Introduction to EHR and MIMIC Data

Anonymous Author(s)
Affiliation
Address
`email`

1 Introduction

An electronic health record (EHR) is a digital record of a patient's interactions with a healthcare system. In the last decade, the worldwide adoption of EHRs has skyrocketed. In the United States, the $19 billion Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 incentivized healthcare institutions to implement EHR systems capable of improving the quality, safety, and efficiency of patient care. As a result, EHR usage in the US has more than quadrupled since 2009. Among their many advantages, EHR systems support comprehensive healthcare documentation, enhanced data security, improved healthcare process management, and integrated medical billing. Additionally, interoperable EHR systems facilitate information exchange and streamlined communication across various healthcare providers involved in a patient's care.

Beyond their uses at the point of care, EHRs have simultaneously transformed biomedical research. In contrast to data collected from traditional clinical studies such as randomized controlled trials, EHR data is readily accessible and available for large and diverse populations. The data is also collected over long periods of time and contains detailed timestamped information on disease diagnoses, laboratory tests, medical procedures, prescriptions, and clinical notes. The depth and volume of EHR data has in turn created unprecedented opportunities for research, including the development of artificial intelligence (AI) systems for rapid and more inclusive clinical trials recruitment, the formation of international research networks focusing on real-time studies of the COVID-19 pandemic, and the introduction of real-world prediction models for sepsis that can identify the condition prior to symptom onset.

In spite of the exciting and expanding uses of EHR data, there are numerous known challenges that arise in its analysis. Fundamentally, EHR data is not designed for research purposes. It is therefore complex in its structure and prone to quality issues such as missing data, erroneous entries, and duplicate data. A substantial amount of work is required to process and transform EHR data before it can be safely and reliably used for a biomedical application.

In this tutorial We will be introducing some of the key data elements stored within an EHR system such as disease diagnoses, lab tests, medications and procedures and certain standard coding systems used to represent them. Further, we will briefly introduce the MIMIC-IV EHR database that is openly available to the research community.

2 Medical Coding Systems

2.0.1 Disease codes

Disease codes encode diseases, symptoms, and phenotypes into a unique code. ICD (International Classification of Diseases) codes [1] and Phecodes [2] are examples of disease codes. ICD codes are typically used in health care settings to code diagnoses made during a patient encounter. However, ICD codes are too detailed to be used for research purposes. Phecodes solves this problem by grouping relevant ICD codes into clinical meaningful phenotypes.

ICD Codes

The International Classification of Diseases (ICD) codes contain codes for diseases, symptoms, findings, and injuries that is maintained by the World Health Organization (WHO). ICD codes are used to record medical findings in a standardized format in EHRs and to track global morbidity and mortality. Different countries and different hospitals may use different versions of ICD codes. WHO periodically releases new versions of ICD codes, among which ICD-9 and ICD-10 are relevant to EHR data. The United States uses an extended version of ICD called the "clinical modification" (CM), e.g. ICD-9-CM. ICD-10-CM has over five times the number of diagnosis codes as ICD-9-CM.

Phecodes

Closely related to the ICD codes, Phecodes are phenotyping codes that group various ICD codes into useful phenotypes. ICD codes are very detailed codes, which can be manually rolled up to phecodes [1, 2]. Phecodes version 1.2 condenses roughly 15500 ICD-9-CM codes and 90000 ICD-10-CM codes into 1867 phecodes. ICD codes and phecodes create a certain hierarchy[3] from general phenotype to detailed findings. An example of this hierarchy is shown in Figure 1.A complete ICD-Phecode system includes ICD codes, ICD strings, corresponding phecodes and phenotypes, and corresponding excluded phecodes range and phenotype. Table 1 shows an example of the ICD-Phecode system.

ICD-9	ICD-9 String	Phecode	Phenotype	Excl. Phecodes	Excl. Phenotypes
250	Diabetes mellitus	250	Diabetes mellitus	249-250.99	DIABETES
250.1	Diabetes with ketoacidosis	250	Diabetes mellitus	249-250.99	DIABETES
250.33	Diabetes mellitus with other coma...	250.1	Type 1 diabetes	249-250.99	DIABETES
250.13	type I diabetes mellitus [juvenile type]...	250.11	Type 1 diabetes with ketoacidosis	249-250.99	DIABETES

Table 1: An example of ICD-Phecode system. ICD codes are grouped by phecodes and phenotypes. Numbers after the decimal point of phecodes show certain hierarchies of phenotypes. For example, 250—250.1—250.11 represents the hierarchy of Diabetes mellitus—Type 1 diabetes—Type 1 diabetes with ketoacidosis.

2.0.2 Medication codes

Medication codes encode drug products and their usage into unique codes. Typical medication coding systems include RxNorm[4] and National Drug Code (NDC)[5]. Health care systems can also have their own local medication codes.

RxNorm

RxNorm, produced by The National Library of Medicine (NLM), provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First Databank, Micromedex, and Gold Standard Drug Database. RxNorm provides a set of codes (RxCUI) for clinical drugs, which are the combination of active ingredients, dose form, and strength of a drug. For example, the

[1] https://www.who.int/standards/classifications/classification-of-diseases

[2] https://phewascatalog.org/phecodes

[3] The figure is obtained from a visualization tool for ICD codes and phecodes. https://hmsrsc.aws.hms.harvard.edu/content/89/

[4] https://www.nlm.nih.gov/research/umls/rxnorm/index.html

[5] https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory

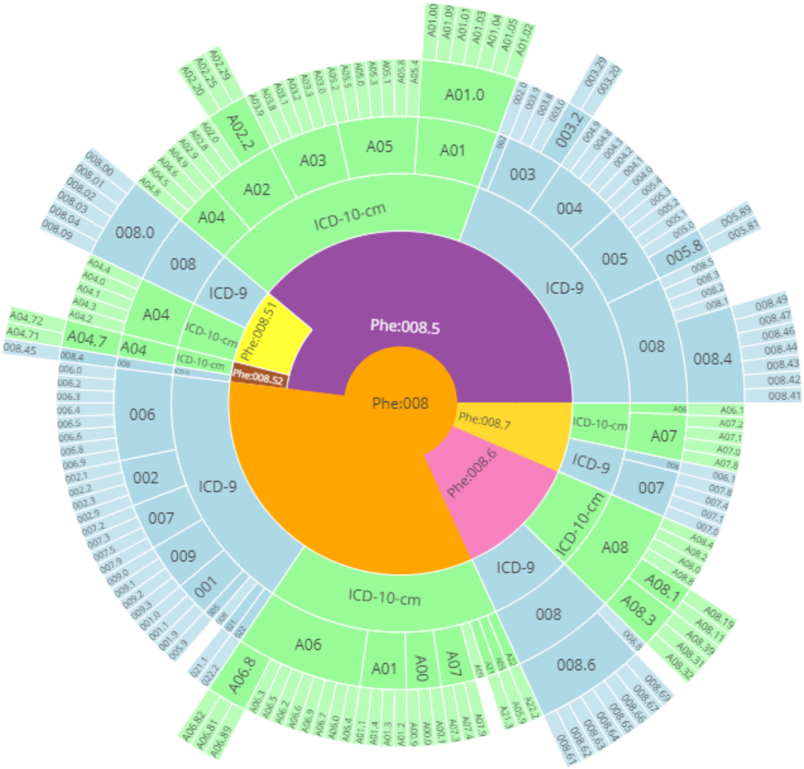


Figure 1: An example of the ICD-Phecode hierarchy. The most general concept (root) is encoded as an integer in the phecode system, e.g. 008. The second level is encoded as a decimal with the same integer part as the root and one digit after the decimal point, while the third level has two digits after the decimal point. Phecode groups multiple ICD codes that falls under the same broader concept.

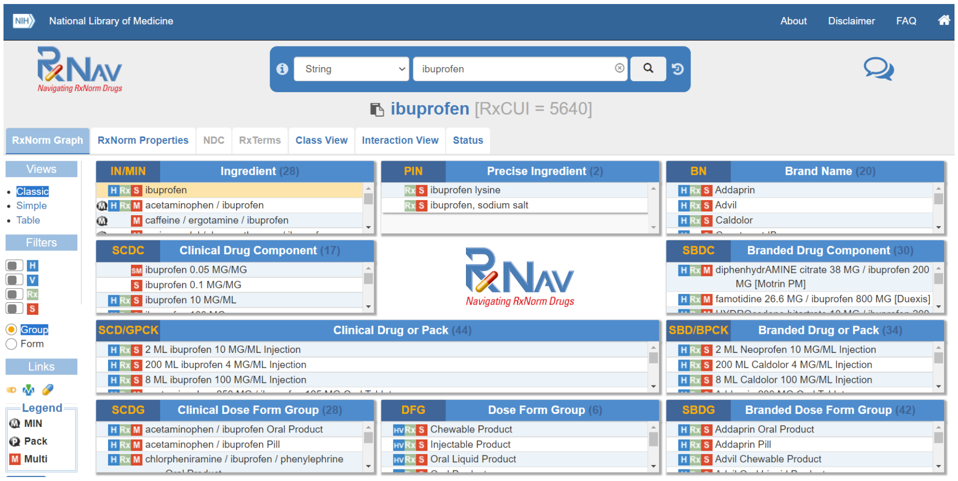
RxCUI for *ciprofloxacin 500 mg 24-hour extended-release tablet* (the generic name for Cipro XR 500 mg) is RX10359383, regardless of brand or packaging. RxNorm is also a hierarchical code, with relationships like "has_precise_ingredient", "has_ingredient", "has_part", and "consists_of".

RxNorm can be also navigated through RxNav, which offers a visualized browser for each RxNorm code. Examples of RxNav is shown in Figure 2.

NDC

The National Drug Code (NDC) is a unique product identifier used in the United States for drugs intended for human use. It's published by U.S. Food and Drug Administration (FDA). Through NDC, drugs are identified and reported using a unique, three-segment number which serves as the FDA's identifier for drugs.

The first segment, the labeler code, is 4, 5 or 6 digits long and identifies the labeler such as the drug manufacturer, repackager, or distributer. The second segment, the product code, is 3 or 4 digits long and identifies a specific strength, dosage form, and formulation for a particular firm. The third segment, the package code, is 1 or 2 digits long and identifies package forms and sizes. For example, the product NDC of *Ibuprofen* produced by Granules India Limited is 62207-0356.



ibuprofen [RxCUI = 5640]

IN/MIN Ingredient (28)
- H Rx S ibuprofen
- H Rx M acetaminophen / ibuprofen
- M caffeine / ergotamine / ibuprofen

PIN Precise Ingredient (2)
- Rx S ibuprofen lysine
- Rx S ibuprofen, sodium salt

BN Brand Name (20)
- H Rx S Addaprin
- H Rx S Advil
- H Rx S Caldolor

SCDC Clinical Drug Component (17)
- SM ibuprofen 0.05 MG/MG
- S ibuprofen 0.1 MG/MG
- H Rx S ibuprofen 10 MG/ML

SBDC Branded Drug Component (30)
- H Rx M diphenhydrAMINE citrate 38 MG / ibuprofen 200 MG [Motrin PM]
- H Rx M famotidine 26.6 MG / ibuprofen 800 MG [Duexis]

SCD/GPCK Clinical Drug or Pack (44)
- H Rx S 2 ML ibuprofen 10 MG/ML Injection
- H Rx S 200 ML ibuprofen 4 MG/ML Injection
- H Rx S 8 ML ibuprofen 100 MG/ML Injection

SBD/BPCK Branded Drug or Pack (34)
- H Rx S 2 ML Neoprofen 10 MG/ML Injection
- H Rx S 200 ML Caldolor 4 MG/ML Injection
- H Rx S 8 ML Caldolor 100 MG/ML Injection

SCDG Clinical Dose Form Group (28)
- H Rx M acetaminophen / ibuprofen Oral Product
- H Rx M acetaminophen / ibuprofen Pill
- H Rx M chlorpheniramine / ibuprofen / phenylephrine Oral Product

DFG Dose Form Group (6)
- HV Rx S Chewable Product
- HV Rx S Injectable Product
- HV Rx S Oral Liquid Product

SBDG Branded Dose Form Group (42)
- H Rx S Addaprin Oral Product
- H Rx S Addaprin Pill
- H Rx S Advil Chewable Product

Figure 2: An example of RxNav for Ibuprofen. The browser shows properties like ingredients and clinical drug components of Ibuprofen. RxNav provides an interactive tool for viewing RxNorm and its constituents.

2.0.3 Procedure codes

Procedure codes identify specific surgical, medical, or diagnostic procedures. Typical procedure codes include Current Procedural Terminology (CPT) and ICD-10 Procedure Coding System (ICD-10-PCS). Procedure codes are used in billing, EHR, and insurance.

CPT

The Current Procedural Terminology (CPT) code set is a procedural code set developed by the American Medical Association (AMA). The CPT code set describes medical, surgical, and diagnostic procedures for administrative, financial, and analytical purposes. There are three types of CPT codes. Category 1 are 5-digit codes that cover evaluation and management, anesthesiology, surgery, radiology, pathology and laboratory, and medicine. These are the most widely used CPT codes. Category 2 are clinical laboratory services and category 3 are emerging technologies, services, and procedures. Examples of CPT codes are listed in Table 2.

CPT	Name
3120F	12-LEAD ECG PERFORMED
4030F	LONG-TERM OXYGEN THERAPY PRESCRIBED
0575F	HIV RNA CONTROL PLAN OF CARE DOCD

Table 2: Examples of CPT codes.

ICD-10-PCS

The ICD-10 Procedure Coding System (ICD-10-PCS) is an international system of medical classification used for procedural coding. ICD-9-CM contains a procedure classification while ICD-10-CM does not. ICD-10-PCS is the procedure classification for ICD-10. Each ICD-10-PCS code consists of seven alphanumeric characters. The first character is the ’section’. The second through seventh characters mean different things in each section. Each character can be any of 34 possible values; the ten digits 0-9 and the 24 letters A-H, J-N and P-Z may be used in each character. The letters O and I are excluded to avoid confusion with the numbers 0 and 1. There are no decimals in ICD-10-PCS. Of the 72,081 codes in ICD-10-PCS, 62,022 are in the first section, "Medical and surgical". Detailed descriptions of the sections can be found in Wiki. For example, 07Q70ZZ is the code for *Repair Thorax Lymphatic, Open Approach.*

2.0.4 Laboratory codes

Laboratory codes identify medical laboratory observations. Logical Observation Identifier Names and Codes (LOINC) is the most widely-used terminology standard for coding laboratory data in US.

LOINC

The LOINC database provides a universal code system for reporting laboratory and other clinical observations. The database currently has over 71,000 observation terms. A LOINC term includes 6 parts: component, kind of property, time aspect, system, type of scale, and type of method. An example of a LOINC code is shown in Table 3.

LOINC has a hierarchical structure. It can be represented as a tree, where each leaf is the LOINC term and the parents are LOINC parts whose codes begin with LP. For example, LP14559-6 stands for *Bacteria*). A visualization of such structure can be browsed in `https://loinc.org/tree/`.

Code	Component	Property	Time aspect	System	Scale	Method
29463-7	Body weight	Mass	Pt	ˆ Patient	Qn	—

Table 3: Example of LOINC term *body weight*. Component indicates what is measured, evaluated, or observed. Property indicates characteristics of what is measured. Time aspect indicates interval of time over which the observation or measurement was made. Pt means point in time. System indicates context or specimen type within which the observation was made, e.g. blood or urine. Scale indicates the scale of measure. Qn means quantitative. Method indicates the procedure used to make the measurement or observation.

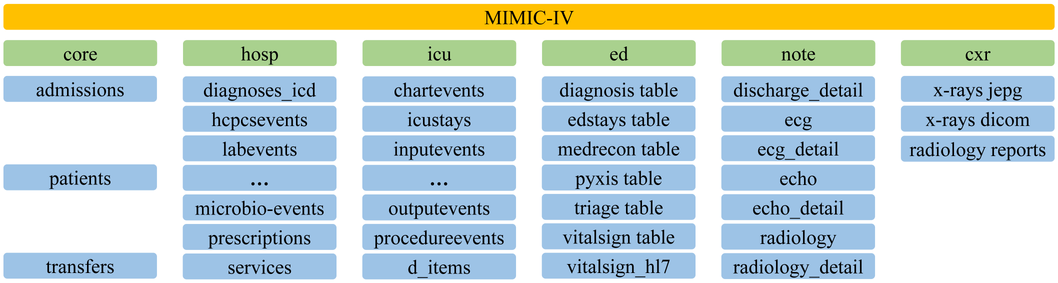
2.1 Aggregation of codified data

EHR data typically include four domains of codified data: diagnosis, procedures, lab measurements, and medications. Due to differential coding practices, the same clinical concepts might be represented by distinct clinical codes at different healthcare systems [3, 4]. For example, acute myocardial infarction (MI) of anterolateral wall and acute MI of the inferolateral wall are separate codes that describe the same concept of MI [5]. To reduce ambiguity and alleviate heterogeneity across different healthcare systems, the individual clinical codes are usually rolled to codes representing general concepts [5, 6]. As discussed previously, ICD codes are often aggregated into PheCodes using the ICD-to-PheCode mapping from PheWAS catalog (https://phewascatalog.org/phecodes). Procedure codes including CPT-4, HCPCS, ICD-9-PCS, ICD-10-PCS are grouped into clinical classification software (CCS) categories based on the CCS mapping (https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp). Medication codes are often aggregated and rolled up into ingredient level RxNorm codes. Laboratory measurements can be aggregated into LOINC codes and further grouped into higher level LOINC Parts codes (LP codes) to reflect broader laboratory code concepts by leveraging the LOINC Multiaxial Hierarchy [7].

3 Introduction to MIMIC

3.1 Why analyze MIMIC?

MIMIC is a large, publicly-available database containing deidentified health-related data associated with a large number patients who stayed in critical care units of the Beth Israel Deaconess Medical Center. The database includes information such as demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality information (including post-hospital discharge). The MIMIC database includes three versions: MIMIC-II, MIMIC III, and MIMIC-IV. The new generation of data has a wider source (from 2001-2019) and better organization . The available data tables in MIMIC-IV are shown in 3.



MIMIC-IV					
core	hosp	icu	ed	note	cxr
admissions	diagnoses_icd	chartevents	diagnosis table	discharge_detail	x-rays jepg
	hcpcsevents	icustays	edstays table	ecg	x-rays dicom
	labevents	inputevents	medrecon table	ecg_detail	radiology reports
patients	...	...	pyxis table	echo	
	microbio-events	outputevents	triage table	echo_detail	
	prescriptions	procedureevents	vitalsign table	radiology	
transfers	services	d_items	vitalsign_hl7	radiology_detail	

Figure 3: Available data tables. of MIMIC-IV.

MIMIC has supported a diverse range of analytic studies spanning epidemiology [8][9][10][11] [12] [13] , clinical decision-rule improvement, and electronic tool (e.g. the electronic medical record management database) development [14] [15] [16] [17]MIMIC is notable for three factors: it is freely available to researchers worldwide; it encompasses a diverse and very large population of ICU patients; and it contains highly granular data, including vital signs, laboratory results, and medications.

This paper describe MIMIC-IV data and can be accessed at the following address.

3.1.1 Data contained in the MIMIC

3.1.1.1 Overview

MIMIC is a multi-modal dataset containing many kinds of medical data recorded on patients in the ICU. The data is present in many different forms and is collected in many ways. Specifically, MIMIC is composed of tabular data, free text data, medical image data, waveform data. Echo reports, electrocardiogram reports, and radiology reports are available for both inpatient and outpatient stays. Almost all record sets include a waveform record contain digitized signals and a “numerics” record containing time series of periodic measurements, each presenting a quasi-continuous recording of vital signs of a single patient throughout an ICU stay (typically a few days, but many are several weeks in duration).

Tabular data: is the most common data type in MIMIC. Tabular data contains structured information. These information usually comes from structured text in the medical record, the electronic medical record system of the hospital, or the automatic recording of the medicine instruments in ICU. There are patient tables and dictionary tables. The patient table records various data of many patients, and the dictionary table records the medical information of the characteristic fields in the patient table.

For example, patient laboratory information, such as white blood cells and red blood cells, are recorded in the labevents file. Each line is a record for a specific patient ID, specific visit ID, specific test item ID and specific time node charttime. If it is a record in ICU, it may also be related to icutayid, which is the identifier of the patient when entering the ICU.

Free text data: ultrasound reports, ECG reports and radiology reports for inpatients and outpatients. The content is the natural language written by doctors, which contains a large number of medical terms.

Image data: the latest version of MIMIC database - MIMIC-IV has X-ray chest films in two directions (with corresponding image reports at the same time).

Waveform data: waveform records include digital signals (usually including ECG, ABP, respiration and PPG, and other common signals) and "numerics" records, which contain time series of periodic measurements. Each record presents a quasi continuous record of individual patient's vital signs during ICU hospitalization (usually several days, but many records last for several weeks).

3.1.1.2 Table columns

Tabular data makes up a majority of MIMIC and consist of laboratory data sheet, ICU data sheet, etc. The laboratory data sheet records the laboratory test items, test results and other data in the form of cross-sectional (2-dimensional) data. Laboratory test items include blood test, urine test, etc. Each row of tabular data is an instance, and each column is a variable. Table 4 provides an example of MIMIC tabular data of simulated data based on the patient table. The table contains an instance and six variables, that is, one row with six columns.

Table 4: An example of mimic tabular data of simulated data based on the patient table. subject_ID denotes the identifier of the patient. anchor_age, anchor_Year denotes the age after time conversion, and dod denotes death

subject_id	gender	anchor_age	anchor_year	anchor_year_group	dod
29463-7	M	100	45	2	True

Columns within a table typically contain identifier, storage time, and variables specific to the table. For example, blood test results will appear in the laboratory table and urine output dose will appear in the ICU table. The unique fields of each table are different, which can be retrieved in the following documents. For details about columns, please refer to the following links.

Tables pre-fixed with "D_" are dictionaries and provide definitions for identifiers and variables. For example, every row of OUTPUTEVENTS is associated with a single ITEMID which represents the concept measured, but it does not contain the actual name of the drug. By joining OUTPUTEVENTS and D_ITEMS on ITEMID, it is possible to identify what concept a given ITEMID represents. As shown in table5 ,an Example of mimic tabular data of simulated data based on the patient table.

Table 5: An example of mimic tabular data of simulated data based on the d_labitems table. itemid denotes the identifier of the laboratory test, label denotes the project name of the laboratory test, fluid denotes the type of test sample, category denotes the type of the test, loinc_code denotes the identifier of the test

itemid	label	fluid	category	loinc_code
42129	Absolute CD3 Count	Blood	Chemistry	8124-0

The specific variables contained in the table and the content and format of the data contained in the table are described in detail in this document.

3.1.2 Identifiers in MIMIC data

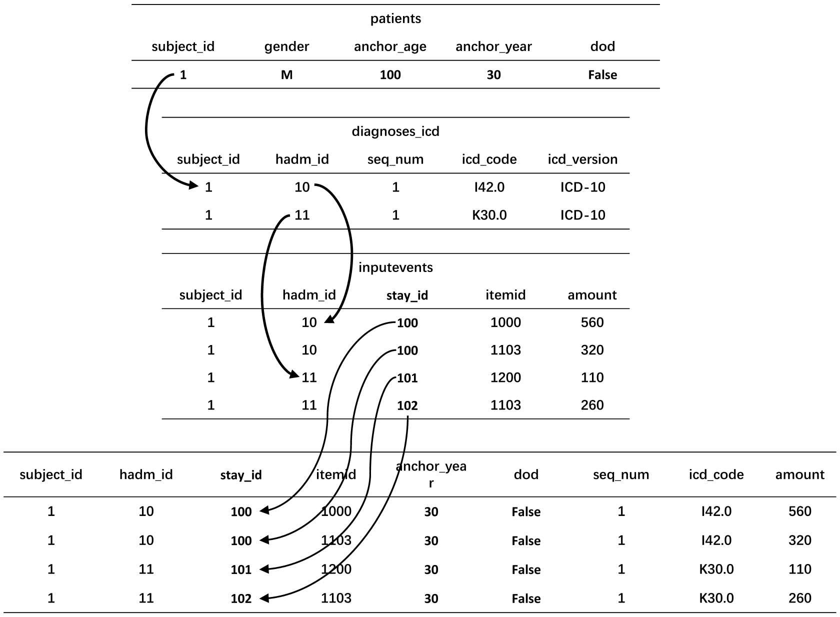
The tables are linked by identifiers which usually have the suffix **"ID"**. The ID identifier is described as follows:

- **SUBJECT_ID** Patient level. This data is constant for a patient.

- **HADM_ID** Hospital level. This data is constant for each admission.
- **ICUSTAY_ID** ICU level. This data is constant every time patients enter the ICU.

One exception is ROW_ID, which is simply a row identifier unique to that table.

Figure 4 shows an example of the correspondence between MIMIC ID identifiers.



patients

subject_id	gender	anchor_age	anchor_year	dod
1	M	100	30	False

diagnoses_icd

subject_id	hadm_id	seq_num	icd_code	icd_version
1	10	1	I42.0	ICD-10
1	11	1	K30.0	ICD-10

inputevents

subject_id	hadm_id	stay_id	itemid	amount
1	10	100	1000	560
1	10	100	1103	320
1	11	101	1200	110
1	11	102	1103	260

subject_id	hadm_id	stay_id	itemid	anchor_year	dod	seq_num	icd_code	amount
1	10	100	1000	30	False	1	I42.0	560
1	10	100	1103	30	False	1	I42.0	320
1	11	101	1200	30	False	1	K30.0	110
1	11	102	1103	30	False	1	K30.0	260

Figure 4: An example of the correspondence between MIMIC ID identifiers

Patient identifiers Each patient has a unique subject_ID. If a patient is admitted multiple times and have multiple hadms_ID, one admission may have multiple access to ICU and multiple ICDs_id A HADM_ID may correspond to multiple icustays_id A HADM_ID is usually used the first icustay_ID corresponding to Carry out relevant research

Medical concept correspondence The dictionary table is used to query a specific detail, for example:the query of white blood cell data of a patient (in the labevents table).

First, you need to find the corresponding three IDS (subject_id, hadm_id, icd_id), and then find the item of leukocyte in the laboratory examination code (d_labitems), and then look it up in the labels table.

Correspondence between patients and features The patient number, medical record number, and ICU number are used as the joint primary key to determine the patient. The item identifier is item_ID, for example, the item identifier corresponding to this item can be in D_Labitems which can be found in the dictionary Record time and storage time. The storage time of corresponding items Recording time can be used to filter specific time window (for example, data within 24 hours of entering ICU) The difference between the previous icustays enrollment time and the current measured time is used to determine the study cohort.

3.2 Tabular Data Cleaning

The MIMIC data was not collected for research purposes, data sets contain a lot of noisy, incomplete, and even inconsistent data. Data sets must be preprocessed before data analysis to improve data quality.

Some preparatory work should be done before data pretreatment, and unified file establishment and storage, naming rules should be followed, so as to find and reproduce others later.

Data preprocessing includes data cleaning, data integration, data conversion and data subtraction. Data cleaning refers to eliminating noise and correcting inconsistent errors in data, including incomplete, erroroneous, and duplicate data. Data integration refers to combining data from multiple data sources to form a complete data set. Data conversion refers to the conversion of data in one format to data in another format. Data reduction refers to the elimination of redundant data by deleting redundant features or clustering, such as data discretization.

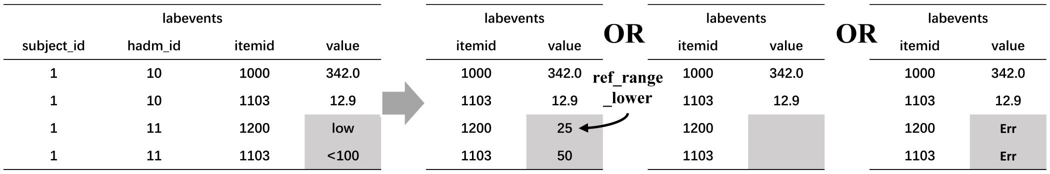
Data preprocessing can help improve the quality of data, which in turn helps improve the effectiveness and accuracy of the medical data mining process.

3.2.1 Exploratory data analysis

Before data analysis, you must understand your goal, that is, what problems need to be solved through mining. After you know your goals or problems to be solved, first back up the data, keep a copy of the original data and never change it. The next step is to examine the data, such as observing the number of samples, dimension size, missing condition and feature type. Finally, clean the data to provide a basis for data analysis.

3.2.2 Confirm legitimacy

In the process of inputting a few data, the data format is not uniform due to format error, format confusion or operator input error, that is, illegal values appear. For example, the test results in the laboratory are usually floating-point numbers. However, string data of < 100 may appear in some data. In order to unify the data format, it is necessary to perform feature processing on the columns that should be numeric variables but are character type (or object type) after data acquisition. By referring to the variable name, that is, the medical entity referred to by the current columns, understand the type of data in the current columns and give corresponding processing, such as converting the string of < 100 to 0 or the average value of 50. Processing example of illegal value as shown in Figure 5.



labevents			
subject_id	hadm_id	itemid	value
1	10	1000	342.0
1	10	1103	12.9
1	11	1200	low
1	11	1103	<100

labevents	
itemid	value
1000	342.0
1103	12.9
1200	25
1103	50

labevents	
itemid	value
1000	342.0
1103	12.9
1200	
1103	

labevents	
itemid	value
1000	342.0
1103	12.9
1200	Err
1103	Err

Figure 5: Processing example of illegal value

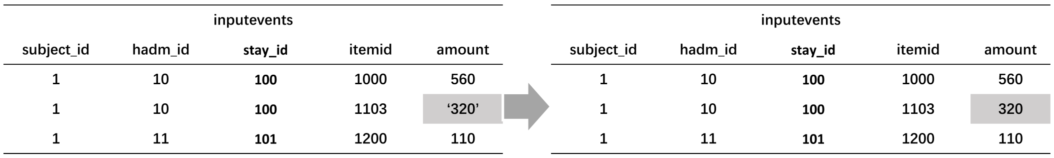
3.2.3 Unified variable type

In the process of data collection, there may be problems such as inconsistent multi-source data format, damaged database integrity or staff errors. There are inconsistent data formats of the same field (feature) in the data, such as string data in the age field (generally int value) or laboratory inspection result field (generally float value). Another example is the occurrence of integer data in a time record (usually string value). In order to avoid the above situation, we need to review and unify the format of data.

First of all, we need to be familiar with the format of each features in the data. For the MIMIC data set, we can find the format of the corresponding feature in the document. For the features that cannot be found, we can observe the format of most of the data to determine the format. Then, the data that does not belong to this type in this feature is processed by code conversion, manual conversion or direct deletion. Processing example of variable type unification as shown in Figure 6.

3.2.4 Code Rollup

Codes are rolled up to group low-level codes into high-level ones to reduce the total amount of codes in the dataset. In section 2.1, we have discussed fundamental concepts and methods of code



inputevents				
subject_id	hadm_id	stay_id	itemid	amount
1	10	100	1000	560
1	10	100	1103	'320'
1	11	101	1200	110

inputevents				
subject_id	hadm_id	stay_id	itemid	amount
1	10	100	1000	560
1	10	100	1103	320
1	11	101	1200	110

Figure 6: Processing example of variable type unification

aggregation. In MIMIC, there are HCPC (CPT), ICD-9, ICD-10-PCS, ICD-10-M, NDC, and some codes defined by MIMIC itself. According to section 2.1, we can convert procedure codes, in- cluding HCPC, ICD-9, ICD-10-PCS, to CCS codes; Diagnosis codes, including ICD-9 and ICD-10-CM, can be grouped to PheCodes; NDC codes can be aggregated to RxNorm. In this way, we are able to compress more than 48,000 codes into about 1,700 ones.

The process of code aggregation is straightforward: we create mapping tables and map the codes in the dataset.

Mapping tables are derived from some public resources. For example, we can download some tables mapping ICDs to PheCodes from PheWAS[6]. They provide tables in the format shown in Table 6.

Table 6: Mapping ICDs to PheCodes

ICD9	ICD9 String	PheCode	Phenotype	Excl. Phecodes	......
003.0	Salmonella gastroenteritis	008.5	Bacterial enteritis	001-009.99	......

r

In this case, we only need information in columns **ICD9** and **Phecode**, so we keep these two columns and remove all the others to create the mapping table. We can also remove those rows with codes that do no appear in our dataset.

In the cleaning program, we load these mapping tables as dictionaries. With these dictionaries, the original clinical codes in the dataset can be mapped into desired healthcare systems directly.

Some codes in the dataset cannot be aggregated with the mapping tables. We can leave these codes untouched in our dataset or simply remove them.

3.2.5 Frequency Filtering

Frequency filtering is to discard codes with low frequency in the dataset. This procedure is not always necessary, but if the frequency of some codes is too low to provide precise information about diseases or treatments, removing these codes from your research can be beneficial. Usually, we set the threshold of filtering according to experience. In MIMIC, we can set the threshold as 1,000, which means we discard all codes with a frequency lower than 1,000.

The concept of frequency filtering is intelligible, but the operation can be tricky. To filter low-frequency codes and reserve high-frequency ones, we have to count the frequency of all codes and record high-frequency codes in a dictionary in advance. In the following procedures, we only deal with codes in this dictionary and neglect the others. In the dictionary, you can record not only codes but also some extra information such as statistics, units of measurement, and labels. Table 7 provides a simple example of dictionary format.

3.2.6 Variables encoding

Categorical variables, also called nominal variable, generally refer to two or more categories, but have no rank order.

[6] `https://phewascatalog.org/phecodes`

Table 7: An example of dictionary format

code	frequency	source_table	unit_of_measurement	label	description
51279	3278470	labevents	m/ul	Red Blood Cells	...
51006	3283231	labevents	mg/dl	Urea Nitrogen	...
51301	3283759	labevents	k/ul	White Blood Cells	...

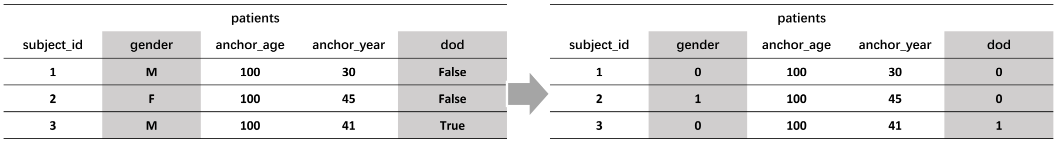
Numericalization of categorical variables. The easiest way to quantify a categorical variable is to enumerate all the values and use an integer map. For example, blood types are divided into A, B, AB, O, which can be directly converted into 0, 1, 2, 3.

However, this processing will cause problems. The mapped values here only represent the commodity categories. Because the commodity categories are out of order, the mapped values are ordered. If these values are involved in modeling, the model will consider the mapped values of the commodity categories to be ordered. Therefore, it is inappropriate to directly map commodity categories to ordered values.

These categories have two principles:

- Different categories must be mutually exclusive, each research object can only be classified into one category
- All research objects belong to each other and cannot be left behind. For example, gender (male or female or None) is mentioned above; all categories of gender are included, while different categories are exclusive

Figure 7 provides an example.



patients

subject_id	gender	anchor_age	anchor_year	dod
1	M	100	30	False
2	F	100	45	False
3	M	100	41	True

patients

subject_id	gender	anchor_age	anchor_year	dod
1	0	100	30	0
2	1	100	45	0
3	0	100	41	1

Figure 7: Processing example of variable encoding

3.2.7 Timestamp conversion

There are a lot of time slice interval features in MIMIC, including fixed time period and non fixed time period. The interval of time slice may be ten minutes, several hours, several days and several months, which will be determined by the time difference between various recording times. For example, most time intervals in ICU records are at the minutes-level and hours-level, and most time intervals in inpatient records are at the days-level and months-level. n time-varying data needs to be considered, time series analysis should be used instead of sample grouping.

Time in the database is stored with one of two suffixes: TIME and DATE. If a column has TIME as the suffix, e.g. CHARTTIME, then the data resolution is down to the minute. If the column has DATE as the suffix, e.g. CHARTDATE, then the data resolution is down to the day. That means that measurements in a CHARTDATE column will always have 00:00:00 has the hour, minute, and second values. This does not mean it was recorded at midnight: it indicates that we do not have the exact time, only the date. **All dates in the database have been shifted to protect patient confidentiality**. Dates will be internally consistent for the same patient, but randomly distributed in the future. This means that if measurement A is made at 2150-01-01 14:00:00, and measurement B is made at 2150-01-01 15:00:00, then measurement B was made 1 hour after measurement A. Most data, with the exception of patient related demographics, are recorded with a time indicating when the observation was made:

- **CHARTTIME:** In order to facilitate efficient observations by nursing staff, a day was separated into hourly blocks, and observations are recorded within these hourly blocks.

Thus, any time one performed a measurement between the hours of 04:00 and 05:00, the data would be charted in the 04:00 block, and so on. Even if data is recorded at 04:23, in many cases it is still charted as occurring at 04:00. In almost all cases, this is the time which best matches the time of actual measurement. In the case of continuous vital signs (e.g. heart rate, respiratory rate, invasive blood pressure, non-invasive blood pressure, oxygen saturation), the CHARTTIME is usually exactly the time of measurement.

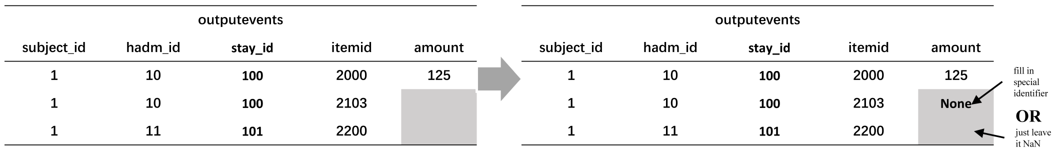
- **STORETIME:** STORETIME data provides information on the recording of the data element itself. All observations in the database must be validated before they are archived into the patient medical record. The STORETIME provides the exact time that this validation occurred. For example, a heart rate may be charted at 04:00, but only validated at 04:40. This indicates that the care provider validated the measurement at 4:40 and indicated that it was a valid observation of the patient at 04:00. Conversely, it's also possible that the STORETIME occurs before the CHARTTIME. For example, while a Glasgow Coma Scale may be charted at a CHARTTIME of 04:00, the observation may have been made and validated slightly before (e.g. 3:50). Again, the validation implies that the care staff believed the measurement to be an accurate reflection of the patient status at the given CHARTTIME.

3.2.8 Missing data preprocessing

Missing values are common in the MIMIC table data. However, it is necessary to distinguish whether the missing value is the default value or the data is missing. For example, some patients who survive at discharge have missing Death FLAG values because survival may be the default value.

For default data, you can fill in the default value, while for actual missing data, you can mark it with ’NaN’ or ’None’. The absence can be brought into the model as an independent category. The key point is to distinguish the actual missing value from the default value to avoid ambiguity in the subsequent analysis of missing values in data mining.

An example of preprocessing missing data is shown in Figure 8.



outputevents

subject_id	hadm_id	stay_id	itemid	amount
1	10	100	2000	125
1	10	100	2103	
1	11	101	2200	

outputevents

subject_id	hadm_id	stay_id	itemid	amount
1	10	100	2000	125
1	10	100	2103	None
1	11	101	2200	

fill in special identifier
OR
just leave it NaN

Figure 8: An example of preprocessing missing data

3.3 What's next?

In the next reading material (MIMIC IV Data cleaning code documentation), you will dive into implementing some of the cleaning steps mentioned above. You will work with MIMIC-IV data and go through the data cleaning pipeline with help of provided python scripts to generate research ready datasets.

References

[1] Wei-Qi Wei, Lisa Bastarache, Robert Carroll, Joy Marlo, Travis Osterman, Eric Gamazon, Nancy Cox, Dan Roden, and Joshua Denny. Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record. *PLOS ONE*, 12:e0175508, 07 2017.

[2] Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C Denny, Evropi Theodoratou, and Wei-Qi Wei. Mapping icd-10 and icd-10-cm codes to phecodes: Workflow development and initial evaluation. *JMIR Med Inform*, 7(4):e14325, Nov 2019.

[3] Penni Hernandez, Tanya Podchiyska, Susan Weber, Todd Ferris, and Henry Lowe. Automated mapping of pharmacy orders from two electronic health record systems to rxnorm within the stride clinical data warehouse. In *AMIA Annual Symposium Proceedings*, volume 2009, page 244. American Medical Informatics Association, 2009.

[4] Swapna Abhyankar, Dina Demner-Fushman, and Clement J McDonald. Standardizing clinical laboratory data for secondary use. *Journal of biomedical informatics*, 45(4):642–650, 2012.

[5] Chuan Hong, Everett Rush, Molei Liu, Doudou Zhou, Jiehuan Sun, Aaron Sonabend, Victor M Castro, Petra Schubert, Vidul A Panickan, Tianrun Cai, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ digital medicine*, 4(1):1–11, 2021.

[6] Doudou Zhou, Ziming Gan, Xu Shi, Alina Patwari, Everett Rush, Clara-Lea Bonzel, Vidul A Panickan, Chuan Hong, Yuk-Lam Ho, Tianrun Cai, et al. Multiview incomplete knowledge graph integration with application to cross-institutional ehr data harmonization. *Journal of Biomedical Informatics*, 133:104147, 2022.

[7] Clem McDonald, Stan Huff, J Suico, and Kathy Mercer. Logical observation identifiers names and codes (loinc®) users' guide. *Indianapolis: Regenstrief Institute*, 2004.

[8] Kejing Yin, William Cheung, Benjamin CM Fung, and Jonathan Poon. Learning inter-modal correspondence and phenotypes from multi-modal electronic health records. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[9] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235, 2020.

[10] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153, 2019.

[11] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*, pages 3512–3520, 2016.

[12] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.

[13] Mrinal Kanti Baowaly, Chia Ching Lin, Chao Lin Liu, and Kuan Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.

[14] Christopher J McWilliams, Daniel J Lawson, Raul Santos-Rodriguez, Iain D Gilchrist, Alan Champneys, Timothy H Gould, Mathew JC Thomas, and Christopher P Bourdeaux. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from mimic-iii and bristol, uk. *BMJ open*, 9(3):e025925, 2019.

[15] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[16] Joon Lee, Evan Ribey, and James R Wallace. A web-based data visualization tool for the mimic-ii database. *BMC medical informatics and decision making*, 16(1):1–8, 2015.

[17] Gaurav Paliwal, Aaquil Bunglowala, and Pravesh Kanthed. An architectural design study of electronic healthcare record systems with associated context parameters on mimic iii. *Health and Technology*, pages 1–15, 2022.