	Adverse Drug Events and Medication Extraction	505 records	Each record includes multiple free-text sections from Partners Healthcare.
2019 n2c2	Smoking and Cohort Selection	~1,000 records	Short clinical notes mainly used for classification and cohort filtering tasks.
2022 n2c2	Clinical Text De-identification	~1,000 records	Extends the 2014 i2b2 dataset with new PHI categories and more complex contexts.

5、AmsterdamUMCdb — ICU database from Amsterdam University Medical Center, Netherlands (free access)

File format: CSV

This is a de-identified ICU database from Amsterdam University Medical Center, covering 23,106 ICU admissions (20,109 patients) from 2003–2016.
Access is free upon registration and ethical approval.

https://github.com/AmsterdamUMC/AmsterdamUMCdb?tab=readme-ov-file

II. Doctor–Patient Dialogues and Medical Question–Answering Datasets

1、PubMedQA

Proposed by **Jin et al. (2019)**, PubMedQA is a biomedical research–oriented question answering dataset.
The questions are derived from medical paper titles or excerpts, and the answers are categorized as **“yes,” “no,” or “maybe,”** based on conclusions drawn from PubMed abstracts.
The dataset contains **1,000 expert-annotated QA pairs**, **61.2k unlabeled examples**, and **211.3k automatically generated QA instances**.

Each PubMedQA entry consists of:

1. A **question**, usually derived from or identical to a research article title;
2. A **context**, which is the corresponding abstract excluding its conclusion;
3. A **long answer**, corresponding to the abstract’s conclusion, which is assumed to answer the question;
4. A **short answer**, summarizing the conclusion as “yes,” “no,” or “maybe.”

https://aclanthology.org/D19-1259/#:~:text=We%20introduce%20PubMedQA%2C%20a%20novel,of%20the%20abstract%20and%2C%20presumably

2、emrQA

Developed by **Pampari et al. (2018)**, emrQA is a large-scale clinical question–answering dataset automatically generated using **i2b2 annotated corpora**.
 It includes **over one million question–logical form pairs** and **more than 400,000 question–answer pairs**.
 All questions and answers are grounded in real discharge summaries, making it a valuable resource for studying reading comprehension and QA within **EHR (Electronic Health Record)** contexts.

The dataset is **publicly available** and designed for research on clinical reasoning and medical QA.
 Questions are generated **through templates**, then **manually verified by experts**.
 For each task category (Medication, Problem, Test, Temporal, Relation, etc.), approximately **100–200 pairs** were randomly sampled and validated.

https://aclanthology.org/D18-1258/#:~:text=annotations%20on%20clinical%20notes%20for,and%20question%20to%20answer%20mapping

3、MediTOD

MediTOD, proposed by **Saley et al. (2024)**, is an English medical task-oriented dialogue dataset focused on **clinical history–taking conversations**.
 It contains **22,503 annotated utterances** covering respiratory and musculoskeletal case scenarios, with **fine-grained intent and slot annotations**.
 The dataset is publicly available on GitHub and is suitable for training and evaluating medical dialogue systems, especially for natural language understanding (NLU), dialogue policy, and generation tasks.https://aclanthology.org/2024.emnlp-main.936.pdf#:~:text=1,designed%20for%20the%20medical%20domain

4、MTS-Dialog

Released by **Abacha et al. (2023)**, MTS-Dialog is a dataset of **doctor–patient conversations** paired with **structured clinical summaries** for outpatient and emergency settings.
 It contains **1,701 short dialogues** along with their corresponding summaries, divided into **training (1,201 pairs)**, **validation**, and **two test sets (200 pairs each)**.
 The dataset is **openly available under the MIT license**, supporting research in **medical dialogue summarization** and **clinical note generation**.

https://github.com/abachaa/MTS-Dialog

III. Biomedical Literature Abstract and Full-text Datasets

1、PubMed Database

Default output format: XML

Maintained by the **National Center for Biotechnology Information (NCBI)**, PubMed contains more than **39 million** citations and abstracts from biomedical and life science journals.
 While it does not host full-text articles, it covers a wide range of disciplines, including **medicine, health, life sciences, behavioral sciences, chemistry, and bioengineering**.
 Researchers can use the **PubMed API** to retrieve and download English titles and abstracts for text mining and biomedical information retrieval.

NCBI provides a suite of API tools called **E-utilities**, which allow programmatic access to PubMed data:

- `esearch`: searches the database and retrieves a list of IDs that match specific criteria.
- `efetch`: fetches detailed records (titles, authors, abstracts, journals, and other metadata) based on those IDs.

https://pubmed.ncbi.nlm.nih.gov/about/#:~:text=PubMed%20contains%20more%20than%2039,PMC

2、CORD-19

File format: JSON

CORD-19 (COVID-19 Open Research Dataset) is a large corpus of scientific papers related to **COVID-19** and other **coronavirus research**, curated by the **Allen Institute for AI's Semantic Scholar team** to support text mining and NLP research.

The **final release (June 2, 2022)** includes over **1 million indexed papers**, among which around **370,000 contain full-text content**.
This dataset has been widely used for **information extraction, question answering, and literature mining** in the biomedical domain.

https://github.com/allenai/cord19