The way to get MIMIC database

1. Step 1. Complete CITI "Data or Specimens Only Research" course

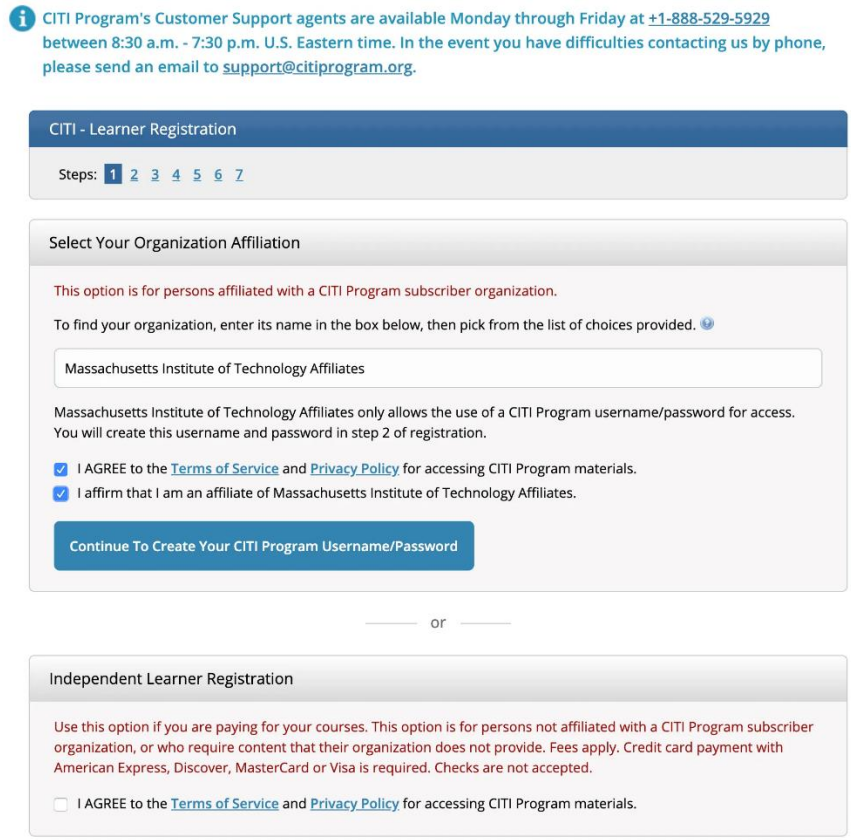
Complete the required training course

Prior to requesting access to MIMIC, you will need to complete the CITI "Data or Specimens Only Research" course:

- First register on the CITI program website, selecting "Massachusetts Institute of Technology Affiliates" as your affiliation (***not*** *"independent learner"*): https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0
- Follow the links to add a Massachusetts Institute of Technology Affiliates course. In the Human Subjects training category, select the "Data or Specimens Only Research" course
- Complete the course and save a copy of your completion report. The completion report lists all modules completed, with dates and scores.

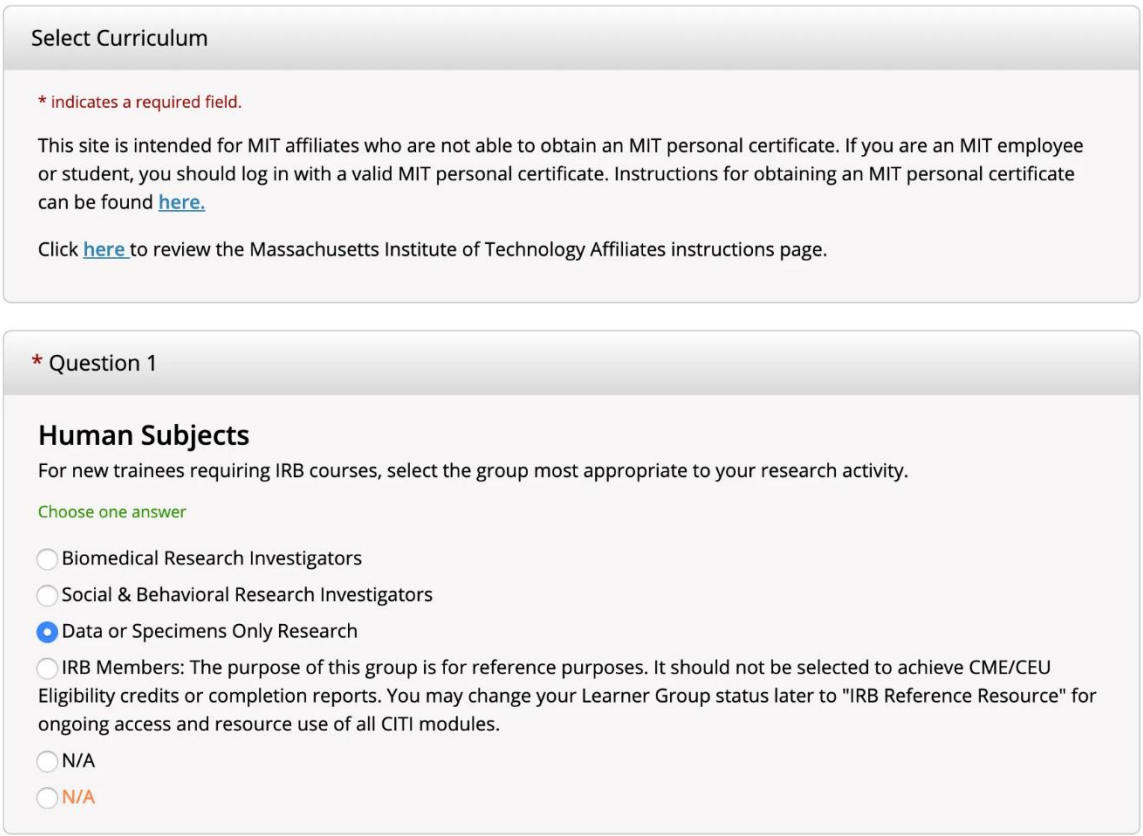
- Visit and create an account https://www.citiprogram.org/index.cfm?pageID=154&icat=0&ac=0

- In step 1, with "Select Your Organization Affiliation", fill in the blank with "Massachusetts Institute of Technology Affiliates"



- Complete step2 to step6

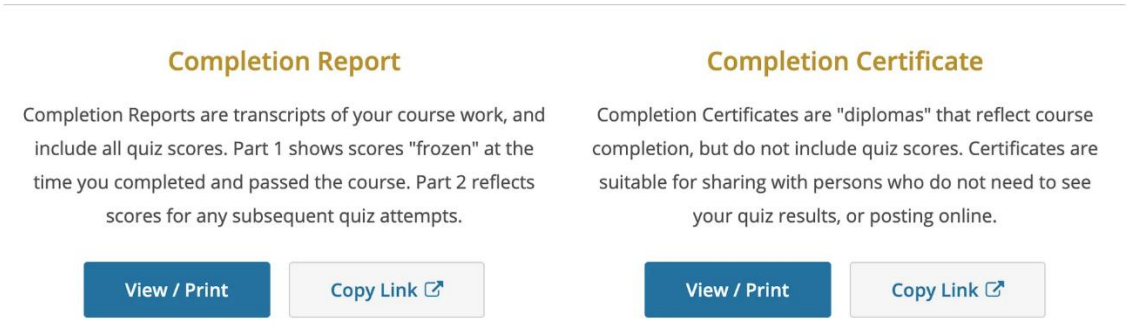
- In step 7, for Question 1, select “Data or Specimens Only Research ” in “Human Subjects”



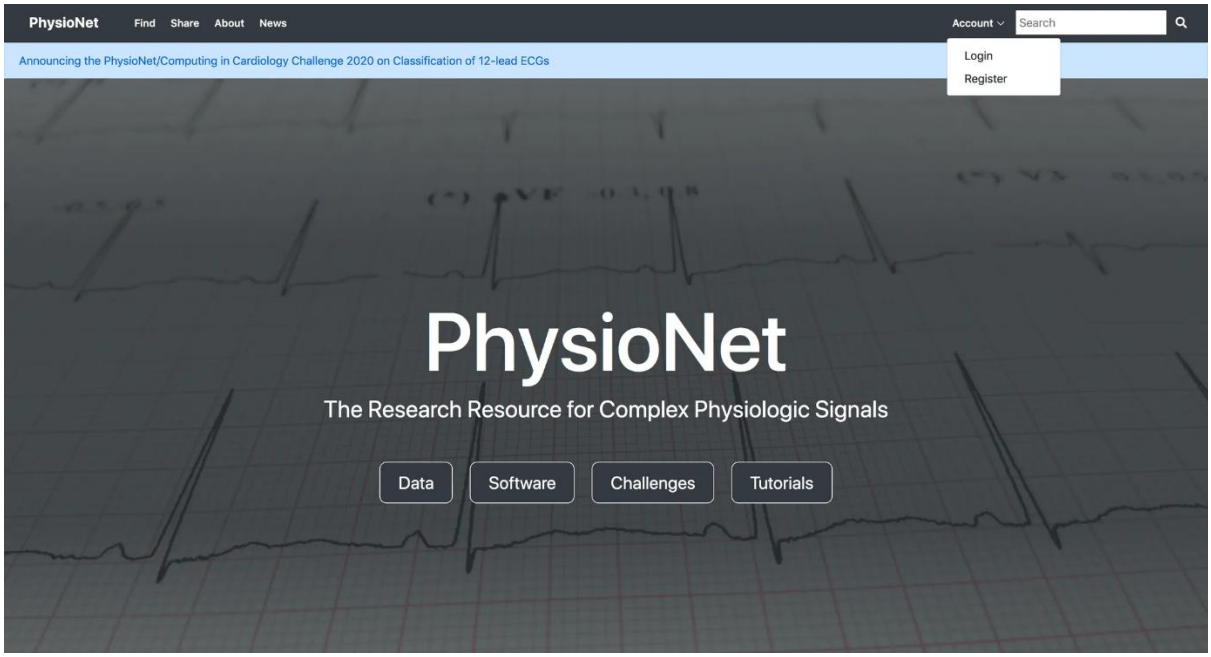
- After creating an account, complete all classes and quizzes. If the total score of quiz exceeds 90, you can complete the course

Modules	Completed	Score	
Belmont Report and Its Principles (ID 1127)	Incomplete	-	Start
History and Ethics of Human Subjects Research (ID 498)	Incomplete	-	Start
Basic Institutional Review Board (IRB) Regulations and Review Process (ID 2)	Incomplete	-	Start
Records-Based Research (ID 5)	Incomplete	-	Start
Genetic Research in Human Populations (ID 6)	Incomplete	-	Start
Populations in Research Requiring Additional Considerations and/or Protections (ID 16680)	Incomplete	-	Start
Research and HIPAA Privacy Protections (ID 14)	Incomplete	-	Start
Conflicts of Interest in Human Subjects Research (ID 17464)	Incomplete	-	Start
Massachusetts Institute of Technology (ID 1290)	Incomplete	-	Start

- Then, you can save "Completion Report" in the following image.

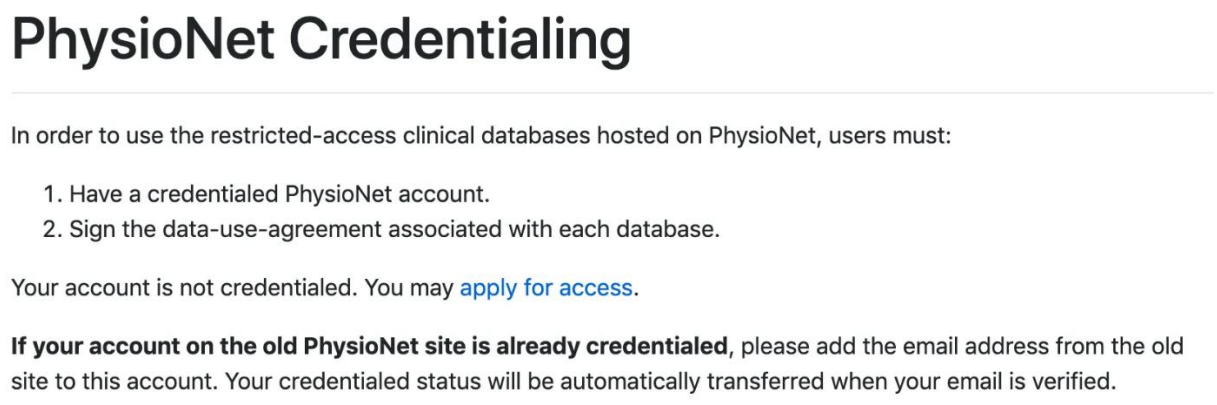


2. Step 2. Create an account in Physionet site https://physionet.org



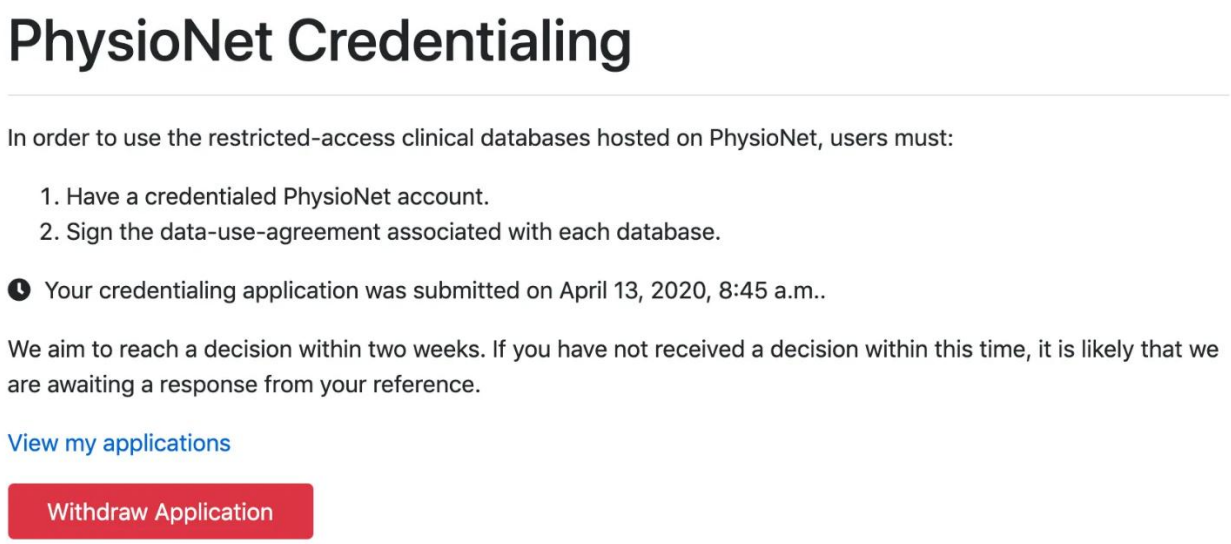
(optional) You can download the demo dataset for MIMIC-Ⅲ https://physionet.org/content/mimiciii-demo/1.4/

- After creating an account in Physionet site, you should do Physionet Credentialing. Click "Account-Settings-Credentialing" and then click the button, "apply for access"



- Fill in the required information and attach the pdf file that you got from CITI program.

- Fill the reference and click the button, “Submit application”



- Credentialing is processed within a maximum of two weeks, and you will be contacted by email.

3. Step 3. Download the dataset

1) MIMIC-Ⅱ Database: The files for this project are no longer available.

2) MIMIC-Ⅲ Database (6.2GB): https://physionet.org/content/mimiciii/1.4/

- Access the files

  - Download the ZIP file (6.2 GB)
  - Request access to the files using the Google Cloud Storage Browser. Login with a Google account is required.
  - Request access using Google BigQuery.
  - Download the files using your terminal: wget -r -N -c -np --user "username"--ask-password https://physionet.org/files/mimiciii/1.4/

- Data description

MIMIC-III is a relational database consisting of 26 tables. Tables are linked by identifiers which usually have the suffix 'ID'. For example, SUBJECT_ID refers to a unique patient, HADM_ID refers to a unique admission to the hospital, and ICUSTAY_ID refers to a unique admission to an intensive care unit.

Charted events such as notes, laboratory tests, and fluid balance are stored in a series of 'events' tables. For example the OUTPUTEVENTS table contains all measurements related to output for a given patient, while the LABEVENTS table contains laboratory test results for a patient.

Tables prefixed with 'D_' are dictionary tables and provide definitions for identifiers. For example, every row of CHARTEVENTS is associated with a single ITEMID which represents the concept measured, but it does not contain the actual name of the measurement. By joining CHARTEVENTS and D_ITEMS on ITEMID, it is possible to identify the concept represented by a given ITEMID.

Developing the MIMIC data model involved balancing simplicity of interpretation against closeness to ground truth. As such, the model is a reflection of underlying data sources, modified over iterations of the MIMIC database in response to user feedback. Care has been taken to avoid making assumptions about the underlying data when carrying out transformations, so MIMIC-III closely represents the raw hospital data.

Broadly speaking, five tables are used to define and track patient stays: ADMISSIONS; PATIENTS; ICUSTAYS; SERVICES; and TRANSFERS. Another five tables are dictionaries for cross-referencing codes against their respective definitions: D_CPT; D_ICD_DIAGNOSES; D_ICD_PROCEDURES; D_ITEMS; and D_LABITEMS. The remaining tables contain data associated with patient care, such as physiological measurements, caregiver observations, and billing information.

In some cases it would be possible to merge tables—for example, the D_ICD_PROCEDURES and CPTEVENTS tables both contain detail relating to procedures and could be combined—but our approach is to keep the tables independent for clarity, since the data sources are significantly different. Rather than combining the tables within MIMIC data model, we suggest researchers develop database views and transforms as appropriate.

3) MIMIC-IV Database (9.9GB): https://physionet.org/content/mimiciv/3.1/

- **Access the files**

- Download the ZIP file (9.8 GB)
- Request access using Google BigQuery.
- Download the files using your terminal: wget -r -N -c -np --user ayeyoung --ask-password https://physionet.org/files/mimiciv/3.1/

- **Data description**

- The *hosp* module contains detailed data regarding 546,028 unique hospitalizations for 223,452 unique individuals. Measurements in the *hosp* module are predominantly recorded during the hospital stay, though some tables include data from outside an admitted hospital stay as well (e.g. outpatient or emergency department laboratory tests in *labevents*). Patient demographics (*patients*), hospitalizations (*admissions*), and intra-hospital transfers (*transfers*) are recorded in the *hosp* module. Other information in the *hosp* module includes laboratory measurements (*labevents*, *d_labitems*), microbiology cultures (*microbiologyevents*, *d_micro*), provider orders (*poe*, *poe_detail*), medication administration (*emar*, *emar_detail*), medication prescription (*prescriptions*, *pharmacy*), hospital billing information (*diagnoses_icd*, *d_icd_diagnoses*, *procedures_icd*, *d_icd_procedures*, *hcpcsevents*, *d_hcpcs*, *drgcodes*), online medical record data (*omr*), and service related information (*services*).
- Provider information is available in the *provider* table. The provider_id column is a deidentified character string which uniquely represents a single care provider. As `provider_id` is used in different contexts across the module, a prefix is usually present in data tables to contextualize how the provider relates to the event. For example, the provider who admits the patient to the hospital is documented in the *admissions* table as `subject_id`. All columns which have a suffix of `provider_id` may be linked to the *provider* table.
- Deidentified dates and aligning stays to year groups
- All dates in MIMIC-IV have been deidentified by shifting the dates into a future time period between 2100 - 2200. This shift is done independently for each patient, and as a result two patients admitted in the deidentified year 2120 cannot be assumed to be admitted in the same year. To provide information about the original time period when a patient was admitted, the *patients* table provides a set of columns with the "anchor_" prefix. The `anchor_year` column is a deidentified year occurring sometime between 2100 - 2200, and the `anchor_year_group` column is one of the following values: "2008 - 2010", "2011 - 2013", "2014 - 2016", "2017 - 2019", and "2020 - 2022". These pieces of information allow researchers to infer the approximate year a patient received care. For example, if a patient's `anchor_year` is 2158, and their `anchor_year_group` is 2011 - 2013, then any hospitalizations for the patient occurring in the year 2158 actually occurred sometime between 2011 - 2013. In order to minimize accidental release of information, only a single `anchor_year` is provided per subject_id. Consequently, individual stays must be aligned to the anchor year using the respective date (e.g. `admittime`).

Finally, the `anchor_age` provides the patient age in the given `anchor_year`. If the patient was over 89 in the `anchor_year`, this `anchor_age` has been set to 91 (i.e. all patients over 89 have been grouped together into a single group with value 91, regardless of what their real age was).

- Out of hospital linkage of date of death

- Date of death is available within the *dod* column of the *patients* table. Date of death is derived from hospital records and state records. If both exist, hospital records take precedence. State records were matched using a custom rule based linkage algorithm based on name, date of birth, and social security number. State and hospital records for date of death were collected two years after the last patient discharge in MIMIC-IV, which should limit the impact of reporting delays in date of death.
- Dates of death occurring more than one year after hospital discharge are censored as a part of the deidentification process. As a result, the maximum time of follow up for each patient is exactly one year after their last hospital discharge. For example, if a patient's last hospital discharge occurs on 2150-01-01, then the last possible date of death for the patient is 2151-01-01. If the individual died on or before 2151-01-01, and it was captured in either state or hospital death records, then the *dod* column will contain the deidentified date of death. If the individual survived for at least one year after their last hospital discharge, then the *dod* column will have a NULL value.
- *icu*
- The *icu* module contains data sourced from the clinical information system known as MetaVision (iMDSoft). MetaVision tables were denormalized to create a star schema where the *icustays* and *d_items* tables link to a set of data tables all suffixed with "events". Data documented in the *icu* module includes intravenous and fluid inputs (*inputevents*), ingredients for the aforementioned inputs (*ingredientevents),* patient outputs (*outputevents*), procedures (*procedureevents*), information documented as a date or time (*datetimeevents*), and other charted information (*chartevents*). All events tables contain a *stay_id* column allowing identification of the associated ICU patient in *icustays*, and an *itemid* column allowing identification of the concept documented in *d_items*. Additionally, the *caregiver* table contains `caregiver_id`, a deidentified integer representing the care provider who documented data into the system. All events tables (*chartevents*, *datetimeevents*, *ingredientevents*, *inputevents*, *outputevents*, *procedureevents*) have a `caregiver_id` column which links to the *caregiver* table.
- The *icu* module contains a total of 94,458 ICU stays for 65,366 unique individuals as of MIMIC-IV v3.0. An ICU stay is defined as a contiguous sequence of transfers within a unit of the hospital classified as an ICU, and the *icustays* table is derived from the *transfers* table. During the creation of the *icustays* table, consecutive transfers within an ICU were merged into the same stay_id for analytical convenience, as these transfers are often bed number changes. Importantly, non-consecutive ICU stays remain as unique `stay_id` in the *icustays* table. In some cases, these could be considered the "same" ICU stay as the patient was transferred out for a planned procedure. In other cases, these are unanticipated readmissions to the ICU. As there was no systematically perfect method to differentiate these cases, we did not attempt to merge non-consecutive `stay_id`, and it is up to the investigator to appropriately handle these cases.

4) MIMIC-CXR Database (4.7TB): http://physionet.org/content/mimic-cxr/2.1.0/files/p10/#files-panel

- **Access the files**

- Download the files using your terminal: wget -r -N -c -np --user ayeyoung --ask-password https://physionet.org/files/mimic-cxr/2.1.0/

- **Data Description**

- A set of 10 folders (p10 - p19), each with ~6,500 sub-folders. Sub-folders are named according to the patient identifier, and contain free-text reports and DICOM files for all studies for that patient
- cxr-record-list.csv.gz - a compressed file providing the link between an image, its corresponding study identifier, and its corresponding patient identifier
- cxr-study-list.csv.gz - a compressed file providing a link between anonymous study and patient identifiers
- cxr-provider-list.csv.gz - a compressed file providing the ordering, attending, and resident provider associated with the given radiology study
- mimic-cxr-reports.tar.gz - for convenience, all free-text reports have been compressed in a single archive file

Folder structure

Free-text reports and images are provided in individual folders. An example of the folder structure for a single patient's images is as follows:

```
files
└── p10
    └── p10000032
        ├── s50414267
        │   ├── 02aa804e-bde0afdd-112c0b34-7bc16630-4e384014.dcm
        │   └── 174413ec-4ec4c1f7-34ea26b7-c5f994f8-79ef1962.dcm
        ├── s50414267.txt
        ├── s53189527
        │   ├── 2a2277a9-b0ded155-c0de8eb9-c124d10e-82c5caab.dcm
        │   └── e084de3b-be89b11e-20fe3f9f-9c8d8dfe-4cfd202c.dcm
        ├── s53189527.txt
        ├── s53911762
        │   ├── 68b5c4b1-227d0485-9cc38c3f-7b84ab51-4b472714.dcm
```

```
        │   └── fffabebf-74fd3a1f-673b6b41-96ec0ac9-2ab69818.dcm
        ├── s53911762.txt
        ├── s56699142
        │   └── ea030e7a-2e3b1346-bc518786-7a8fd698-f673b44c.dcm
        └── s56699142.txt
```

Above, we have a single patient, `p10000032`. Since the first three characters of the folder name are `p10`, the patient folder is in the `p10/` folder. This patient has four radiographic studies: `s50414267`, `s53189527`, `s53911762`, and `s56699142`. These study identifiers are completely random, and their order has no implications for the chronological order of the actual studies. Each study has two chest x-rays associated with it, except `s56699142`, which only has one study.

Metadata files

The cxr-record-list.csv.gz file lists all DICOM images available in the dataset. It also provides a mapping of these DICOM images to their corresponding anonymous study and subject identifier.

The cxr-study-list.csv.gz lists all studies available in the dataset, and provides a mapping of these anonymous study identifiers to the patient identifier.

The cxr-provider-list.csv.gz file has four columns, namely study_id, ordering, attending, and provider. These provide deidentified identifiers for the providers of care. All studies have an associated ordering and attending provider, and 83,021 of the studies have a resident radiologist associated with the study. These provider IDs represent the same individuals as the provider identifiers in MIMIC-IV v2.2 and later versions.

The mimic-cxr-reports.zip file is a compressed archive containing all text reports in the dataset. While the text reports are available within each patient folder, users may be interested in examining only the text without the images. This archive file is intended to make this process simpler.

5) MIMIC-IV ED: https://physionet.org/content/mimic-iv-ed/2.2/

- **Access the files**

- Download the ZIP file (116.3 MB)
- Request access using Google BigQuery.
- Download the files using your terminal: wget -r -N -c -np --user ayeyoung --ask-password https://physionet.org/files/mimic-iv-ed/2.2/

- **Data description**

MIMIC-IV-ED is composed of a single patient tracking table, *edstays*, and five data tables: *diagnosis*, *medrecon*, *pyxis*, *triage*, and *vitalsign*.

- **edstays**

Patient stays are tracked in the *edstays* table. Each row of the *edstays* table has a unique `stay_id`, which represents a unique patient stay in the ED. The *edstays* table contains the following columns: `subject_id`, `hadm_id`, `stay_id`, `intime`, `outtime`, `gender`, `race`, `arrival_transport`, and `disposition`. The `intime` indicates the time at which the patient was admitted to the ED, and the `outtime` indicates the time at which the patient was discharged from the ED. If the patient was admitted to the hospital following their ED stay, the `hadm_id` column will be populated with an identifier representing their hospital stay. `hadm_id` can be linked with the `hadm_id` in MIMIC-IV to obtain further detail about the patient's hospital stay. Each individual is assigned a unique `subject_id`, and patients with multiple ED stays will have the same `subject_id` across stays in the *edstays* table. Patient demographics including race and gender are provided in the respective columns. The mechanism of patient admission is provided in `arrival_transport`, and is coded into one five values: AMBULANCE, HELICOPTER, WALK IN, UNKNOWN, or OTHER. Patient discharge location is coded in `disposition`, and is one of eight values: ADMITTED, ELOPED, EXPIRED, HOME, LEFT AGAINST MEDICAL ADVICE, LEFT WITHOUT BEING SEEN, TRANSFER, and OTHER.

Note that `subject_id` can be used to link MIMIC-IV-ED with MIMIC-IV to obtain additional information regarding individuals, e.g. age. `subject_id` can also be linked with the PatientID DICOM attribute in MIMIC-CXR to obtain chest x-rays for patients if they are available [3].

- **diagnosis**

The *diagnosis* table provides coded diagnoses for the patient in the International Classification of Diseases (ICD) Ninth or Tenth revision (ICD-9 or ICD-10). These diagnoses are determined by trained coders after discharge from the emergency department and are used for billing purposes. There are six columns in the *diagnosis* table: `subject_id`, `stay_id`, `seq_num`, `icd_code`, `icd_version`, and `icd_title`. A maximum of 9 ICD codes are available for a single stay. The `seq_num` column provides a pseudo-order for the ICD codes, with a value of 1 usually

indicating highest relevance and a value of 9 indicating least relevance.
The `icd_code` provides the coded representation of the diagnosis using the ICD ontology, the `icd_version` column is either 9 or 10 indicating whether the ontology used is ICD-9 or ICD-10, and the `icd_title` column provides the textual description of the ICD code.

It is important to note that the billed diagnoses in the *diagnosis* table are exclusively related to the patient's emergency department stay. If the patient is subsequently admitted to the hospital, they will have a separate set of billed diagnoses for their hospital stay, which are not recorded in this table. See the usage notes for details regarding linking MIMIC-IV-ED to MIMIC-IV, which would facilitate comparison of the billed ED diagnoses with billed hospital diagnoses.

- **medrecon**

The *medrecon* table provides medicine reconciliation for each patient, that is a list of the medications which the patient was taking prior to their ED stay. The *medrecon* table has nine columns: `subject_id`, `stay_id`, `charttime`, `name`, `gsn`, `ndc`, `etc_rn`, `etccode`, and `etcdescription`. The `charttime` provides the date and time at which the medicine reconciliation was documented. The `name` column provides a text description of the medicine, the `gsn` column provides the Generic Sequence Number (GSN), and the `ndc` column provides the National Drug Code (NDC). Note a `gsn` or an `ndc` of 0 indicates that the value is missing. Columns prefixed with `etc` provide an ontology for grouping together drugs of a similar class. Note that as a medicine can be classified in multiple groups in the ontology, there may be more than one row for a single medication. For example, the medication Adderal is (1) a CNS stimulant, (2) an Attention Deficit-Hyperactivity Therapy, and (3) a narcolepsy therapy. As a result, patients taking adderal prior to their admission will have three rows in the *medrecon* table, delineated by the sequential monotonically increasing integer `etc_rn`. The `etccode` provides the coded form of the ontology group, and the `etcdescription` proides the textual description of the ontology group.

- **pyxis**

The *pyxis* table provides dispensation information for medications provided by the BD Pyxis MedStation, an automated medication dispensing system present in the ED [7].
The *pyxis* table has nine columns: `subject_id`, `stay_id`, `charttime`, `med_rn`, `name`, `gsn_rn`, and `gsn`. The `charttime` provides the time at which the medication was dispensed. If multiple medications were dispensed at the same time, the `med_rn` column delineates these medications. The `name` column provides a textual description of the medication dispensed, and may additionally contain auxiliary information such as the formulation. The `gsn` column provides the Generic Sequence Number (GSN) if available, and `gsn_rn` delineates multiple GSN values associated with the same medication. Note that a `gsn` of 0 indicates that the GSN is missing. Not all medications are dispensed by the Pyxis MedStation, and as a result not all medications are recorded in the *pyxis* table. For example, large fluid volumes (such as those used for resuscitation) are not present in this table.

- **triage**

The *triage* table provide information collected from the patient at the time of triage. All patients who present to the ED are immediately triaged, a process which involves assessing their health status and ascertaining the reason for their visit. The *triage* table has eleven

columns: subject_id, stay_id, temperature, heartrate, resprate, o2sat, sbp, dbp, pain, acuity, and chiefcomplaint. Vital signs collected at triage include patient temperate (temperature), heart rate (heartrate), respiratory rate (resprate), oxygen saturation (o2sat), systolic blood pressure (sbp), and diastolic blood pressure (dbp). Although vital signs can be documented as free-text, the deidentification approach retained only numeric vital signs. A patient reported pain level is available in the pain column. The chiefcomplaint is a free-text field which contains the patient's reported reason for presenting to the ED. The chiefcomplaint field is usually a comma separated list of entries. PHI present in the chiefcomplaint field has been replaced by three underscores ("___"). Based upon the triage assessment, the care provider will assign an integer level of severity (acuity), where 1 indicates the highest severity and 5 indicates the lowest severity.

- **vitalsign**

The *vitalsign* table contains aperiodic vital signs documented for patients during their stay. The *vitalsign* table has eleven columns: subject_id, stay_id, charttime, temperature, heartrate, resprate, o2sat, sbp, dbp, rhythm, and pain. Vital signs in the *vitalsign* table are similar to those collected in the *triage* table. The rhythm column additionally provides the hearth rhythm for the patient. The charttime provides the time at which the vital signs were recorded.